UNIVERSITY OF TWENTE.

MSc Interaction Technology - Thesis

Unlocking domain-specific image annotations with AI

A human-in-the-loop approach for generating interior-design insights using large language models and vision language models

September 2024

Author:
Sandhiyaa B Y

Examination Committee:
dr. Doina Bucur
dr. Lorenzo Gatti
Dan Porder, IKEA

Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

In compliance with confidentiality agreements with partnering company, certain texts in this document have been redacted.

Abstract

The home-furnishing brand IKEA prioritizes its core ability to analyze the interior-design elements of rooms. These analyses enhance its digital capabilities such as image tagging, product recommendations, and personalised interior-design advice. This study attempted to automate such room analyses using automated room image annotations with the GPT-4 series of large language models (LLM) and vision-language models (VLM). The three stages being - 1. Defining a *conceptual framework* for the annotations (determining the key set of interior design aspects the annotations must contain information for); 2. Generating interior design annotations with GPT-4; 3. Checking the validity of the generated annotations. The *conceptual framework* for the annotations was defined through a combination of LLM-driven taxonomy generation and key concepts identification from a workshop with interior designers. The LLM-driven taxonomy generation was a sub-exploration that revealed insights about the knowledge engineering capabilities of GPT-4. It was observed that although GPT-4 is capable of identifying key concepts of a domain, it lacks the expertise to self-assess the outputs and organise these concepts into a robust taxonomy. 10 key concepts (aspects of bedroom interior design) identified, and finalised by experts, in the first stage formed the *conceptual framework* for the annotations. In the second stage, prompts for GPT-4 were drafted for each of these 10 interior design aspects informed by insights gathered from a manual interior design analysis workshop with experts. A prompting strategy of first generating a description of the image focused on one aspect at a time, then converting the description into annotations for that aspect, produced interior design annotations to the desirable level of details. The drafted set of prompts with this strategy were run for 60 bedroom images to generate a sample set of image annotations. In the third stage, this sample set was evaluated with interior design experts. A quantitative evaluation for *incorrectness* and *incompleteness* of the annotations was conducted followed by a qualitative post-evaluation interview with the evaluators. Low *incorrectness* and *incompleteness* of <10% coupled with the qualitative feedback from the evaluators led to the inference that this process of generating interior design annotations with GPT-4, given a *conceptual framework*, is able to produce 'good enough but not expert-like' image annotations. It was therefore identified that the applicable use-cases for this process could be customer room metadata generation, domain-specific dataset creation, etc. Beyond an understanding of how well the AI model can analyse and generate insights about the interiors from a room image, this work also contributes a method for generating domain-specific image annotations with LLMs and VLMs. The *human-in-the-loop* approach in this work, incorporating experts across the process, gave rise to insights that are adaptable for AI Developments beyond interior design. Among them were the value of collaborating with domain-experts for prompt designing and importance of combining quantitative and qualitative methods for human expert evaluations. Recommendations such as co-prompt-design sessions, domain-specific evaluation criteria, increased number of expert-evaluators, were made to improve the process.

Keywords— Artificial intelligence, automated image annotations, domain-expert evaluation, GPT-4, human-in-the-loop, IKEA, image-to-text, interior design, LLM, prompt engineering, taxonomy, VLM

Preface

Over the past few years, generative AI tools like ChatGPT and DALL-E, have gained significant popularity, sparking an interest and enthusiasm in me for this technology. For this thesis, I wished to tap into the potential of generative AI for applications beyond conventional ones like chatbots and image generation. As someone who has always taken a personal interest in interior design, always in awe of aesthetic spaces, this project being a collaboration with a home-furnishing company was a unique opportunity to merge these interests.

With this enthusiasm, I aimed to explore how IKEA's processes could be enhanced through the application of generative AI. The goal was to investigate ways to automate or assist in tasks that are typically resource-intensive, mundane, or tedious, without diminishing the value of human expertise. This project was about strategically utilising the advancements in AI to complement human skills, making it possible to achieve a level of detail and accessibility that might otherwise be impractical.

Reflecting on this journey, I was given great flexibility and freedom to let my curiosity guide the research, which led me to adopt a more exploratory approach to the research question. While a straightforward path might have been less stressful given the time constraints, this exploratory approach resulted in a series of intermediate sub-explorations. Each of these offered valuable insights in areas such as taxonomy generation with LLMs, expert-driven prompt engineering, and the integration of human feedback loops into AI development. These experiences not only expanded my understanding of AI's capabilities but also helped improve my skills in conducting research.

This thesis is the result of a journey to harness the advancements in generative AI to enhance IKEA's digital processes, driven by the belief that a comfortable and personalised home is something everyone deserves. As AI continues to impact everyday life, I hope this research lays a foundation for automation that remains human-centric, not just within interior design but across other domains as well.

Contents

List of Acronyms

AI artificial intelligence

AIA automated image annotations

COT chain-of-thought

CF *conceptual framework*

gen AI generative artificial intelligence

HCI human-computer interaction

HITL *human-in-the-loop*

ID interior design

LLM large language model

ML machine learning

NLP natural language processing

VLM vision language model

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An IKEA Vision

Interior design is a highly visual discipline involving aesthetics and functionality. When it comes to upgrading a space, interior designers make intricate observations about its interior elements such as furniture, lighting, color schemes, etc. See Figure 1.1, illustrating various kinds of elements interior designers consider.

Figure 1.1: Example interior design observations of a bedroom. (Room image source : Lsun Public Bedrooms Database [1])

Understanding room interiors is crucial to the Swedish home-furnishing brand IKEA, which leverages interior design as one of its core expertises. This task of observing and inferring interior-design information from images of rooms is applied in various ways within IKEA. For example, in-house interior designers receive images of customer rooms and make summarised analysis of their interiors to provide personalised design recommendations. Also, in IKEA's online design-your-own-space platform, *IKEA Kreativ* [10], the layout of an uploaded room image is extracted so customers can redesign their room with IKEA products. Furthermore, for IKEA's internal data collection, coworkers add metadata[1] to customer room images, for instance by indicating the type of room (bedroom, kitchen, bathroom, etc), which are used as assets for various purposes like content creation, annual reports analysis, etc.

[1]Metadata refers to supplementary information that accompanies data such as images, serving as an informative identity of the data or even as a substitute for the data. Metadata can include various types of information, including dates, locations, symbolic descriptions, and physical properties of the data. It can be expressed in either free-text format or within more constrained, structured formats [3].

Recent advancements in the field of generative artificial intelligence (gen AI)[2] indicate an automation opportunity for the task of interior-design analysis. Particularly, large language models (LLMs) and vision language models (VLMs)[3] have been shown in studies to possess strong image-to-text capabilities (i.e., generating textual information about an entire image or parts of an image) better than traditional, less advanced machine learning (ML) algorithms, which are trained on significantly smaller amounts of data [12]. LLMs and VLMs having been trained on internet-scale data, also showcase expertise in various domains [13, 14, 15, 16]. Additionally, they demonstrate contextual understanding of images to generate rich and detailed captions or annotations[4] [12, 17].

These capabilities indicate the potential for generating domain-specific information about an image, in the form of image annotations, which has largely been unexplored for the domain of interior design. This work explored that very potential as a means of yielding interior-design insights. Such a capability could optimise the aforementioned use-cases within IKEA by augmenting the work of in-house interior designers, enhancing the *Kreativ* platform with personalised product recommendations, automating room-image metadata generation, and more.

A major advantage of these pre-trained models[5] over traditional machine-learning models is their ability to generate desired output without needing to train the models with datasets containing similar outputs. Gen AI models can be instructed to produce a specific output using a natural-language instructions, known as *prompt*[6]. Generating domain-specific image annotations purely via prompting an LLM/VLM is an untapped area in research that this thesis incorporated.

The domain-specific nature of this exploration called for close collaboration with relevant domain experts. The study therefore was guided by *human-in-the-loop* (HITL) principles [22, 23] ensuring human experts provide validation and refinement throughout the process, maintaining the quality and accuracy of the generated data. This is yet again an under-researched area especially stakeholder-collaboration practices for *gen AI prompt design*[7].

In short, this thesis aimed to develop a HITL process for applying LLMs and VLMs to generate annotations containing interior design insights for images of rooms via *prompts*. To manage the scope of the thesis, the images were only of *bedrooms* and the gen AI models used were from the *GPT-4* series of LLMs and VLMs [24]. To achieve the research objective, a three-stage approach, guided by the sub-research questions, was designed and implemented. The process involved first defining the key aspects of interior design that the annotations must contain insights about, then generating image annotation samples by designing prompts for these key concepts, and finally evaluating these annotations with interior designers within IKEA. Intermediate analyses of the process combined with the evaluation results revealed to what extent this process/method can generate accurate interior design image annotations, thereby to what extent it can be applied to automate the interior design analysis task. The study also led to several key insights regarding the knowledge engineering capability of *GPT-4* for interior design, effective prompting strategies for annotation generation, designing domain-expert-evaluation for AI outputs.

The research question(s) that guided this study can be found in Section 1.2 while Section 1.3 describes

[2] *Gen AI* is a class of artificial intelligence models that can generate new content from learned patterns in existing sets of data. These models can generate various types of content, including text, images, music, and even videos [11].

[3] Common types of gen AI models include large language model (LLM)s, which offer human-like text generation capabilities, and vision language model (VLM)s, which are essentially LLMs with vision capabilities (i.e., image-to-text models).

[4] Image annotations, in this context, are texts that provide meaningful information about the contents of an image. These could be general descriptions of an image, a list of objects found in an image, facial expression descriptions of a photo, etc. The texts seen around the room image in Figure 1.1 can be considered annotations of that particular image.

[5] A pre-trained model is a machine learning model often trained on large or internet-scale datasets, affording it diverse abilities in generic tasks like natural language generation. A pre-trained model can be used as is or customised to suit specific application requirements [18].

[6] A prompt is a set of instructions, in natural language text, provided to a generative AI model that programs it to perform a specific task [19, 20]. These tasks could be answering questions, language translations, image generations, image captioning, etc [21].

[7] Prompt designing is the "art" of curating and tweaking the natural language instruction passed to a *pre-trained* gen AI model to generate the desired output. "Prompt designing" is often used interchangeably with "prompt engineering" [21].

the approach taken to answer the questions and the outline of the document thereby.

1.2 Research Questions

Given the premise described above, the following research question was defined:

RQ 1: How can LLMs and VLMs be applied to generate domain-specific image annotations for interior design with a *human-in-the-loop* approach?

The following sub-research questions were identified to answer the above question:

Since the annotations would not contain a direct description of the image, the initial step involved defining the domain-specific elements of information that the annotations must include. For the use case in this work, interior design analysis of room images, it was essential that the annotations consistently provide insights on various aspects of interior design, such as lighting, furniture, and color schemes. This collection of interior design aspects, or 'concepts,' is referred to as the *conceptual framework* CF[8] for the annotations. In other words, the CF outlines the key aspects of interior design that the annotations must capture.

The term "*conceptual framework*," rather than simply referring to them as a set of interior design aspects, is used in this thesis in an effort to make the approach adaptable to domains beyond interior design. For instance, if agricultural annotations were needed from crop images, the CF could be a set of agricultural aspects. Thus, for this specific use case, the appropriate conceptual framework would consist of the key concepts necessary to thoroughly understand the interior elements in a room image. These key concepts had to be carefully defined with input from domain-specific resources and experts.

The first sub-question to be addressed is, therefore:

SQ 1.1: How can a CF be derived for the annotations to ensure all necessary information is included in them?

The defined CF for the annotations can then be used as a base to construct the instructions (prompts) to a VLM to generate interior design insights. This leads to the second sub-question:

SQ 1.2: How can prompts for a VLM be effectively designed to generate interior design image annotations?

The generated annotations had to then be validated. This required identifying appropriate criteria and designing an appropriate evaluation process to check the quality of the annotations. The quality is crucial as it determines whether the annotation process aligns with human expertise it attempts to replicate and can thereby reliably serve as an automation solution for interior design analysis of images.

SQ 1.3: How can the validity of the generated annotations be checked?

Incorporating human expertise in this process remains vital and unavoidable specifically due to the domain-specific nature of this exploration. Human experts provide invaluable insights, especially in specialized fields like interior design,Moreover, the importance of collaboration with relevant stakeholders in any AI development process has been increasingly highlighted by various experts in the research field [26, 27, 28]. This brings us to a sub-research question auxiliary to all the above questions:

[8] *conceptual framework* in the field of research stems from the philosophical idea of what something is, the idea of every concept being a delineation defined by its components. It has varied definitions among which we shall use the one by Miles and Huberman [25]. They state it is an artifact that "explains, either graphically or in narrative form, the main things to be studied and the key factors, concepts, or variables" [25]

Auxilary-SQ 1.1: How can a *human-in-the-loop* approach be effectively integrated throughout the process?

By addressing these questions, the study explores a methodology that employs a human-in-the-loop approach to generating annotations of interior design insights using LLMs and VLMs.

This exploration ultimately led to a discussion on the effectiveness of this method, thus addressing the overarching research question. To help the reader visualise and better follow how the research question(s) have been explored, Figure 1.2 shows a flowchart.

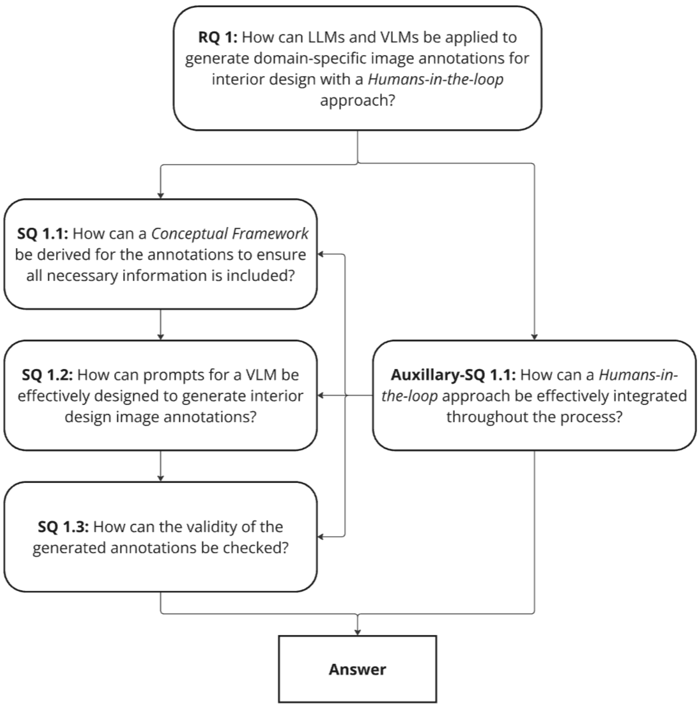


Figure 1.2: Visualisation of flow of answering the research question(s)

1.3 Approach and Outline:

The sub-research questions naturally led to the development of a methodology for this study. The research was structured into three stages, with each stage designed to answer a sub-question. Additionally, the auxiliary sub-question about applying a HITL approach was addressed across all three stages (see Figure 1.2). Further, specific steps were defined to achieve the objective within each stage. These stages and the rational behind the defined steps are explained in the following chapter - Chapter 3. Later in Chapter 4 the implementation of these stages and steps within have been elaborated. In Chapter 5, the insights gained along the way have been discussed, leading to an answer to the main research question. Finally, Chapter 6 encloses this thesis study with a summary of the research process and key insights derived.

1.3.1 Choice of LLM & VLM:

The *GPT-4* series of models were used for this exploration. Particularly, the model *GPT-4-1106-preview*[9] was the LLM in use and *GPT-4 vision preview*[10], the VLM [24]. This choice was made due to the high-performance of the GPT-4 series of models across several benchmarks of various natural language generation and visual data processing tasks [31, 32]. Moreover they outperform other public and proprietary *large (vision) language models* [33, 34, 35]. Large context window [11] of about 90,000 words, availability within the IKEA digital infrastructure were other reasons for the choice.

1.3.2 Choice of Type of Rooms:

A strategic decision was made to make the domain narrower and focus to *bedroom images*. This decision was made to align with other AI explorations within the team. This narrowed scope provided a starting point for the study which can be expanded on in future research.

Interior Design vs Interior Decoration:
The focus of this research lies within the domain of Interior Design, with a specific emphasis on space decoration. While interior design encompasses a wide range of activities, including furniture design, architectural layout planning, and building, it must be acknowledged that this work hones in only on the decorative aspect of the domain. The task of interior analysis performed by AI, which this research explores, aligns closer with the work of interior decorators, a sub-category of interior design professionals who focus on selecting and arranging furniture, colors, textures, and accessories to create a visually pleasing and comfortable environment [37, 38].
The term "interior design" is used throughout this work, rather than "interior decoration," for its broader recognition and to ensure clarity and accessibility for the reader, even though the scope here is more narrowly defined.

* These red boxes throughout the document contain supplementary notes and reflections from me, the author, to help readers better understand my thinking, as well as the context and constraints that shaped this research.

---

[9] All interactions with the *GPT-4-1106-preview* model LLM were through the OpenAI API. As stated in the OpenAI Enterprise Privacy Policy [29], these inputs and outputs are not utilized as training data for other models. Therefore, there is no enterprise level vulnerability due to the data passed to the AI model.

[10] The *GPT-4 Vision Preview* VLM was available as a Microsoft Azure service within the digital infrastructure of IKEA. *zure GPT-4 Vision Preview* is a managed service of the OpenAI models offered by Microsoft, with enterprise level security and data privacy [30].

[11] Context window, in the field of LLMs and VLMs, refer to the amount of text the model can be given as input when generating or understanding natural language. It is measured in tokens which are machine-readable representation of words or parts of words. 1 token can be considered to be 0.75 words on an average [36].

Chapter 2

Related Work

This thesis work taps into various facets of literature. This section outlines related work that informed the exploration of the sub-research questions both directly as well as indirectly.

2.1 Automated image annotations and Conceptual Framework

Automated image annotations AIA in early years were predominantly for object detection, image classification, image tagging with keywords. (An example of image annotations is shown in Figure 2.1). The objective of this study is similar to the image-tagging-with-keywords category of AIA, but it extends beyond simply tagging with keywords. Instead, it focuses on generating detailed textual information (ID insights) about the contents of an image. Another computer vision task this work is akin to is image captioning, which involves generating descriptive text related to an image. However, the task that the AI performed in this work was not purely captioning, as the aim was to produce insightful phrases linked to specific interior design aspects of the image, rather than general descriptions. This places the task somewhere between captioning and annotating. In a slightly non-classical sense, annotation can be considered a broader term that can include image captioning, hence the task that the AI performs is referred to as image annotation in the document.

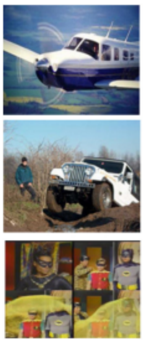


Figure 2.1: Example of image annotations [2]

This process of transcribing visual content into a linguistic expression is commonly referred to in research as *semantic analysis* of images [2]. Early vision computing systems (ML models for image processing) faced the issue of the *Semantic Gap*. *Semantic gap* refers to the difference between low-level features (e.g. colours, shapes) that computers can automatically extract from data, such as images or audio, and the high-level concepts (e.g. facial expressions, scene understanding) or meanings that humans perceive in the same data [39].

To address the semantic gap, early research incorporated knowledge structures such as *taxonomies*, *ontologies*, and *thesauri*. These are forms of representing the concepts within a domain/topic in an organised structured manner. Using these as a guide for the annotations, helped break down the complex content of annotations into more manageable components that computer vision models could effectively map.

- **Taxonomy:** A structured hierarchy used to classify objects or concepts, typically showing ”Is-A” relationships, like a subtype/super-type connection (e.g., ”car” is a type of ”vehicle”).
- **Thesaurus:** An expanded version of a taxonomy that not only shows hierarchical relationships but also connects synonyms, explains word usage, and includes related terms within the same domain.
- **Ontology:** A representation of concepts, their properties, and the diverse relationships between them, allowing for more complex connections than those in a thesaurus.

Figure 2.2 developed by Tousch et al. [3] shows a sample representation of a knowledge structure incorporating different types of relationships predominantly used for object recognition applications.

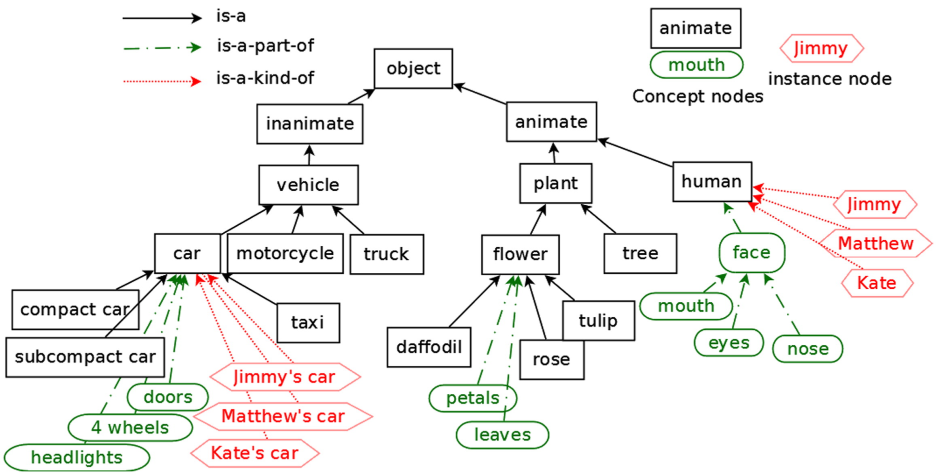


Figure 2.2: Example of a knowledge structure with relations between concepts usually met in object recognition. Image by Tousch et al. [3]

One example of such work is presented in a study by [4], where the authors train a model to automatically annotate parts of images using a defined hierarchical ontology specific to the domain of chest radiography. This ontology includes categories corresponding to the various anatomical parts that might appear in the images, allowing for precise and domain-specific image annotations (see Figure 2.3).

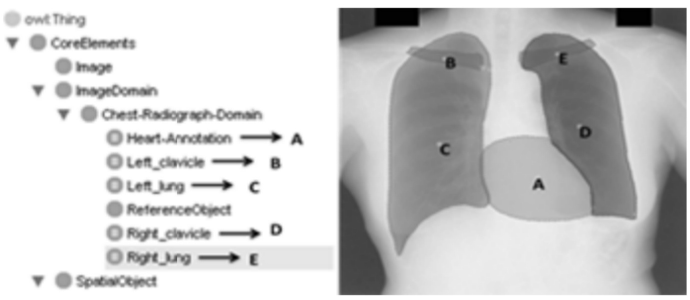


Figure 2.3: An image from the work by Smailis and Iakovidis [4] showing the use of a hierarchy of anatomical parts to label segments of an x-ray.

This semantic gap has largely been bridged by the advent of gen AI models with vision capabilities. The combination of enhanced natural language processing and contextual understanding of image data, VLMs are capable of generating more nuanced and contextually relevant annotations, thereby reducing the earlier need for knowledge structures breaking down the high-level concepts to low-level machine understandable concepts.

However, even with gen AI, the necessity to define ”what the annotations must contain” persists for domain-specific image annotations. Recent AIA studies using advanced ML models address this challenge

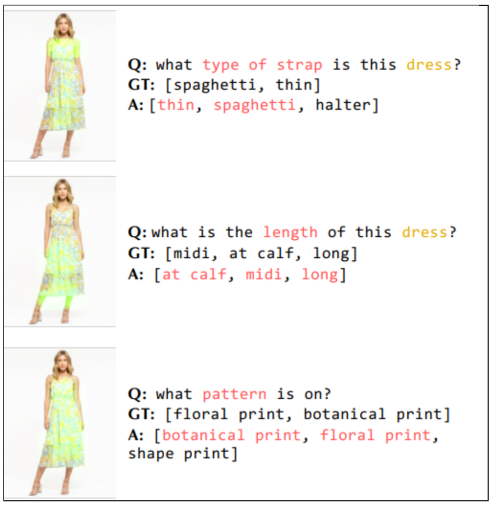


Figure 2.4: Image from visual question answering study by Wang et al. [5]

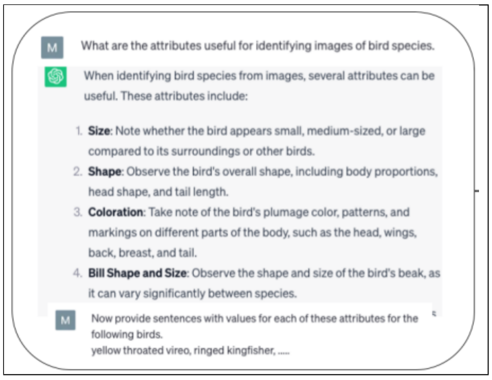


Figure 2.5: Image from study by Maniparambil et al. [6] showing a prompt used to identify all 'useful' attributes of an image

by employing visual question answering, which focuses on extracting and describing specific attributes of an image. Wang et al. [5] used questions about specific attributes of information found in the ground truth (GT) dataset to generate descriptions for the specific attribute. Maniparambil et al. [6] and Zhu et al. [40] use an LLM to automatically define these questions to generate a comprehensive image caption.

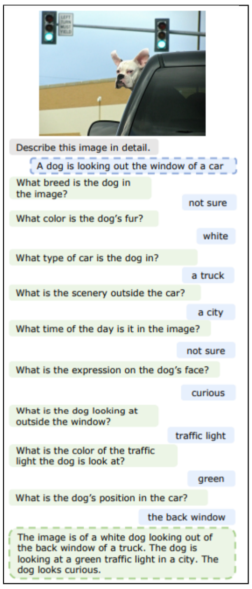


Figure 2.6: Image from work by Ghosh and Anupam [7] on auto-captioning with visual question answering using an LLM

It is evident the above image annotation studies focused on extracting information related to specific aspects of an image is more effective than directly instructing an AI model to generate annotations. However these aspects, in these studies, are not verified to ensure that they comprehensively cover all aspects of the image which is required for the objective of this study. Picking a generic list of interior design concepts will not be sufficient to ensure a complete interior design analysis of the image. This is a gap that this seeks to address by developing an approach that could be used to define a set of interior design concepts that could be used to get a sufficient understanding of the interiors of a room image, hence developing an approach to define a CF for domain-specific annotations.

Recalling the structured knowledge used in early AIA efforts, a *taxonomy* of interior design would encompass all key aspects of the field, hierarchically categorized to collectively represent the domain. One approach to identifying these key aspects for the CF could be from an interior design taxonomy. However, at present, there is no established classification system for interior design concepts. The manual generation of such a taxonomy is highly iterative and labor-intensive as it would require extensive expert input and repeated revisions to accurately categorize and refine a comprehensive set of concepts and relationships. Here, an opportunity to develop such a taxonomy using an LLM was identified.

2.1.1 LLM-Driven Knowledge Engineering:

In the field of structured knowledge generation, the use of LLMs have been largely studied specifically for knowledge graphs construction [41, 42, 43, 44]. Knowledge graphs are large-scale graph-structured representations of information using named entities (nodes) and relationships (edges) and are extensively used in computing systems as structured knowledge is easily machine-readable. The studies that automated knowledge graph construction demonstrate the capability of LLMs to semantically comprehend named entities within a corpus of unstructured text and extract them along with their relationships more meaningfully compared to traditional algorithms.

A taxonomy can be regarded as a specialized type of knowledge graph, with entities (concepts) organized in a hierarchical super-class/sub-class structure.

Particularly for taxonomy generation, Shah et al. [45] employed an LLM to construct taxonomies of user intentions from log data comprising user requests expressed in natural language.

It was noted that these studies primarily relied on unstructured data to identify concepts and relationships, without much leveraging the pre-trained domain knowledge of the models.

The objective of this thesis was to develop a taxonomy of interior design in order to extract key concepts of interior design for the CF, specifically tailored for room analysis, rather than directly structuring unstructured documents. Here, GPT-4 was used to generate an interior design taxonomy, with unstructured documents serving as reference data rather than the primary source for structuring.

Another motivation for this approach to generating a CF for the annotations was that it provided an opportunity to investigate the potential of LLMs for taxonomy construction, thereby contributing to the understanding of their capabilities in the field of knowledge engineering and automation research.

2.2 VLMs for Interior Design

Perhaps, due to interior design being a domain about aesthetics and visual appeal, much of the research works in this area have been focused on image generation and rendering with generative capabilities [46, 47, 48, 49]. A high concentration of studies on image or visual media generation related solutions highlights a lack in innovative explorations of the plethora of generative AI capabilities in interior design. Yanhua [50] in their article state that there remains a gap in comprehensive understanding of how these technologies can be effectively integrated into design practices. To this end they put together an analysis of various applications of AI in interior design within current research. These include generating alternative design solutions, making sustainable design recommendations, and synthesizing large datasets to understand user preferences. This analysis indicates that automatic analysis of images for interior design insights is an untapped application of AI in this domain, which is the focus of this study.

The study by Hou et al. [51], however, is a good example of looking at the use of generative AI in interior design from a different lens. They utilised an LLM to effectively translate user textual inputs into interior color design prompts which were then used as instructions to image rendering system.

This thesis research is yet another innovative application of LLMs and VLMs within the domain of interior design, aiming to derive domain-specific insights by leveraging their image analysis capabilities.

There have been several studies that show proficient image analysis capabilities of VLMS [40, 52, 7, 53], showing promise for its application for interior design analysis task.

Although several studies explore the knowledge of pre-trained LLMs and VLMs in various domains such as law [54, 14] and medicine [55, 56], but there is none for interior design. This gap also points to the lack of datasets within the domain against which the model results can be evaluated, making this research heavily reliant on evaluations by domain experts. By generating interior design insights and having them evaluated, this thesis research takes a step towards understanding the interior design expertise and capabilities of these models, specifically GPT-4, which is the suite of models used in this study.

2.3 AI Output Evaluation

Image annotation tasks have traditionally relied on established datasets of input-output pairs for evaluation purposes. These evaluations typically employ quantitative metrics such as precision, recall, F1-score, etc., to assess the performance of annotation systems against the ground truth dataset [2].

However, in the domain of interior design, there is a notable absence of a validated dataset of image-annotation pairs to enable such large scale quantitative evaluation. Thus, in order to gauge the AI's image analysis potential, expert-evaluations were the only way.

Sometimes, human-evaluations are used to complement the evaluations with ground truth datasets. The criteria used differ based on the specific use-case and type of output. For example:

- Tariq et al. [57] developed a domain-specific LLM specialised in prostrate cancer and evaluated its performance for question answering using a user evaluation on the criteria - *correctness* according to clinical guidelines, *completeness* of the response covering all aspects of the ground truth answer and *relevance* of the response to the question.
- Otani et al. [58] in their review about human evaluation for 'text-to-image' mention that *overall quality*, *correctness of object location*, *consistency across multiple image generation* are popular criteria.
- Wu et al. [59] in their work about image annotation 'like humans' specifically studied the annotations for *distinctiveness* and *relevance*.
- Karpinska et al. [60] survey multiple open ended text generation studies and put together the most commonly used criteria for this task - *fluency*, *gramaticality*, *overall "quality"*, *relevance*, *coherance* and *likeability*.

Given the lack of studies specifically focused on domain-specific image annotation tasks like the one undertaken in this research, the evaluation criteria were developed by drawing inspiration from the criteria used in related studies. *Correctness* to evaluate how well the annotations matched the precise details of the room image and *completeness* to ensure that the annotations are not missing any necessary details, were deemed most relevant to the use-case of this research. The criteria of evaluation was therefore defined by building upon these.

While the other criteria are valuable in different contexts, they seemed too nuanced or irrelevant for the scope of this study among other reasons such as criteria like *overall quality* and *relevance* carry high risk of ambiguity for the interior-design domain.

2.4 Prompting Strategies

Several works have explored and collated techniques for good prompts. Brown et al. [61] demonstrated various approaches for in-context learning of an LLM. These include *Few-shot prompting*, where model is given two or more demonstrations of the task as part of the prompt; *One-shot prompting*, where the model is given only one demonstration of the task and output in the prompt and *zero shot prompting* , where the model is prompted with no demonstrations of the task or output.

While such works highlight the efficacy of example-based prompts, the niche downstream tasks in this current study would have demanded meticulous preparation of such examples. Moreover, the domain-specific nature of the tasks made it resource-intensive to define appropriate examples. Hence, *zero-shot prompting* was predominantly employed in this study.

Outside of the work by Brown et al. [61], Wei et al. [62] illustrate how generating a chain of thought, a series of intermediate reasoning steps, can enhance an LLM's reasoning abilities. They introduce the concept of chain-of-thought prompting, where the LLM is instructed on 'how to think', enabling the model to reason its way to a desired result. Furthermore, Zhou et al. [63] explored least-to-most prompting a strategy to break down complex problems into simpler sub problems. Lampinen et al. [64] showcase that explanations in prompts enhanced the LLM's learning capabilities on complex tasks.

These prompting strategies informed the emphasis on breaking down instructions into simpler steps across various stages of this research.

Moreover, Bsharat et al. [8] designed 26 guiding principles for better prompts which were referred for defining the prompts throughout this study. See Figure 2.7.

#Principle	Prompt Principle for Instructions
1	If you prefer more concise answers, no need to be polite with LLM so there is no need to add phrases like "please", "if you don't mind", "thank you", "I would like to", etc., and get straight to the point.
2	Integrate the intended audience in the prompt, e.g., the audience is an expert in the field.
3	Break down complex tasks into a sequence of simpler prompts in an interactive conversation.
4	Employ affirmative directives such as '*do,*' while steering clear of negative language like '*don't*'.
5	When you need clarity or a deeper understanding of a topic, idea, or any piece of information, utilize the following prompts: o Explain [insert specific topic] in simple terms. o Explain to me like I'm 11 years old. o Explain to me as if I'm a beginner in [field]. o Write the [essay/text/paragraph] using simple English like you're explaining something to a 5-year-old.
6	Add "I'm going to tip $xxx for a better solution!"
7	Implement example-driven prompting (Use few-shot prompting).
8	When formatting your prompt, start with '###Instruction###', followed by either '###Example###' or '###Question###' if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data.
9	Incorporate the following phrases: "Your task is" and "You MUST".
10	Incorporate the following phrases: "You will be penalized".
11	Use the phrase "Answer a question given in a natural, human-like manner" in your prompts.
12	Use leading words like writing "think step by step".
13	Add to your prompt the following phrase "Ensure that your answer is unbiased and avoids relying on stereotypes."
14	Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output (for example, "From now on, I would like you to ask me questions to ...").
15	To inquire about a specific topic or idea or any information and you want to test your understanding, you can use the following phrase: "Teach me any [theorem/topic/rule name] and include a test at the end, and let me know if my answers are correct after I respond, without providing the answers beforehand."
16	Assign a role to the large language models.
17	Use Delimiters.
18	Repeat a specific word or phrase multiple times within a prompt.
19	Combine Chain-of-thought (CoT) with few-Shot prompts.
20	Use output primers, which involve concluding your prompt with the beginning of the desired output. Utilize output primers by ending your prompt with the start of the anticipated response.
21	To write an essay /text /paragraph /article or any type of text that should be detailed: "Write a detailed [essay/text /paragraph] for me on [topic] in detail by adding all the information necessary".
22	To correct/change specific text without changing its style: "Try to revise every paragraph sent by users. You should only improve the user's grammar and vocabulary and make sure it sounds natural. You should maintain the original writing style, ensuring that a formal paragraph remains formal."
23	When you have a complex coding prompt that may be in different files: "From now and on whenever you generate code that spans more than one file, generate a [programming language] script that can be run to automatically create the specified files or make changes to existing files to insert the generated code. [your question]".
24	When you want to initiate or continue a text using specific words, phrases, or sentences, utilize the following prompt: o I'm providing you with the beginning [song lyrics/story/paragraph/essay...]: [Insert lyrics/words/sentence]. Finish it based on the words provided. Keep the flow consistent.
25	Clearly state the requirements that the model must follow in order to produce content, in the form of the keywords, regulations, hint, or instructions
26	To write any text, such as an essay or paragraph, that is intended to be similar to a provided sample, include the following instructions: o Use the same language based on the provided paragraph[/title/text /essay/answer].

Figure 2.7: Guiding principles for prompt design by Bsharat et al. [8]

These previous literature informed this study's emphasis on breaking down instructions into simpler steps across various stages of this research. Among the different ways in which the output of a generative AI model be manipulated, prompting is highly resources and time efficient yet effective [65, 66],given the right prompting strategies, in the above-mentioned literature. This research thus explored the potential of generating interior design annotations of images using an LLM and a VLM by manipulating the model output via prompts.

2.5 Humans in the Loop

HITL refers to the strategic incorporation of manual human intervention at various stages of an artificial intelligence development process [23, 67]. As generative AI tools producing human-like outputs proliferate, the importance of implementing relevant HITL strategies becomes increasingly critical to ensure these tools do not steer away from human values and needs. This approach reflects a commitment to the principles of Human-Centered AI, which emphasize the need for human oversight, ethical considerations, and the complementary strengths of human and artificial intelligence [26, 68].

An HITL approach in the era of generative AI is crucial for many reasons including Quality Assurance and Error Correction, ethical supervision, contextual understanding of the use-case, continuous improvement [69, 70]. In this study, which focuses on AI performing niche expert-level tasks, the HITL approach is primarily utilized to ensure that AI outputs reflect human expertise. This has been done via a set of interviews, workshops and consultation sessions throughout the thesis study. By keeping humans involved in the AI development process, researchers can ensure that human expertise and intuition continue to play a vital role, preventing over-reliance on AI and maintaining a balance between artificial and human intelligence. [69, 70]

A survey by Wang et al. [71] examines various ways HITL methods have been employed in recent research in natural language processing (NLP) tasks. The methods generally involve gathering feedback from stakeholders on raw training data for machine learning, on intermediate outputs for iterative improvement, or on the final outputs of ML models. Since some form of an interface is generally required for collecting feedback for AI generated data, the authors emphasize the importance of human-computer interaction (HCI) principles in designing user-friendly feedback interfaces, as these impact the quality of collected feedback and, hence, the performance of downstream tasks. Notably, their survey observes a tendency in research to favor numerical or quantitative feedback over natural language feedback. While numerical feedback is easier to incorporate into AI systems, it often provides limited information compared to natural language feedback. Therefore, the authors recommend future research to adopt a combination of quantitative and qualitative feedback mechanisms, particularly for complex feedback scenarios. This recommendation has been adopted in this study.

Amershi et al. [72] conducted a comprehensive survey on user feedback loops in interactive machine learning through rapid, focused and incremental evaluations of model outputs allowing users to interactively examine the impact of their actions and adapt their subsequent inputs to obtain desired outputs. By synthesizing these interactive machine learning studies, they highlight some key learning about HITL that informed the expert-evaluation design of this work:

- The feedback systems must account for human factors like frustrations and interruptibility as they have the potential to bias the feedback provider.
- People generally tend to want to provide more unstructured feedback than just data labels.
- People tend to give richer feedback if they have an understanding of the workings of the system they are evaluating. They highlight that users are not always satisfied by "black box" learning systems-sometimes they want to provide nuanced feedback to steer the system.

This research is heavily focused on utilsing the capabilities of LLMs and VLMs through *Prompts*. There is, however, little to no existing literature on having humans-in-the-loop within the development stages of a generative AI system, specifically in prompt design. Zamfirescu-Pereira et al. [73] highlight the challenges in having non-experts, in this case people who do not work with LLMs, directly design the prompts. They highlight that the non-experts were generally unsure of what modifications they could make to prompt templates, that they expected human behaviour in the outputs, and that they exclusively chose polite language for prompting. It is clear that gen-AI prompting, although a task in natural language is not an evolved skill-set among general population.

As a way to tackle this and to fill this gap in the literature, this study takes a more collaborative approach with interviews and workshops to translate expertise into effective prompts. This study therefore is one of the firsts to incorporate expert collaboration in the prompt design stages without requiring their expertise in prompt engineering.

Chapter Summary

In this chapter the key gaps in research that this work addresses, has been elaborated. Notably, there is no established method or framework for obtaining standardised, consistent annotations from VLMs, which is tackled in this thesis by creating a *conceptual framework*(CF) for the annotations. By using an LLM to generate a taxonomy of interior design and derive the CF, this study also discovers to what extent an LLM could be used for such a knowledge engineering task. It was also identified that such an image-to-text application of gen AI employed in this work, is under-researched within the domain of interior design. The thesis also draws insights from related works on AI evaluation and prompt engineering. Furthermore, there is limited research on *human-in-the-loop* considerations for intermediate phases of AI development like prompt-design, which this research work seeks to address.

Chapter 3

Method Design

Informed by the sub-research questions, a methodology to apply LLMs and VLMs to generate interior design image annotations was designed. A three-stage process with further steps within was defined. This section maps out the methodology which is proposed as *an* answer to the overarching research question RQ 1.

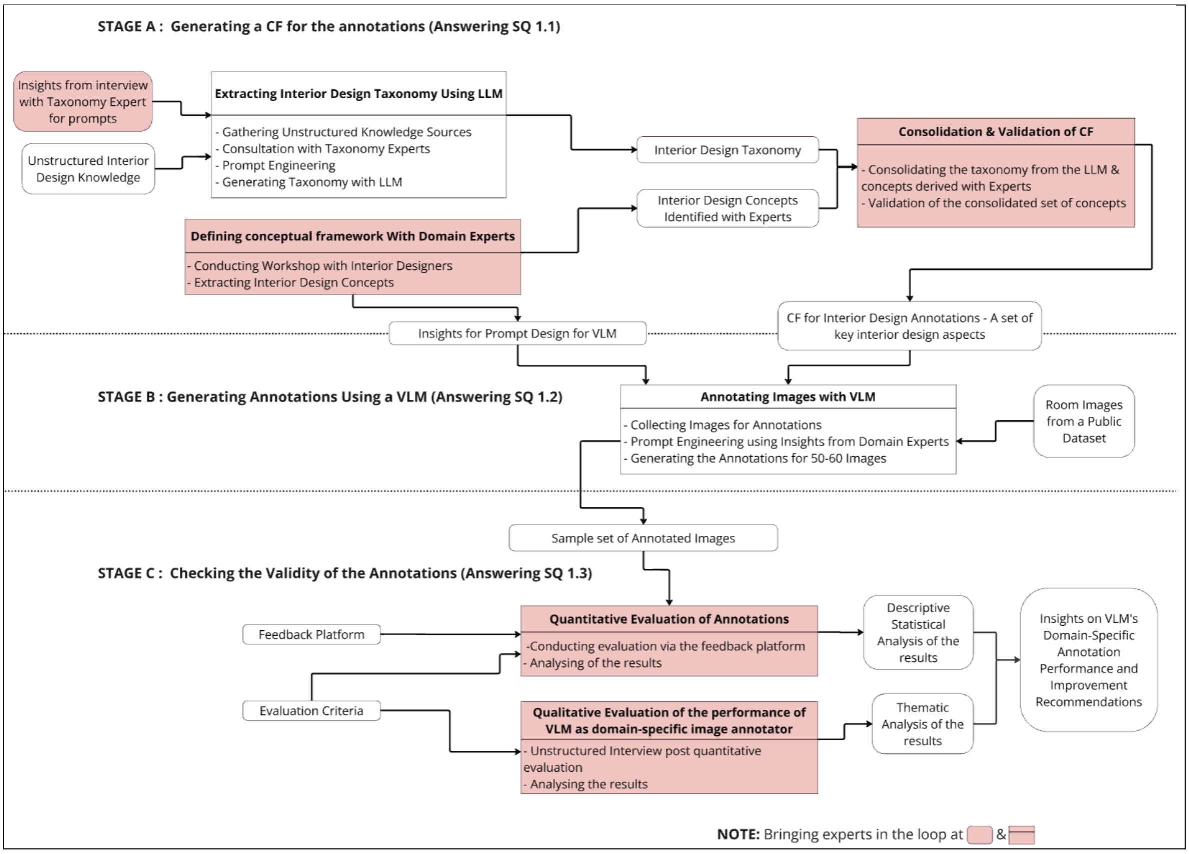


Figure 3.1: Methodology showing the three stages and steps within

3.1 Stage A - Defining a *conceptual framework* for the annotations:

3.1.1 Why is a CF needed for the annotations?

- **To capture all necessary information:** Defining a framework for what the annotations must contain will ensure that all relevant aspects of the target domain get addressed in the annotations. In the case of the interior design domain, it would help guide the AI in generating annotations that cover all elements of the room, thereby leading to richer and more useful annotations.

- **To steer the inductive bias:** AI models have *inductive biases*. These are inherent assumptions and tendencies that an AI model uses to make generalizations and make predictions of the output [74, 75]. In this context, this means that without specific guidance, the AI might generate generic descriptions that align with its training on diverse datasets. For example, a simple prompt was passed to a VLM[1] to generate interior design insights for two different images of rooms, as shown in Figure 3.2. However, the aspects the AI focused on for each image were different, highlighting challenges in achieving reliability and consistency in the annotations.

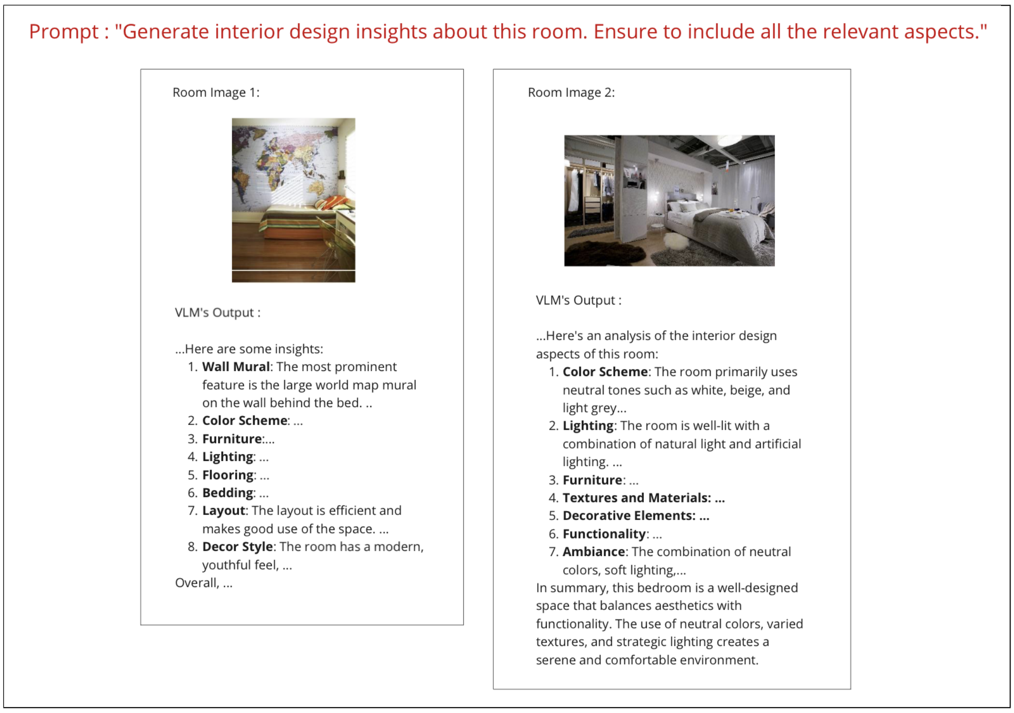


Figure 3.2: Output of a VLM for two room images using the same prompt. The aspects focused on Room Image 1 is different from those of Room Image 2. (See Appendix A.1 for full outputs)

By defining the key concepts of the required information, the AI's inductive bias could be steered towards focusing on the specific aspects of interior design that are most relevant to the task at hand and to IKEA's business need. This ensures that the generated annotations are consistent and rich in domain-specific details, capturing the full scope of interior design insights.

[1]The VLM used for this mini experiment was GPT-4.

3.1.2 Key steps within this stage:

Extracting Interior Design Taxonomy using an LLM:

A *taxonomy* is a hierarchical structure that organizes all the concepts within a domain [76, 77]. It can be seen as a blueprint of the domain. As mentioned in Section 2.1, a taxonomy of interior design could be useful to identify a key set of interior design aspects that could form the *conceptual framework* for the annotations. Due to the absence of an established interior design taxonomy, the ability of LLMs to generate structured knowledge could be utilised to generate such a taxonomy. This step involved using GPT-4 to extract all key aspects of interior design through a taxonomy. In order to enhance the domain knowledge of the LLM unstructured interior design knowledge sources were also considered to be used as reference data along with the prompts[2]. This step required collaboration with taxonomy experts to improve instructions to the VLM and assess the results.

Defining the CF with interior design experts:

Another source of domain information is the expertise of professionals i.e., interior designers. Conducting tasks that involve designers brainstorming and discussing key concepts of interior design or having them analyze interiors of a room are effective methods for deriving implicit concepts. These activities allowed experts to articulate their thought processes and reveal the key elements they consider important, which can then be integrated into the annotations' CF.

Consolidation and Validation:

Since two sources of key interior design concepts were considered, the last step within this stage involved consolidating the insights from the LLM and human experts to create a final CF for the annotations (The final step in Stage A. See Figure 3.1). A validation of the resulting framework from the interior designers would help ensure that it is accurate and complete.

3.1.3 Input & Output of the Stage:

The inputs to this stage were unstructured knowledge sources (e.g., IKEA documents, articles) and the results of the consultations with experts (both taxonomy experts and interior designers). These inputs provide the raw material and expert insights necessary for defining the key elements of interior design information the annotations must contain i.e., the CF. The output of this stage is the CF, a set of interior design concepts/aspects, for the annotations that will help guide the VLM to generate standardized, consistent interior design insights.

3.2 Stage B - Generating Annotations Using a VLM:

With the generated CF for the annotations, the next stage was to generate the annotations for images using a VLM. This stage predominantly involves constructing the prompts for GPT-4V to generate interior design insights from images, for which the following steps were defined. Upon successful implementation, these steps helped answering the sub question SQ 1.2.

[2]Using proprietary IKEA documents for this purpose also had the potential advantage of the taxonomy terminology following that used within IKEA, hence making the taxonomy more business-relevant.

3.2.1 Key steps within this stage:

Constructing the initial prompts:

The first step was to construct the prompts that guide the VLM to generate annotations based on the predefined set of interior design concepts/aspects (CF). Various prompting styles and techniques were explored to optimize the design of the prompts. Techniques such as *chain-of-thought prompting* [62], where the AI is guided step-by-step through its reasoning process, will be considered. Experimenting with different types of prompts helped determine an effective method for the task.

Expert collaboration for prompt design:

The involvement of domain experts yet again were imperative for this phase. It must be considered that although prompts to the VLM are instructions in natural language, it may not be ideal to have experts directly design these prompts as they may not be fully acquainted with the technical aspects of the AI. Instead, their expertise were leveraged to inform the prompt design through alternative methods. Conducting workshops where experts analyze a space and provide their insights was one such approach. The information gathered from these workshops then informed the design of the prompts (one of the inputs shown in Stage B of Figure 3.1).

Generating Data Samples for evaluation :

With the defined prompts, the final step in this stage was to generate a sample set of image annotations that could be evaluated by the experts. The number of images for generating interior design annotations depended on the availability of domain experts for evaluation.

3.2.2 Input & Output of the Stage :

The input for this stage was the collected images and set of interior design concepts, the CF for the annotations. The output was a small dataset of annotated images, each annotation containing detailed interior design insights generated by the aVLM.

3.3 Stage C - Checking the Validity of the Annotations :

The stage tries to answer the sub-research question SQ 1.3 by defining a process to evaluate the annotations generated with the help of domain-experts. Human evaluations were the main source of feedback regarding the AI's interior design analysis performance, due to the absence of a dataset for ground truth.

3.3.1 Key steps within this stage:

Defining the evaluation criteria:

Criteria for evaluation must be chosen according to the requirements for the domain-specific annotations and defined clearly to ensure they can be understood by the domain experts who are not familiar with technical terms.

Conducting the Validity Check:

The validity of the annotations is recommended to be checked both quantitatively and qualitatively [71] (see stage C of Figure 3.1).

- Quantitative evaluation involved asking interior design experts to evaluate each annotation against a set criteria.

- Qualitative evaluation involved asking open-ended questions about the overall performance and results, capturing the more qualitative opinions of the domain experts. This step was defined in an attemped to gather deeper insights into their perceptions of the AI's ability to perform comprehensive image analysis of bedrooms.

Analysing the Results:

The results were later analyzed to derive an understanding of the performance of the VLM as domain-specific image annotators and identify areas for improvement (final output of stage C shown in Figure 3.1). This analysis also helped determine whether the method was successful in generating valid annotations, in turn helping determine if the proposed process can be deemed as an answer to the research question.

3.3.2 Input & Output of the Stage :

Input to this stage was the set of images with corresponding interior design insights as annotations. Other requirements for checking the validity of the annotations were the evaluation criteria and an interface, a feedback platform for effective quantitative evaluation.
The output of this stage were results of the evaluation, leading to an answer to the research question.

With the stages and steps laid out, this methodology was implemented This method is proposed as an answer to the research question (see Section 1.2). This means that a successful implementation of this proposed process would mean that this is indeed an effective method to generate domain-specific annotations using a AI. The implementation and results of the process is explained in the following chapter.

The methodology was designed such that each stage directly addresses a sub-research question in the same order that it has been presented in the Section 1.2. The auxiliary sub-research question regarding the Humans-in-the-loop was addressed in every stage of the methodology through collaboration with experts in the form of interviews, workshops, consultations.

Chapter Summary

This chapter elaborated on the methodology designed to systematically address the research question of this study. The approach was structured into three stages:

1. Defining a conceptual framework for the annotations.
2. Generating the Annotations Using a VLM.
3. Checking the Validity of the Annotations.

Each stage was broken down into executable steps, detailing the respective inputs and outputs.

Recap of Sub Research Questions

To help readers follow along better, this recurring section recalls the research questions asked and explains if and how they have been answered so far.

SQ 1.1: How can a CF be derived for the annotations to ensure all necessary information is included in them?
Not yet answered. The steps in stage A were defined to be a potential answer to the question.

SQ 1.2: How can prompts for a VLM be effectively designed to generate interior design image annotations?
Not yet answered. The steps in stage B were defined to be a potential answer to the question.

SQ 1.3: How can the validity of the generated annotations be checked?
Not yet answered. The steps in stage C were defined to be a potential answer to the question.

Auxiliary SQ 1.1: How can a *human-in-the-loop* approach be effectively integrated throughout the process?
Not yet answered. Opportunities for where HITL practices could be incorporated were identified when designing the methodology.

Chapter 4

Realization

In the previous chapter, the design of the methodology has been explained. This chapter will elaborate how each of those steps were implemented, documenting the decisions made along the way to the desired outputs at each stage, and eventually leading to the answer of the research question.

4.1 Stage A : Defining The Conceptual Framework Answering SQ 1.1

This section details the implementation of the first stage - the approach taken to generate a CF for the interior design annotations. To recall, Figure 4.1 zooms in on the first stage from the overall methodology. This stage elaborates the various ways in which the domain knowledge of interior design was captured from multiple sources.

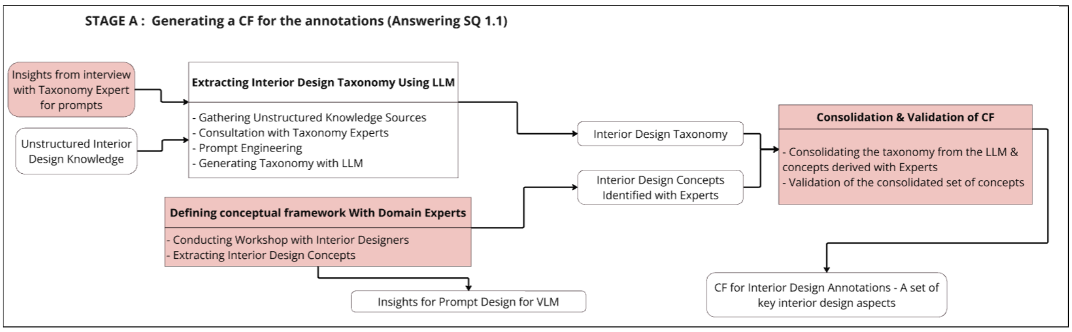


Figure 4.1: Snippet of Stage A from the Overall Methodology shown in Figure 3.1

4.1.1 Identifying key concepts of Interior Design Using LLM - An LLM-driven Taxonomy Experiment

As discussed in Section 2.1.1, the potential of an LLM to generate a taxonomy for interior design could be utilised for developing the CF. This potential led to a sub-exploration of LLM-driven taxonomy generation using prompts and unstructured documents (to enhance its understanding of the domain). To recall, the LLM used was *GPT-4-1106-preview*. This sub-exploration contributes an understanding of the possibility of automating taxonomy generation process with GPT-4.

Gathering Knowledge Sources:

Proprietary IKEA documents that were relevant to this domain and use-case were gathered [1]. These were chosen by gauging how comprehensively they covered the domain of interior design and if they contained applicable information for the specific use-case of room interiors analysis. For example, it was considered if the document held information regarding renovating a space, if it held information regarding understanding the interiors of the space, etc.

Below is the list of documents that were used for this exploration :

Proprietary IKEA Documents:

1. Scrape of the content of over a 100 *'How-To'* articles from the IKEA US website.[78]. These articles cover a wide-range of interior design topics like decorating tips, maximising small spaces, storage solutions, Lighting design, eco-friendly design, etc.

2. A document detailing Home Furnishing and Interior Design Ideas used by the IKEA experts. It details steps that interior interior designers could use to creatively design a space for a client. It therefore contains descriptive information regarding various aspects of ID that contribute to the aesthetics and functionality of the space. (This is an internal IKEA document)

The rationale for using these unstructured documents was that when directly input with the prompt, the LLM would interpret their content as part of the instructions to follow, thereby equipping it with domain-specific knowledge to produce more "informed" outputs. Ideally, a comprehensive guide detailing expert methods for space analysis would be optimal as a context for the LLM. However, in the absence of such specialized resources, the use of less specific, unstructured data such as the documents mentioned above provide a valid alternative. An informed assumption was made here that the pre-training of the GPT 4 model with large amount of data and it's contextual understanding capability would potentially compensate for gaps in the provided unstructured information.

Prompting trials to test taxonomy generation ability of GPT-4:

Multiple trials of prompting were made to generate a taxonomy specifically for interior design room analysis with the unstructured documents as reference data for the LLM.

Despite the availability of guidelines for prompt design, the process of developing an effective prompt remains largely one of trial-and-error until the desired outcome is achieved.

Nonetheless, these prompt were in general informed by the best practices outlined by Bsharat et al. [8].

The following bullet points summarise the most distinct prompt trials:

1. Generating a taxonomy without providing any contextual documents. This was to compare and analyse the impact of the unstructured knowledge sources in the output. See Figure 4.2 for an example of this trial.

2. Attaching the proprietary IKEA documents in the prompt and instructing the LLM to extract the key concepts that is relevant for room interiors into a taxonomy. (The large context window of *GPT-4-1106-preview* enabled the documents to be directly attached with the prompt). See Figure 4.3 for an example of this trial.

3. Instructing the LLM to first summarize the documents, then to generate a taxonomy with the summaries. This approach was motivated by GPT-4's strong text summarisation capabilities [79]. If the summaries

[1] For this study, proprietary IKEA documents were utilized with explicit permission from the company as part of their ongoing AI research initiatives. While non-proprietary interior design resources, such as publicly available books and articles, could potentially contribute to this task, their inclusion would have necessitated obtaining permissions from the respective copyright holders. Furthermore, even documents in the public domain, open for anyone to use, might present legal ambiguities. Therefore, to ensure ethical and legal compliance, this study only utilised the authorized proprietary documents.

proved sufficient to produce similar taxonomies, it could be utilised to significantly conserve the context window and processing power, allowing for the inclusion of more information or documents within the context. See Figure 4.4 for an example of this trial.

4. *Chain-of-thought* multi-stage prompt breaking down the task of taxonomy generation, designed with insights from interviews with a taxonomy specialist (The process of developing this prompt has been elaborated in the below section - Taxonomy-COT Prompt Design.

NOTE All prompts in this stage often included an instruction for the response to be given in JSON format[2], as it is a suitable representation format for hierarchical structures like a taxonomy.

The *temperature* parameter, which determines the randomness of the model's output [3], was experimented with and a range between 0.5 to 0.7 was found to provide an ideal balance of detail and relevance.

[2]JSON is a information-exchange format in which data is represented in key-value(s) pairs where a key is a labelled group and the values are the potential pieces of data associated with that label. These values can be simple (like a number or text) or more complex (like lists or other groups of data), allowing for a clear and flexible representation of structured information.

[3]A low temperature value gives more consistent outputs, while a high value generates more diverse and creative outputs. The temperature parameter value is given within the rande of 0.1 to 1.0. [80]

Without Unstructured Knowledge

PROMPT TO THE LLM:

You're a taxonomist specialising in interior design.
Generate a taxonomy of interior design concepts for the use case of analysing the interiors of a bedroom.
Represent the taxonomy in JSON

RESPONSE FROM THE LLM:

```
{
  "Layout": {
    "BedPlacement": ["Center", "Corner", "AgainstWall"],
    "FurnitureArrangement": ["Symmetrical", "Asymmetrical"],
    "TrafficFlow": ["FreeFlowing", "Restricted"]
  },
  "Furniture": {
    "Bed": {
      "Types": ["Platform", "Canopy", "Murphy", "Bunk", "FourPoster"],
      "Sizes": ["Twin", "Full", "Queen", "King", "CaliforniaKing"]
    },
    "Storage": {
      "Types": ["Wardrobe", "ChestOfDrawers", "Nightstand", "Dresser"],
      "Styles": ["BuiltIn", "Freestanding"]
    },
    "Seating": {
      "Types": ["Armchair", "Ottoman", "Bench"],
      "Placement": ["NearBed", "WindowSide", "Corner"]
    }
  },
```

[redacted]

Figure 4.2: Example Taxonomy generation trial - without including the proprietary IKEA documents

With Unstructured Knowledge

PROMPT TO THE LLM:

<Proprietary documents>

Based on the above information, identify concepts or categories of observations that an interior designer would do of a space? These categories should be those that can be infered from the image of a bedroom without interacting with the inhabitants of the room.
Organise it in JSON format in the form of taxonomy. Also give potential values these concepts/attributes could take.
In the format of {Concept : [possible values]}
Use nested keys wherever necessary.

RESPONSE FROM THE LLM:

```
{
 "Bedroom Design Elements": {

   "Spatial Layout": {
     "Dimensions": ["Length", "Width", "Height"],
     "Zoning": ["Sleeping Area", "Work Area", "Relaxation Area"],
     "Circulation": ["Entry and Exit Points", "Pathways"]
   },
   "Furniture": {
     "Essentials": ["Bed", "Nightstands", "Wardrobe", "Dresser"],
     "Additional": ["Desk", "Chair", "Bookshelves", "Seating Area Furniture"],
     "Customization Options": ["Materials", "Colors", "Sizes"]
   },
```

[redacted]

```
}
```

Figure 4.3: Example Taxonomy generation trial - using the proprietary IKEA documents as reference data”

With LLM-Summarised Unstructured Knowledge

PROMPT TO THE LLM:

<Summaries of the documents - generated by the LLM>

Based on the above information, identify concepts or attributes that interior designers observe to study a bedroom? These attributes should be those that can be infered from the image of a bedroom without interacting with the inhabitants of the room.
Organise it in JSON format in the form of taxonomy. Also give potential values these concepts/attributes could take.
In the format of {Concept : [possible values]}
Use nested keys wherever necessary.

RESPONSE FROM THE LLM:

```
"BedroomAttributes": {
"ColorScheme": {
  "WallColor": ["White", "Beige", "Gray", "Blue", "Green", "Yellow", "Red"],
  "AccentColor": ["Black", "Silver", "Gold", "Copper", "Bronze", "Bright Colors"],
  "FurnitureColor": ["Natural Wood", "White", "Black", "Brown", "Colorful"]
},
"Furniture": {
  "Style": ["Modern", "Contemporary", "Traditional", "Rustic", "Industrial", "Mid-Century Modern"],
  "Material": ["Wood", "Metal", "Glass", "Leather", "Fabric", "Plastic"],
  "Arrangement": ["Centered", "Against Wall", "Symmetrical", "Asymmetrical"]
},
"Flooring": {
  "Material": ["Hardwood", "Tile", "Carpet", "Laminate", "Vinyl", "Concrete"],
  "Pattern": ["Solid", "Striped", "Geometric", "Floral", "Abstract"]
},
```

[redacted]

Figure 4.4: Example Taxonomy generation trial - using summaries of the IKEA documents as reference data

Multi-stage Chain-of-thought (COT)

SUMMARISED VERSION OF THE COT PROMPT TO THE LLM:

Stage 1:
<Proprietary Documents>
Given the unstructured knowledge, recall your knowledge about analysing all things interior design of a bedroom. Output your knowledge in free-associative paragraphs.

Stage 2:
<Output of Stage 1>
1. Read the content given and recollect your knowledge about interior design.
2. Identify all "things" within interior design for the specific use case of bedroom analysis.
3. Come up with consistent logic for grouping and/or subdividing these "things"
4. Group and subdivide these based on the logic.
5. Make revisions.
6. Output your thinking of the above steps.

Stage 3 :
<Output of Stage 2>
Above is information to generate a taxonomy for interior design specifically for bedroom analysis.
Generate a taxonomy. Ensure taxonomy naming conventions are followed.

RESPONSE FROM THE LLM:

```
- BedroomInteriorDesign
 - BedFrames
        - Types : Platform, Canopy, Sleigh, Murphy, Bunk
        - Materials : Wood, Metal, Upholstered
        - Features : WithStorage, WithoutStorage
    - Mattresses
        - Types : Innerspring, MemoryFoam, Latex, Hybrid, Airbed
        - FirmnessLevels : Soft, Medium, Firm
        - Sizes : Twin, Full, Queen, King, CaliforniaKing
    - StorageSolutions
        - UnderBedStorage : Drawers, Bins
        - HeadboardStorage : Shelves, Cabinets
        - ModularSystems : Shelves, Baskets, Rails
        - Closets : WalkIn, SlidingDoor, Wardrobe
        - Other : Nightstands, Dressers
```

[redacted]

Figure 4.5: Example Taxonomy generation trial - Chain-of-thought designed using expert insights. (Full prompt is given in Figure 4.7)

Taxonomy-COT Prompt Design

This section elaborates how the *chain-of-thought* prompt was created with the help of taxonomy expert.

Interviewing Taxonomy Expert:

An interview was conducted with a Knowledge Engineering specialist within IKEA, who manually builds knowledge structures like taxonomy, ontology, knowledge graph, etc. The interview broadly tried to capture their taxonomy process with a few unstructured questions. The interview was preceded by a detailed explanation of the goal of the research. A detailed information sheet (See Appendix B.1) was handed to them and verbal consent was received for their participation and the usage of their responses anonymously.

The questions for this interview were designed with the aim to understand the process of creating taxonomies in detail (See Appendix B.2 for interview guide). Key questions explored the role and perspective of a taxonomist, and initial steps and criteria for understanding and categorizing concepts within a domain. The interview also covered criteria and methods for validating taxonomies, and the importance of collaboration with domain experts and stakeholders.

Developing a 'chain-of-thought' for taxonomy building:

Unlike traditional analysis methods, this interview was analysed with a targeted approach focused on extracting actionable insights specifically relevant to taxonomy generation. The primary objective was to identify instructions, guidance, or steps that could be directly translated into effective prompt design for AI-driven taxonomy creation. Hence, a targeted analysis was performed on the interview by first semantically coding it (see Appendix B.3 for the codes), similar to traditional qualitative analysis methods and then inferring them with added focus on the semantic codes that reflected instructions or guidelines for taxonomy generation.

The following comprises key insights from the interview:

- A taxonomist is a "pedantic" person who tries to model a domain.
- Generic steps to develop a taxonomy:
  1. Map out the domain with edge-cases.
  2. Identify all "things" (concepts) within the domain.
  3. Establish a logic for categorisation and sub-classification.
  4. Categorise and sub-classify based on the logic.
  5. Critically analyse and revise for coverage and mutual exclusivity.
- The key criteria for a good taxonomy is "necessary and sufficient" i.e., there are no overlaps among the concepts at the same time the mapped concepts sufficiently cover the domain for the use-case.
- Involve stakeholders "as much as you can and then some".

A set of steps for taxonomy generation emerged as an outcome of this targeted analysis. Figure 4.6 shows an overview of the steps identified.

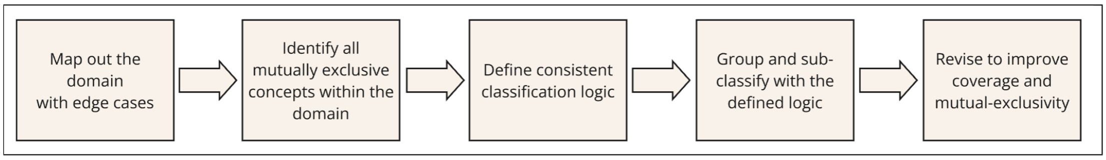


Figure 4.6: Steps for building a taxonomy derived from the interview with expert

Drafting a multi-stage prompt with *chain-of-thought* insights:

The steps identified and the terminology from the interviews informed the development of a *chain-of-thought* prompt specifically designed to create a taxonomy for bedroom interior design analysis. See Figure 4.7.The prompt was structured in a multi-stage format. First, the LLM was instructed to recollect relevant knowledge, enriched by the unstructured IKEA documents, and compile a knowledge pool to facilitate easier concept extraction. Next, the AI was asked to explain how it would apply the defined steps of taxonomy building. This step aimed to gather comprehensive information that the AI could then use to construct and finalize the taxonomy tailored to this specific use case. The final stage used this information to present a finished taxonomy.
This prompt was implemented to derive a taxonomy for bedroom interior analysis with GPT-4. A sample output has been shown in Figure 4.5.

STAGE 1: Knowledge resurfacing enhanced by IKEA documents

You are an expert in interior design and taxonomy construction. You are extremely **pedantic**.

The following is a knowledge source about interior design that you must base your answer on:

<Proprietary Knowledge Sources>

Recall your knowledge about interior design. What knowledge can you gather regarding how a bedroom is analysed for interior design insights. Use the above text as reference.
Your output should be only your train of thought, written in unstructured, free-associative paragraphs.

STAGE 2: Step-by-step detailing of the specific taxonomy to be generated

You are an expert in interior design and taxonomy construction for the purpose of bedroom interiors analysis. You are extremely **pedantic**.

Perform the following tasks:

1. Read the content below, and recollect your general interior design knowledge and determine **the domain and its edges.** In other words, figure out what concepts are included in this domain for the specific use-case of bedroom analysis, and what are not. Also, understand the edge-cases to **stress-test the borders** of the domain that you establish, to confirm or adjust them.

2. Identify all the possible **"things"** in this domain. What are ALL the possible, mutually-exclusive concepts that exist within the domain you've defined, and how can they be termed in a way that they are distinguishable, understandable and mutually-exclusive from a semantic perspective.

3. Come up with **consistent rules/logic** for how you will group/subdivide the **"things"** into classes all the things you've identified in this domain. The logic of your approach should make sense for the nature of this particular usecase.

4. Following the logic/rules you've decided on, group these "things" into classes and subclasses. Use nesting wherever necessary. Then, subdivide each class and sub-class **further into subclasses of that. And then subclasses of that. And so on**, until you have reach a taxonomy that has mutually exclusive concepts and sufficiently covers the domain of interior design specifically for bedroom interiors analysis.

5.Critically analyze the organisation of the concepts you've identified. **Make revisions** to enhance clarity, coverage, and practical utility.

6. Output your thinking for each of these steps.

CONTENT:
<Stage 1 response>

STAGE 3: Extracting the final taxonomy for interior design analysis

You are an expert in interior design and taxonomy construction. You are extremely **pedantic**.
Below you will find the information to generate a taxonomy for bedroom interiors analysis.
Generate a taxonomy with categories/concepts relevent specifically for the use case. Ensure the naming of the classes and all the nested subclasses are according to the naming conventions of taxonomies.

CONTENT:
<Stage 2 response>

Figure 4.7: Multi-stage, Chain-of-thought prompt designed using insights given by taxonomy expert. The words/phrases highlighted in bold text were directly taken from the interview i.e., expert terminology.

Consultation & Analysis of resulting Taxonomies :

As a HITL practice, the most distinct taxonomies from the prompt trials were collected and analysed in collaboration with a taxonomy expert within IKEA, who qualitatively assessed the validity of the outputs as a sufficient interior design taxonomy. The following are some observations:

- Change in the level of detail of input knowledge did not significantly enhance the taxonomy's detail. For example, taxonomies generated from summaries of the documents were similar to those produced from the entire document set. A plausible explanation for this could be that the source documents did not contain sufficiently granular information to warrant a more detailed taxonomy.
- The impact of the unstructured documents was more prominent with the COT prompt trial. Unlike other trials, the outputs generated with COT frequently omitted 'Furniture,' a key super-category for this use case, according to the taxonomy experts. Instead, it prioritized categories such as 'Mattresses' and 'Bedframes' at the top level, which are more appropriately considered sub-categories. This observation could be linked to the fact that IKEA's *How-To* articles [78] (one of the unstructured document used) had many product suggestions related to mattresses and bedrooms. While this suggests that GPT-4 used the unstructured document more with the COT prompt, it highlights that the unstructured documents do not sufficiently align with the type of content ideally that significantly enhance the pre-trained domain knowledge of the model to develop an interior design taxonomy.
- Expert's observations about the categorisation and organisation of concepts:
  - Overlapping concepts diminishing mutual exclusivity. Example: [redacted] similar to those of *Personalisation*.
  - Inconsistent/incorrect categorisation logic. Examples: [redacted]
  - Ambiguity in categorisation and sample values. [redacted]
  - Poor nomenclature. [redacted], [redacted]

Overall, the general opinion of the taxonomy expert towards the LLM's outputs was that although it seems to identify various important aspects of interior design, it lacks in efficient categorisation and organisation of the taxonomy.

These trials give us an understanding of the AI's capability to generate a taxonomy for interior design in one go with prompting (Discussed further in the reflections section below).

Although this step of the stage was to efficiently define a *conceptual framework* for the annotations leveraging the LLM's capability to process unstructured corpus of text, it was pursued with a focus on taxonomy generation. This was due to the potential of such an experiment(if successful) to significantly enhance knowledge engineering processes beyond the domain of interior design. LLM-driven taxonomy generation is an untapped field of research to which this sub-exploration contributes the following key insights:
- GPT-4 is able to identify key concepts of interior design but falls short in organising them to form a gap-free non-overlapping taxonomy. This opens up an opportunity for future research for an iterative expert-supervised taxonomy generation process with the LLM identifying key concepts from large corpus of texts and the experts tweaking the organisation of these concepts. This also highlights the inherent iterative nature of taxonomy building.
- Future iterations should also consider using more detailed and contextually relevant unstructured documents to see if they contribute more effectively to the taxonomy generation process.
This exploration was not pursued further because the CF for the use-case of this study only required a list of key aspects that sufficiently encompasses interior design information of a bedroom. The LLM was in fact able to identify several key aspects for the CF.

Creating a list of key concepts from the results of the LLM :

Nonetheless, the concepts of interior design analyzed through these trials were valuable in identifying the essential aspects of interior design. From the taxonomies generated by the LLM, the most prevalent and distinct top-level categories of interior design were selected. These categories formed the the key topics that needed to be addressed in the annotations.

These key aspects were compiled into a list (See Figure 4.8). Even though the examples provided by the LLM in the taxonomies clarified what exactly these categories were, to further avoid ambiguity, the LLM was asked to define these categories.

This list formed the *conceptual framework* for the annotations, to ensure that the model consistently considered the necessary aspects of interior design.

However, it was crucial to validate this list with domain experts. This validation was conducted through two methods: 1. A workshop where domain experts analyzed a room, and the implicit categories they used were extracted and compared with the above list from the LLM; 2. By direct consultation with experts to validate if the list sufficiently covered all the key aspects given the use-case.

It must be noted that deciding on a list of key aspects as the conceptual framework for the annotations, i.e., the information the annotations must comprise of , was a strategic decision based on what is necessary for the use-case within IKEA. A list of aspects (a one-level taxonomy) was a more streamlined and less convoluted way to collect the data which would be flexible for multiple use-cases within IKEA, some of which might require simple or minimal annotations (for example, creating simple image metadata). Moreover, a larger taxonomy would have required more time and effort from a limited number of experts.
Hence defining the CF for an AI's image annotation task is therefore use-case specific. Multiple factors such as the processing method of these insights, their applicability, and the cost of maintenance must be considered while defining the framework for the annotations.

4.1.2 Deriving key concepts from the minds of human experts :

Workshop with Interior Designers:

In addition to textual knowledge sources, the expertise and insights of interior design professionals are other sources for understanding the domain of interior design. Instead of directly asking the experts to jot down a set of interior design concepts, a more effective approach was to have them perform the same task as the AI i.e., analyzing the interiors of spaces using images of rooms. This method was deemed to provide rich insights for both defining the necessary CF for the annotations as well as informing the prompt design for the VLM in the next stage (which will be elaborated upon in a later section).

To this end, a workshop was conducted with 5 interior design experts from IKEA. The experts were presented with an image of a room and were asked to drop sticky notes of what they notice about the room as an interior designer. Post this activity, a small discussion session was conducted where the experts elaborate on their analysis points. This was useful to clear out any ambiguity in the sticky notes. The experts performed this task for 3 different images of bedrooms. See Appendix D.1 for a snippet from the infinite canvas board used in the workshop.

It must be noted that, similar to the interview with taxonomists, an information sheet was provided prior to the workshop to introduce the research, their consent for participation was received, and permission for recording the session was obtained.

Workshop Analysis for Key Interior Design Aspects

Since the goal of the workshop (for this stage) was to extract key aspects of interior design that the experts implicitly analysed the interiors of a room image with, the workshop artifacts (Sticky notes containing insights

[redacted]

2. Furniture
    Definition: The movable objects intended to support various activities within the room.
    Examples: Upholstered bed with storage, Wooden desk, Metal chair.

3. Style
    Definition: The aesthetic and design theme of the room.
    Examples: Modern, Minimalist, Rustic, Traditional, Eclectic.

4. Colours
    1. **Definition:** The specific colours and colour palette seen in the space and a description of the tone, color-theme.
    2. **Examples:** Neutral tones, Earthy hues, Bright Orange Accents

[redacted]

9. Storage
    Definition: Solutions and systems for organizing and storing items within the room.
    Examples: Modular storage units, Built-in wardrobe, Storage boxes under the bed.

[redacted]

Figure 4.8: List of Key Aspects of Interior Design Consolidated from the LLM Driven Taxonomies

about the room images) were synthesized to form common themes/categories. The following are the aspects of interior design that emerged from the workshop analysis (Figure 4.9).

Here one can note that there is significant overlap in the set of aspects that was derived using the LLM in the previous section. The aspects in red can also be found in the list obtained from the LLM. From these lists we can already see a conceptual framework for the annotations forming. We can note that with both the results of the LLM and that of the workshop, the annotations must contain information about a set of key

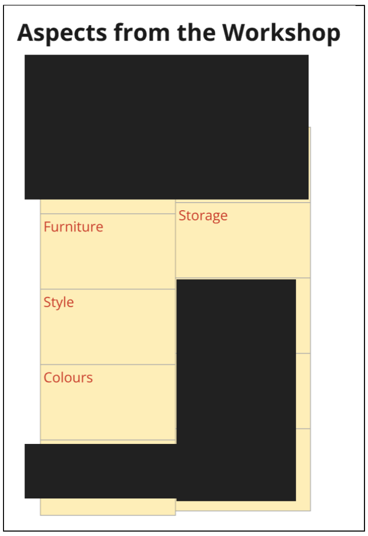


Figure 4.9: List of Key Aspects of Interior Design Consolidated from the Workshop

aspects of interior design.

4.1.3 Consolidating the Lists and Consultation with experts :

To finalise the conceptual framework for the annotations, the results from the LLM taxonomy experiment and the workshop were combined into one list of aspects of interior design. Figure... A follow-up consultation session was conducted with the interior designers to ensure that the consolidated list covered all aspects of interior design.

During this consultation the interior designers were informed of the exact task the AI would be performing and they were presented with the set of key aspects of interior design insights gathered from both the LLM processing IKEA proprietary documents and the workshop.

They were asked if these aspects encompass everything interior design about the room and if they would add modify or remove any aspect.

Upon consultation, the following changes were made to the key aspects :

- [redacted] is a subjective aspect that cannot be understood by simply looking at the image. [redacted] Hence, it was removed from the list.
- [redacted] is more relevant to the design of each article in the room and less to the overall interior design of the room. Hence, it was removed from the list.
- [redacted] has a lot of overlap with [redacted] and [redacted] Everything that falls under [redacted] fall under these two aspects as well. Hence, to avoid redundancy, this aspect was removed.
- It can be hard to distinctly categorise articles in a room into either [redacted] for example, [redacted] It was identified that materials as an aspect can encompass both hard textures as well as cloth textures in a space. hence [redacted] were combined into on single aspect -[redacted] and the definition was modified accordingly.
- [redacted] was an aspect that did not appear in the workshop but was in the taxonomies produced by GPT-4. Hence, the experts were specifically consulted regarding this category to know its relevance and they agreed that this would be an important aspect to analyse the interiors of a room for. Hence, the aspect of [redacted]

Upon applying the above modifications to the consolidated list, a final list of 10 key aspects was obtained that would be form the *conceptual framework* that the VLM will follow to generate the annotations. Figure

4.10 shows the final set of aspects along with definitions and example. This was then used as a guide to generate the domain-specific annotations of room images using a VLM which is elaborated in the next section.

Final Set of Aspects for the Annotations

[redacted]

3. Furniture
    Definition: The movable objects intended to support various activities within the room.
    Examples: Upholstered bed with storage, Wooden desk, Metal chair.

4. Style
    Definition: The aesthetic and design theme of the room.
    Examples: Modern, Minimalist, Rustic, Traditional, Eclectic.

5. Colours
    1. **Definition:** The specific colours and colour palette seen in the space and a description of the tone, color-theme.
    2. **Examples:** Neutral tones, Earthy hues, Bright Orange Accents

[redacted]

9. Storage
    Definition: Solutions and systems for organizing and storing items within the room.
    Examples: Modular storage units, Built-in wardrobe, Storage boxes under the bed.

[redacted]

Figure 4.10: The final set of aspects that form the conceptual framework for the image annotations

Stage A Summary

This section explains how the conceptual framework for the annotations was defined, forming the first stage of the methodology. Initially, an LLM (GPT-4) was leveraged through prompt engineering to derive a key set of interior design concepts. Unstructured proprietary documents were fed to the LLM to potentially enhance its pre-trained domain knowledge of the LLM. This was conducted as a sub-exploration to tap into LLM-driven taxonomy generation. The findings from this exploration were discussed, and the most recurring concepts were compiled to form the conceptual framework for the annotations. This conceptual framework essentially comprised a list of interior design aspects.
These aspects were then validated with domain experts through two methods: First, by analyzing a workshop where domain experts examined a room, and then, through direct consultation with experts to ensure the list sufficiently covered all key aspects of interior design for the given use-case of image analysis. Following the consultation, a list of 10 interior design aspects was finalized to generate standardized image annotations in the next stage. The list also included definitions and examples generated by the LLM to avoid ambiguity.

Recap of the Sub-Research Questions

SQ 1.1: How can a CF be derived for the annotations to ensure all necessary information is included in them?

By first gathering key interior design aspects from LLM generated taxonomies for the domain. Although the taxonomies fell short in organisation and mutual-exclusivity the concepts identified can be useful in forming a foundation set of key aspects for the CF. These concepts can then be validated by synthesizing an image analysis workshop with interior design experts, followed by a direct consultation with them. This way a CF containing 10 key aspects of interior design was defined.

SQ 1.2: How can prompts for a VLM be effectively designed to generate interior design image annotations?

The CF has been obtained which is the input for next Stage. The upcoming section (Stage B) explores how this question was addressed.

SQ 1.3: How can the validity of the generated annotations be checked?

Not yet addressed.

Auxiliary SQ 1.1: How can a *human-in-the-loop* approach be effectively integrated throughout the process?

The prompt design and conceptual framework was closely informed by domain expertise through multiple touch points in the form of interviews, workshop and consultations.

4.2 Stage B - Generating Annotations with VLM Answering SQ 1.2

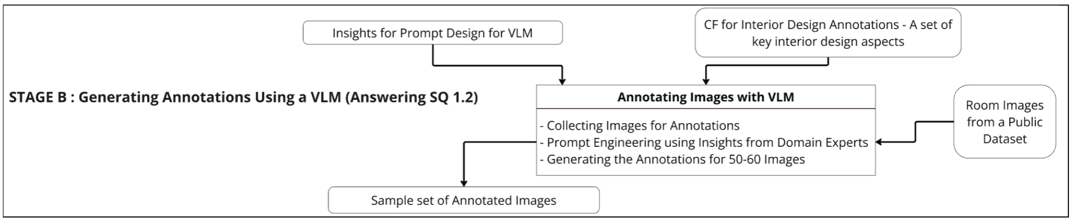


Figure 4.11: Snippet of Stage B from the Overall Methodology shown in Figure 3.1

With the CF, the interior design aspects that the annotations must contain insights about were defined. GPT-4V, specifically *GPT-4 Vision Preview*, was then used to generate interior design insights for room images thereby generating image annotations. To recall, Figure 4.11 shows the snippet of this stage B from the Overall methodology design chart shown in Figure 3.1 This section will elaborate how this stage was implemented.

4.2.1 Gathering Images for Annotation Generation:

A subset of approximately 80 images were taken (random selection) from the LSUN-Bedrooms Images dataset [1]. This is a publicly available dataset that comprises images of bedrooms of real people. Out of the subset, 20 images were for the prompt design phase, for testing through trial-and-error and 60 images for creating a sample set of image-annotation pairs for validation with domain experts)

4.2.2 Initial Prompt Attempts :

Multiple initial attempts with different prompting strategies were made to instruct the VLM to generate interior design annotations for the set of 10 key ID aspects. For example, an image of a room and the list of aspects along with definitions and example values [4] was passed to the AI, instructing it to fill the list with corresponding interior design insights (refer to the Figure 4.12). In another attempt, the definitions were converted into questions such that the answer is interior design insights for the specific aspect (refer to the Figure 4.13). In this attempt with questioning, the response of the VLM was descriptive despite adding examples of the values. Perhaps because the prompt did not specifically mention to follow the format of the examples. Hence, another prompt was used as a second stage to convert the descriptive answer into set of terms that capture the key insights (shown in Figure 4.13).

Moreover multiple minor tweaks were made to the prompts by following tips from the guidelines given by Bsharat et al. [8] to see if they make the responses more detailed, articulated. Overall it was observed that there were no obvious differences between the results generated with these prompting trials. This is probably because the task the AI was performing was ultimately the same, i.e., to generate key insights for the room in short phrases. The insights generated were mostly 2-3 single terms for each aspect. Even if the AI took different paths to get to the output for different prompting styles, it was not reflected in the results as can be seen by comparing the results shown in the two Figures 4.12 and 4.13).

An attempt was also made to use the documents used in the previous stage (Refer Section 4.1 for details of the documents) as contextual knowledge for the VLM. However, the results were not significantly different from the other attempts. This lack of enhancement could have been due to the documents not containing much details about the variety of objects and aesthetics that could be found in different rooms. Hence, perhaps there was little for the model to use from these documents.

[4] These were the definitions and example values generated by the LLM in the previous stage

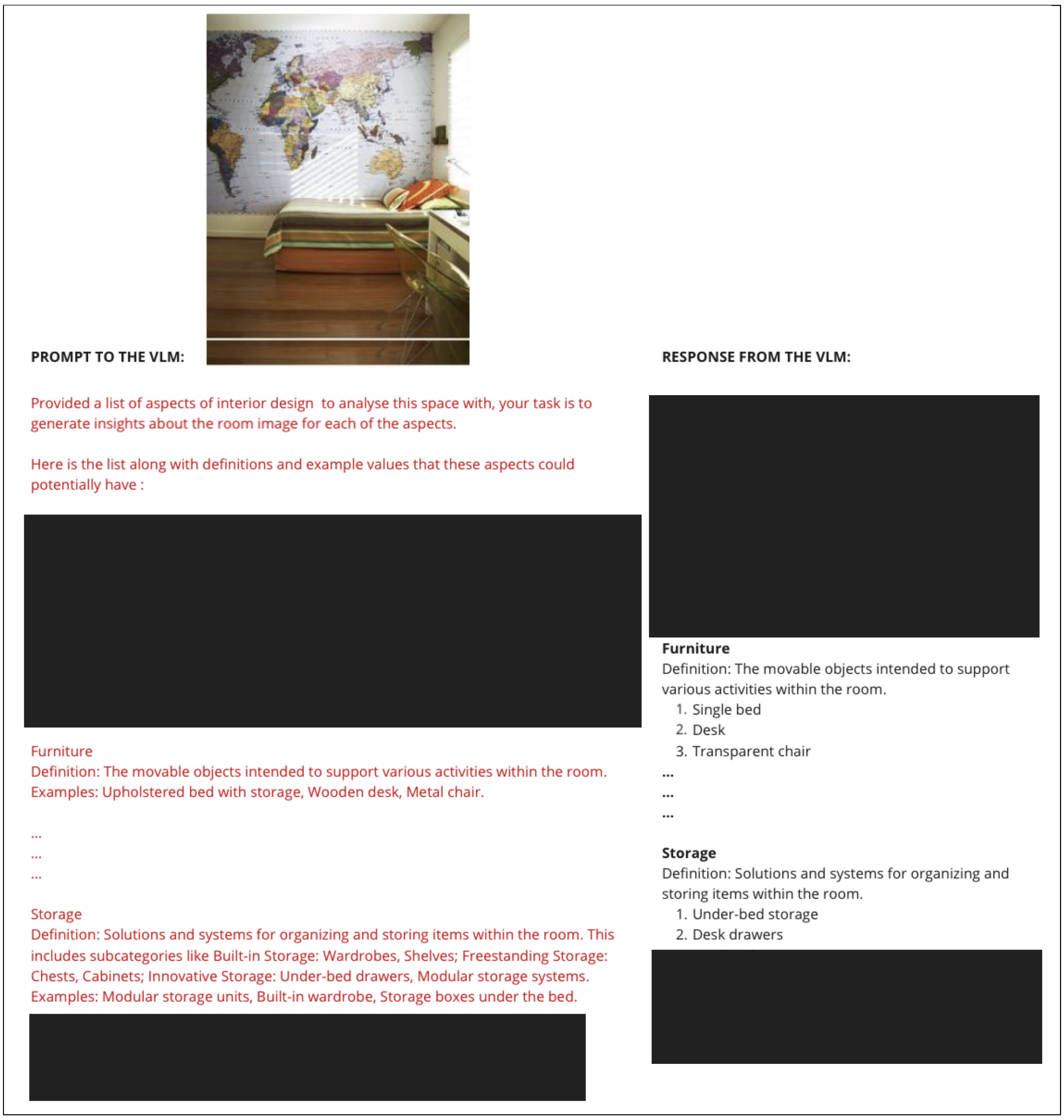


Figure 4.12: An initial attempt to prompt the VLM by directly instructing it to generate insights for all the interior design aspects defined. (Full prompts and outputs can be found in Appendix C.1.1)

4.2.3 Multi-Stage Prompt:

In another attempt, a two-stage prompt approach was developed and tested. This was inspired by the initial *question-answering style* prompts, where the VLM provided descriptive analysis on different aspects before condensing them into key points.

First Stage:

In the first stage, GPT-4V was asked to describe the image based on specific interior design aspects from the conceptual framework. Each time the prompt focused on only 2-3 aspects to ensure the VLM could provide a more detailed analysis. Although *GPT-4 Vision Preview* has a response token limit of 4096 tokens [24] (approximately 3500 words), it typically generated only around 800 to 1000 words in the other initial attempts. Hence limiting the number of aspects per prompt allowed the VLM to allocate more response tokens to each aspect, leading to more thorough and specific descriptions. Figure 4.14 shows two examples from this attempt.

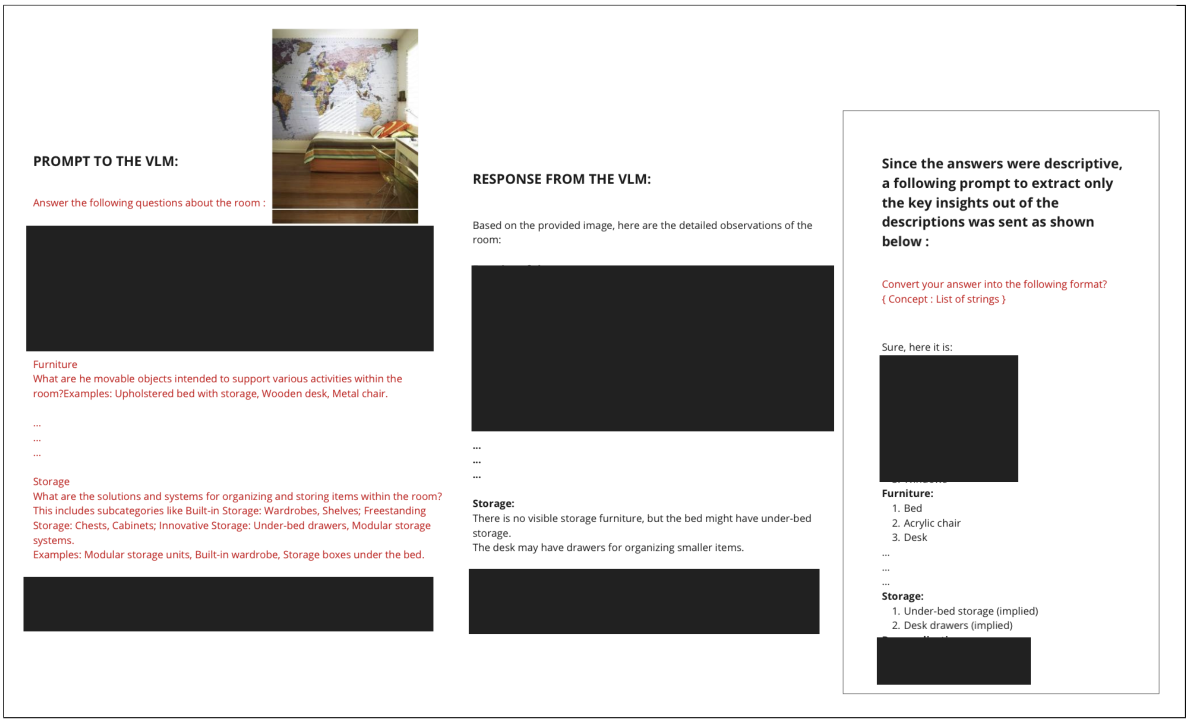


Figure 4.13: An initial attempt to prompt the VLM by forming questions for all the interior design aspects defined. (Full prompts and outputs can be found in Appendix C.1.2)

Second Stage:

In the second stage these descriptions were instructed to be converted into key insights in the form of short phrases (See Stage 2 in Figure 4.14). Additionally, to push the AI for more detailed insights, a string (word) count was specified, instructing the VLM to generate lists of 1-5 word strings for each aspect.

> One might wonder if these descriptions from the first stage could be directly as insights. Descriptive, verbose insights for 10 different aspects of interior design would mean a bulky text accompanying the bedroom image, which would make it less efficient in for further processing. Condensing the information into "bullet-points-like" insights would make the annotations more applicable for diverse use-cases and also facilitate easier evaluation of results. Moreover, the LLM itself in the previous stage demonstrated this preference by providing single terms and phrases as example values for the aspects.

This two-stage prompting approach was tested on 10 images, with aspects decomposed into groups of 2 or 3. Figure 4.14 shows two instances of the results of these two stages. The model performed well in capturing the room's aesthetics descriptively. Even with manual reviews by the researcher, who is not a domain expert, it was evident that the descriptive analyses were clear and comprehensive. In the second stage, although still in short phrases and rarely reaching the 5-word maximum, the insights were notably more descriptive and detailed compared to previous attempts. [redacted] This improvement could be attributed to both the decomposition of the task into fewer aspects and the more detailed analysis in the initial stage enhancing the AI's *chain-of-thought*.

This 2-stage prompting approach gave the best outputs, in the sense of details and comprehensiveness, among all attempts. However, there were still some ambiguous insights, such as "minimal approach" in color. To further improve and define the final prompts to generate the synthetic annotations with, the next step involved incorporating domain expertise to enhance the prompts and address edge cases.

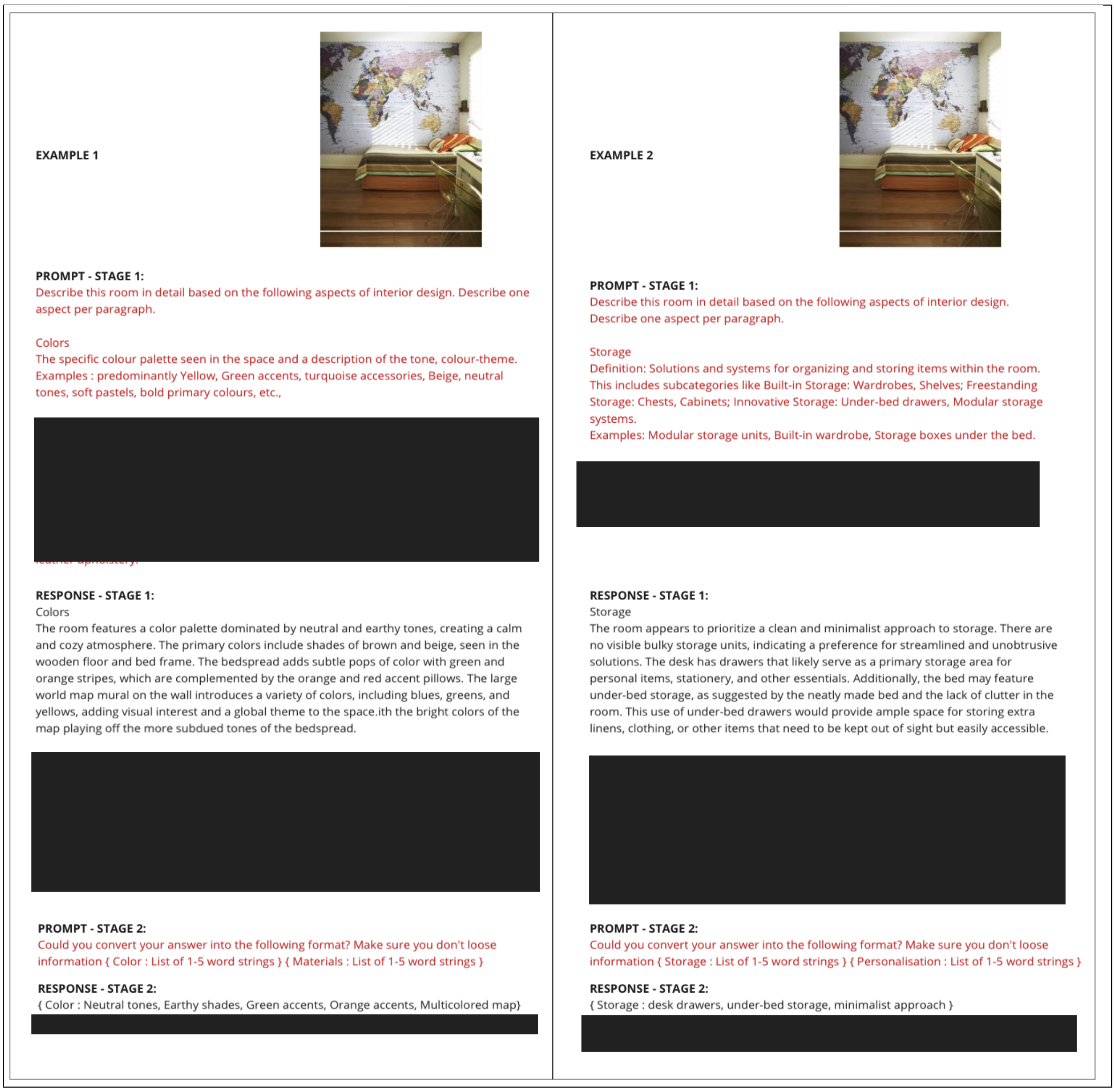


Figure 4.14: Two examples from the multi-stage prompting attempts

4.2.4 Insights for prompts design from workshop with the interior designers:

The workshop, in which interior designers analyzed and described images of rooms, proved useful in developing the conceptual framework for the annotations, as described in previous section (Subsection 4.1.2. This exercise directly aligns with the task that the AI had to perform, i.e., generating interior design insights by analysing an input image of a room. The significant overlap between the aspects generated by the LLM and those from the workshop indicated that the workshop contained valuable information about how designers analyze images based on these aspects. Therefore, the workshop was analyzed to extract guidelines that could inform the 2-stage prompts. A content analysis[5] [81] was performed on the workshop data.

The insights generated by the experts were grouped based on the aspects in the CF. These groups were then reviewed to infer guidelines the designers used when analyzing the images. Inferential summaries were generated for every group, which could serve as a basis to enhance the prompts for the respective aspects. This

[5]Content analysis is a research technique for making systematic, credible, valid, and replicable inferences from texts or other media [81]. There is not a specific set of steps for this technique, but most involve coding pieces of data to facilitate categorization, then grouping them and making inferences by finding patterns in these groups.

method ensured that the prompts incorporated expert-level insights and analysis patterns used by professional interior designer experts. The inferential summaries for the groups of *Storage* and *Colour* are shown in Figures 4.15 and 4.16

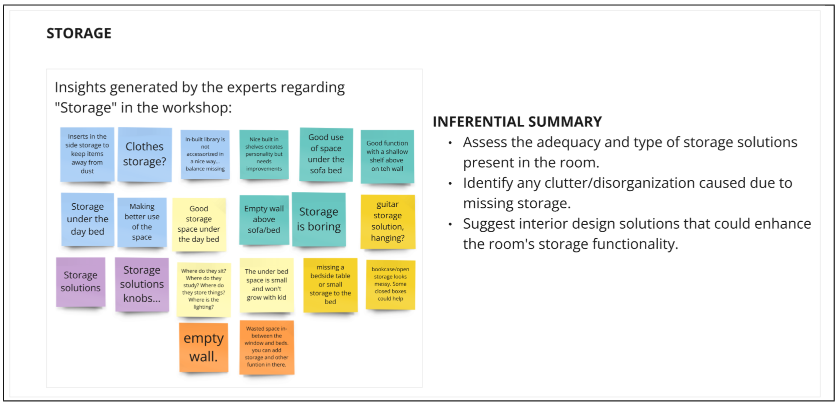


Figure 4.15: The content analysis of the workshop with interior designers for the aspect - *Storage*

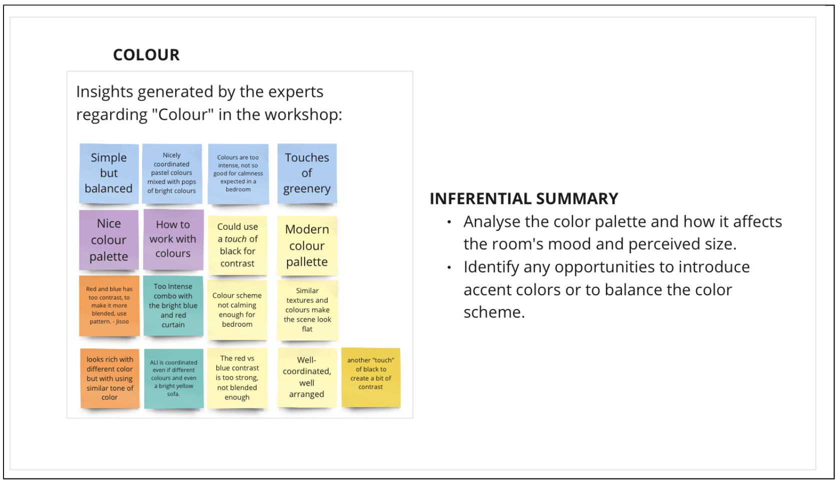


Figure 4.16: The content analysis of the workshop with interior designers for the aspect - *Colour*

From the inferential summaries, it was noted that experts rarely stated things as they were; instead, they evaluated the room, focusing on how various aspects could be improved rather than merely noting their existence. To replicate this evaluative approach of interior designers, prompts included instructions about assessing the space for the aspects and providing improvement suggestions. However, it was identified that not all aspects require such improvement insights. For instance, providing improvements in [redacted] is impractical given the scope of the domain.

[redacted] was one aspect that the interior design experts did not directly address during the workshop [6]. Since there were no direct insights for this aspect, a cumulative understanding of the inferential summary along with the definition helped in designing the prompt for [redacted] For example, it was inferred that the interior designer predominantly looked with a critical perspective and often suggested improvements. This insight coupled with the definition of the aspect helped develop a full prompt explaining how the [redacted] aspect must be analysed.

[6] However, as discussed in the previous section, the experts agreed that it is an important aspect of interior design during the final consultation, so it was included in the conceptual framework.

4.2.5 Final Prompts design :

Using insights from the analysis of the workshop with interior designers, two-stage prompts were drafted for each interior design aspect of the conceptual framework. Through multiple trials and errors, additional conditions were incorporated into the prompts to minimize errors and improve reliability. The following are the common conditions included in the final prompts:

- Stage 1 - Describing the image based on a specific aspect:
  - The task of analysing the image for a specific interior design aspect was broken down using the inferential summaries analysed from the workshop. This was helpful to enhance the *chain of thought* of the VLM and provide specific topics to address in the description that would reflect the way a human expert would perform the task.
  - Most off-the-shelf VLMs are constrained to provide positive responses. Due to this initial insights generated were as if the rooms needed no improvements, contrary to how human experts generated these insights. To address this, phrases like *"Be critical whenever necessary"* were added to the prompt.
  - To prevent the AI from generating information not present in the image (hallucinations), phrases such as *"Look at every part of the image"* and *"Do not make any inferences beyond what is incredibly clear and obvious"* were included.
  - The VLM was assigned the role of *"interior designer with expertise in <interior design aspect>"* to ensure efficient use of its pre-trained knowledge about the subject.
- Stage 2 - Extract key insights from the descriptions from stage 1:
  - The VLM was instructed to summarize the description from Stage 1 and represent it in a given format of phrases.
  - In cases where the VLM encountered issues, it was instructed to provide an exception message: "Error".

These common conditions were applied to all prompts for each aspect. Each aspect was meticulously tested and improved multiple times with 10 images from the dataset, each time tweaking the phrasing of the instructions. The prompts were finalised once further tweaks did not show any significant improvement in the outputs.

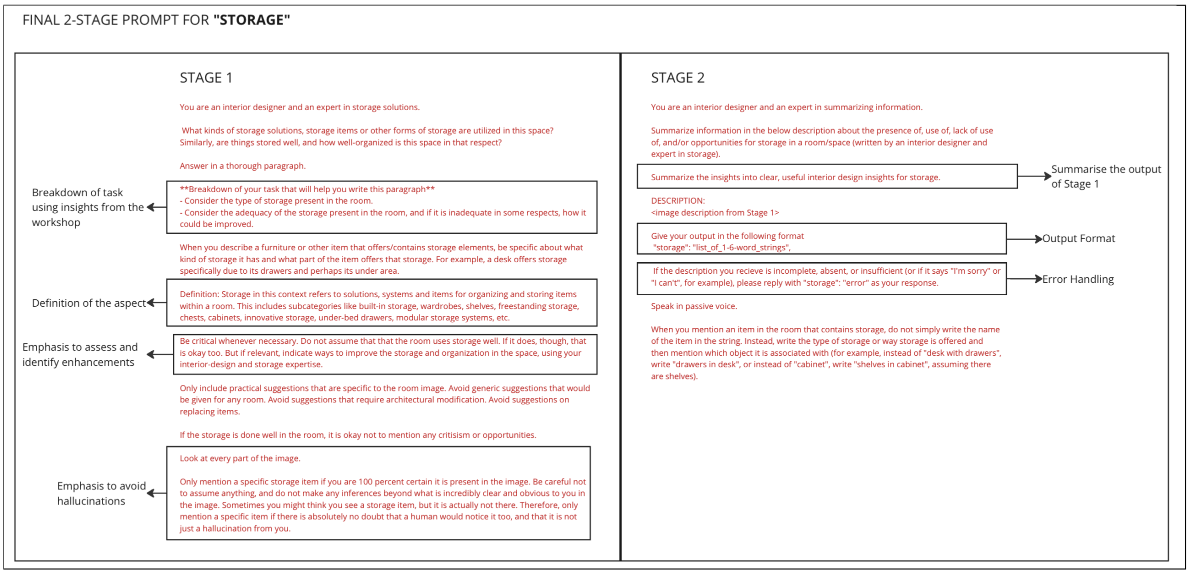


Figure 4.17: The 2-Stage prompt used to generate the insights for the interior design aspect - *Storage*.

The Figure 4.17 illustrates the final prompt used for the aspect *Storage*. Such a prompt was developed for all the 10 aspects. Due to the multiple tweaks and conditions, the final prompts were much more detailed than the initial attempts. Moreover, as the prompts for each aspects are different, the task was decomposed to generating insights for only 1 aspect per prompt, unlike previous attempts where prompts were defined for groups of 2 or 3 aspects.

4.2.6 Generate a sample set of annotations for 60 images:

Keeping the availability of experts for evaluation in mind, The finalized prompts were repeatedly run for 60 room images from the dataset. Although a dataset of 60 images and corresponding annotations is small when compared to the scale at which this capability could eventually be used, it provides a starting point to assess and analyse the off-the-shelf VLM's performance. The 2-stage prompt for each aspect was applied to all 60 images, resulting in a dataset comprising 60 images with corresponding outputs for the 10 aspects of ID. An example of annotation data for one bedroom image can be found in Appendix E.1.

Stage B Summary

This section elaborates how the second stage of the methodology was realized to generate synthetic image annotations of interior design insights. After multiple trials, a multi-stage prompt where the VLM first descriptively analyses the image given a specific aspect and this description is then passed back to the model to be converted into "bullet-points" like insights in the second stage. Detailed prompts were drafted for all the 10 aspects within the contextual framework, with this multi-stage approach. These prompts were enhanced by insights from the workshop with interior design experts. These 2-stage prompts were run for 60 images from a public dataset of bedroom images, to form a samples set of images with standardised annotations of interior design insights.

Recap of the Sub-Research Questions

SQ 1.1: How can a CF be derived for the annotations to ensure all necessary information is included in them?

By first gathering key interior design aspects from LLM generated taxonomies for the domain. Although the taxonomies fell short in organisation and mutual-exclusivity the concepts identified can be useful in forming a foundation set of key aspects for the CF. These concepts can then be validated by synthesizing an image analysis workshop with interior design experts, followed by a direct consultation with them. This way a CF containing 10 key aspects of interior design was defined.

SQ 1.2: How can prompts for a VLM be effectively designed to generate interior design image annotations?

Through a set of multi-stage prompts focusing on one aspect at a time, enhanced by insights derived from domain-experts' analysis of rooms (workshop).

SQ 1.3: How can the validity of the generated annotations be checked?

Not yet answered. How the sample set of images and corresponding interior design insights were evaluated is explained in the following section.

Auxiliary SQ 1.1: How can a *human-in-the-loop* approach be effectively integrated throughout the process?

Stage A : The prompt design and conceptual framework was closely informed by domain expertise through multiple touch points in the form of interviews, workshop and consultations.
Stage B : Domain expertise captured from the workshop was translated into insights that could be incorporated into prompts.

4.3 Stage C : Checking the Validity of the Annotations Answering SQ 1.3

The annotations generated for the set of images were then checked for their validity with human experts. Figure 4.18 gives a closer look at this stage from the overall methodology chart shown in Figure 3.1. This section elaborates how this stage was implemented.

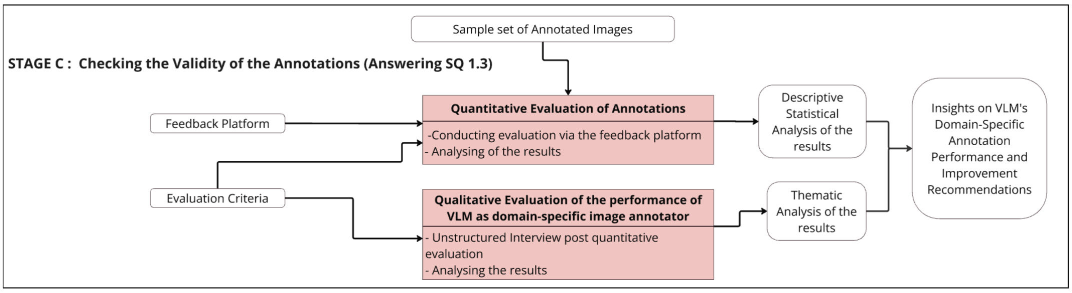


Figure 4.18: Snippet of Stage C from the Overall Methodology shown in Figure 3.1

4.3.1 Defining the Evaluation Criteria:

As discussed in Section 2.3, calculating traditional AI metrics such as accuracy require a ground truth dataset. Due to the absence of a validated dataset for interior design annotations, the annotations had to be directly evaluated by humane experts. Specific criteria that experts could evaluate the annotations on had to be defined.

Out of the most common criteria used by other studies in human evaluation of annotation data, *completeness* and *correctness* [57, 58] inspired the definition of the evaluation criteria for the use-case of this study. These criteria aligned best with the scope of this study as having experts evaluate whether the generated insights are correct and thorough would provide an understanding of the practical applicability of this process for automating interior design analysis in various use cases at IKEA.

It stands to reason that experts might find it easier to pinpoint what is wrong or missing rather than assess the overall correctness or completeness of the annotations, making the evaluation process more straightforward and efficient. Hence, the criteria of *Incorrectness* and *Incompleteness* were defined. Moreover, highlighting specific incorrect and incomplete elements provides more actionable insights for refining and improving the AI model.

Hence, the criteria established for the evaluation are :

Incorrectness: A value in the annotation is considered incorrect if it is clearly and demonstrably false based on the visible content of the image. For example, an annotation mentioning a blue sofa when it is clearly a red sofa, the annotation containing a window where there is none.

Incompleteness: The annotation is deemed incomplete if it misses values that clearly impact the understanding of the room. For example, not mentioning a prominent beach-themed style, omitting a large bed in the center of the room, etc.

4.3.2 Quantitative Evaluation:

Building the platform:

The evaluation required experts to examine each image and identify incorrectness and incompleteness in the corresponding annotation. Given that each annotation consists of interior design insights for 10 different aspects, and these specific insights need to be checked and compared to the image, this task is inherently tedious.

Hence, an evaluation platform was built using Python and the Streamlit[7] library, with special considerations to make the process user-friendly for the evaluators for example, side-by-side image and annotations panels for easy cross-checking of the insights. A screenshot of the platform can be seen in Figure 4.19.

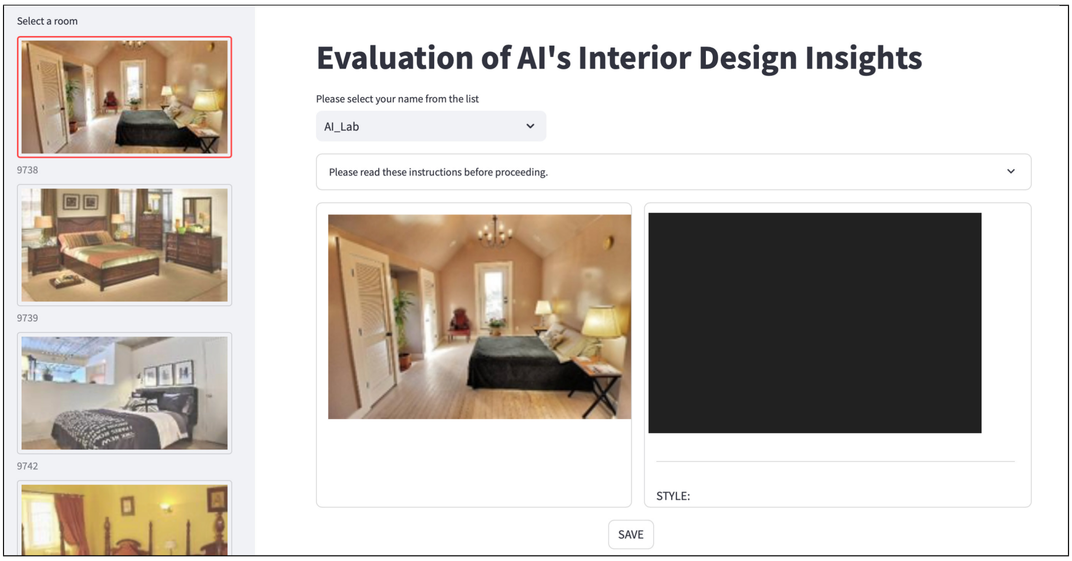


Figure 4.19: Evaluation Platform

Platform Description

- **Expandable Top Panel:** The top-most panel, which can be expanded, contains a brief explanation of the platform and the evaluation criteria.
- **Side Panel:** The side panel displays the list of images available for evaluation.
- **Image and Annotation Display:** When an image is selected, it populates in the bottom left panel, and the corresponding annotations appear in the bottom right panel. The annotations are displayed in distinct sections, with interior design insights for each of the 10 aspects presented separately.
- **Indicating Incorrectness:** Experts can indicate the incorrectness of an insight by turning off the toggle buttons next to each value.
- **Indicating Incompleteness:** Experts can indicate the incompleteness of an aspect by selecting the checkbox at the bottom of each section. Collecting feedback on the incompleteness for each aspect, rather than the entire annotation, was designed to provide a deeper understanding of the AI's performance on the incompleteness criterion.
- **Save Results:** The feedback can be saved by pressing the save button before moving on to the next image.
- **Allocation of data for evaluators:**
  Each evaluator was allocated approximately **30 images** to increase the number of samples evaluated while keeping the workload minimal for the experts. Moreover, **7 images** were the same across different participants to calculate their *inter-rater reliability*[8].

On-boarding and conducting the evaluation:

2 Interior Design Experts were sought to perform the evaluation task. It is worth noting that these experts had over 10 years of Interior Design experience. They were on-boarded onto the platform through an online

[7]Streamlit is a python library used for creating custom, user-friendly web applications from Python scripts using pre-made easy-to-use components.
[8]inter-rater reliability is the degree of agreement among the evaluators

meeting, during which they were explained the task, walked through the platform, and introduced to the evaluation criteria. They evaluated 2 images during the call to familiarize themselves with the platform and clear any doubts.

Since the platform was hosted on a cloud service, the experts could access it at their convenience and perform the evaluations in parts. They completed the evaluation within one week.

> With the proliferating developments that apply gen AI, human feedback and evaluation is becoming increasingly crucial for AI performance validity checking against our expectations and values. While simpler methods such as pen-and-paper sessions with interior designers to evaluate the image annotations could have been employed for the evaluation, it would have been tedious for the experts to evaluate all images in one sitting. Moreover, supervising them during such sessions might have caused them to feel intimidated, especially since this study involves AI, an area outside their expertise.
> As an AI researcher, I strongly believe that evaluation processes, given how frequently they are required, should be considerate towards the evaluators. Ensuring that the process is straightforward and non-tedious for the evaluator should be top priority. This is a subtle but crucial aspect of the HITL concept that is often overlooked. Most evaluations focus on maximising the feedback counts and less on the quality of the experience for the feedback-givers [72]. By allocating extra effort in this research to create a specific feedback platform such that the interior designers can evaluate the sets of images in parts with convenient side-by-side comparisons, I wish to emphasize this subtle aspect of HITL.
>
> The platform was built using the features available in the Streamlit library, within the constraints of the time allocated for its development. The user experience of the platform is an important factor in the quality of inputs. Some suggestions for enhancing it would be through features such as progress tracking, auto-saving feedback, and displaying higher-quality images, etc.
> To minimize the effort required for evaluations, only binary feedback on incompleteness was solicited, without requiring detailed explanations of what was missing. While this streamlined the evaluation process, it also limited the depth of feedback. If feasible, a key recommendation for future iterations would be to allow for more detailed feedback options, even if optional, to capture more specific insights that could be valuable for refining the AI models.

Results :

The experts were able to evaluate a total of **55 image ids**. The evaluations of the both the experts were conjoined and a descriptive statistical analysis was performed. The results of the evaluation for the two criteria - *incorrectness* and *incompleteness* is shown in Table 4.1 and 4.2.

About **9%** of the AI-generated values were found to be evaluated incorrect. About **5%** of the AI-generated aspects were assessed to be incomplete. The distribution of incorrect values and correct values is shown in the chart (see Figure 4.20).

To get a deeper understanding of the AI's performance at the interior design aspect level, A frequency chart of incorrect values for each aspect was plotted (see part (a) of Figure 4.21) and part (b) of Figure 4.21 shows the percentage of incorrectness per aspect. The percentage chart was prepared as percentages provide a normalized measure that allows for easy comparison across different categories, regardless of the total number of instances in each category. The charts show that the experts most disagreed with the AI on [redacted] and *Style*, while they most agreed on Colors and [redacted].

Similarly a frequency chart for incompleteness was also plotted (part (a) of Figure 4.22) to visualise how many times an aspect was marked incomplete for the images compared to the total number of aspects for which insights were generated by the AI for the 55 images. Part (b) of Figure 4.22 shows this frequency separately, indicating that most incompleteness was identified for aspects - [redacted]

Out of the 10 aspects, we can loosely categorize some as more objective and others as more subjective.

Interior Design Aspects	No of Insights Marked Incorrect	Total insights generated per aspect
[redacted]	25	188
Furniture	48	429
[redacted]	54	312
[redacted]	26	544
[redacted]	55	276
[redacted]	8	330
Storage_solutions	55	385
Colors	16	385
[redacted]	22	385
[redacted]	21	426
Total	330	3660

Table 4.1: Evaluation Results : Incorrectness

Interior Design Aspects	No of times an aspect was marked incomplete	No. of times insights were generated per aspect
[redacted]	5	55
Furniture	2	55
Style	2	55
[redacted]	4	55
[redacted]	6	55
[redacted]	2	55
Storage_solutions	2	55
Colors	2	55
[redacted]	2	55
[redacted]	2	55
Total	29	550

Table 4.2: Evaluation Results : Frequency of Incompleteness

For instance, [redacted] and *Storage* tend to be more objective, while *Style* and [redacted] are more subjective. This observation is made because the charts were plotted to explore potential differences in the evaluation of these aspects. However, as the charts show, no definitive conclusions could be drawn about any distinct patterns in how interior designers assess correctness or completeness between these two categories of aspects.

Inter-Rater Reliability:

There were 7 images from the sample that were evaluated by both experts. The inter-rater reliability coefficient, *Cohen's kappa* , was calculated using the evaluation results of these 7 images to understand the level of agreement between the two experts. Cohen's kappa (K) is a statistical measure used to calculate the degree of agreement between two raters [82].

Incorrectness was calculated to be **0.55** and *incompleteness* to be **0.56**. Using the table by Landis and Koch [9] (see Table 4.3) the Cohen's Kappa Coefficient values for the both the criteria map to a moderate agreement between the experts.

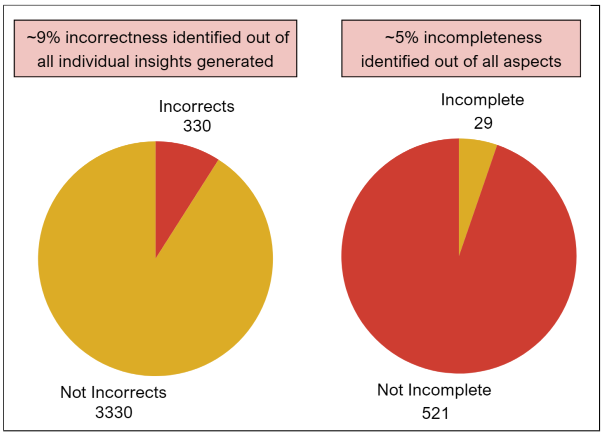
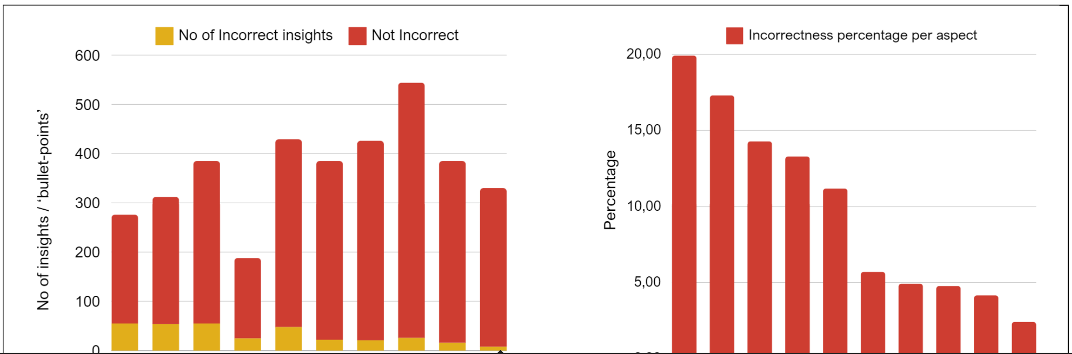


Figure 4.20: Overall Results



[redacted]

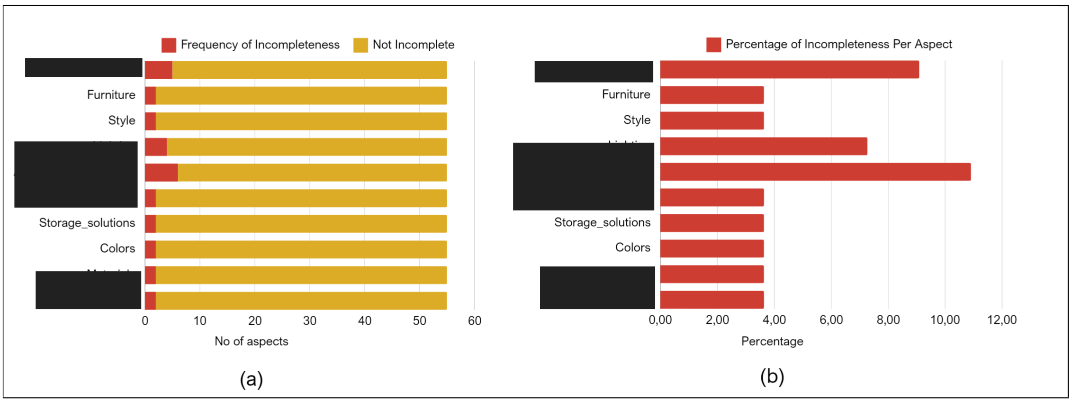


Figure 4.22: Incompleteness Analysis per aspect (a) no of times an aspect was marked incomplete (b) frequency of incompleteness per aspect

4.3.3 Qualitative Evaluation:

To gain an overall qualitative understanding of the AI's performance as a domain-specific image annotator, a post-evaluation interview was conducted with both interior design experts who evaluated the annotations. A

	Kappa
>0.8	Almost Perfect
>0.6	Substantial
>0.4	Moderate
>0.2	Fair
0-0.2	Slight
<0	Poor

Table 4.3: Cohen's Kappa Interpretation Table by Landis and Koch [9]

semi-structured interview format was chosen with questions to help capture the experts' opinions and insights that quantitative criteria might have missed.

Interview Guide:

At the beginning of the interview the experts were given a brief description of the results of the quantitative evaluation in order to give them a closure to their efforts. They were then given a brief introduction to the interview and A verbal consent was requested and received to record the interview and use the insights in the report. Since the same interior design experts who participated in the workshop were also the evaluators, this interview did not require a specific information sheet. They were already aware of the research details and how their information would be processed. See Appendix D.1 for Information Sheet used for the Workshop. Nonetheless, the key points about information processing were reiterated verbally during the interview.

They were then asked questions regarding their general experience with the evaluation, their understanding of the criteria - incomplete and incorrect, and their observations on the AI's handling of both subjective and objective aspects. The experts were also asked about how the AI's insights compared to those of a human interior designer and their overall impressions of the AI's knowledge of the field. See Appendix F.1 for complete interview guide.

Interview Analysis:

A thematic analysis was performed with the transcripts of the interview. The transcript was semantically coded. These codes were grouped based on similarity of context and then emerging themes were identified. See Figure 4.23.

Incorrectness :

- Criteria Understanding: The experts looked for obvious mismatch of insights in the annotations and the image for identifying incorrectness, indicating that they correctly grasped the definition of the criteria.
- [redacted] with incorrect technical details and contradictions within the insights for Style were highlighted as significant issues in the AI's annotations.
- Hallucinations related to subjective aspects were also identified as a recurring reason for incorrectness.

Incompleteness :

- Criteria Understanding: Incompleteness was identified based on obvious omissions, indicating that they correctly grasped the definition of the criteria.
- The experts indicated that they rarely felt annotations were incomplete.

Terminology & Language :

- Technical Language: The experts found some of the language used in the annotations to be overly technical, particularly for the aspects [redacted]
- Repetitive Terms : They are identified that the terms used in the insights were repetitive and there was a lack of creativity in the annotations, which is unlike a human domain expert.

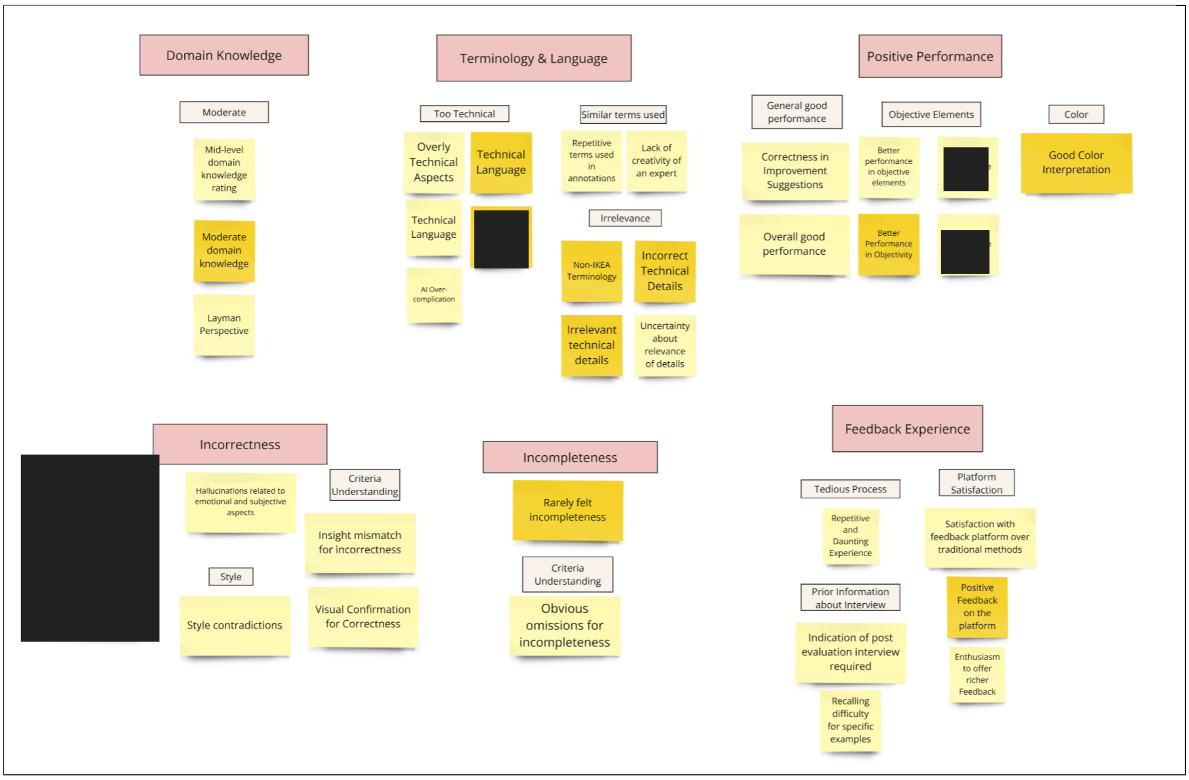


Figure 4.23: Thematic Analysis of Post-Evaluation Interview

- Non-IKEA Terminology: The use of non-IKEA terminology and irrelevant technical details caused some confusion among the experts. They were uncertain about the relevance of these details.

Following are few quotes from by the interior designers to further support the above analysis:

> *”For example, I remember the part about* [redacted] *that was like way, way too technical for what we are normally used to considering in the work we do” — An evaluator of the AI’s outputs.*

> *“Not English I mean, but the wording so to say, but the words are very complicated, we don’t hear those words often” — An evaluator of the AI’s outputs.*

> *“but I think also the language is not that language that we are, you know, connecting home furnishing to. It’s (the annotations are) much more technical language and also a lot of words that are not IKEA. We use simple words to express things, but it used complicated terms a lot” — An evaluator of the AI’s outputs.*

Domain Knowledge :

- Mid-level expert domain knowledge: The experts rated the AI’s domain knowledge as moderate. There was a perception that the AI provided insights from a layman’s perspective, and its understanding was considered to be at a mid-level i.e., not entirely incorrect at the same time not to the level of an expert. The following quote exemplifies this opinion from the interior designers.

> *“I think it felt a bit more like a customer or someone with little interior design knowledge was evaluating a space and saying this room is cozy or if I would ask my aunt for example, how would you describe this room?” — An evaluator of the AI’s outputs.*

Positive Performance :

- Overall Good Performance: Despite the issues, the experts acknowledged the AI had a good understanding of the interiors overall.
- Objective Elements: Better performance was noted in objective elements such as [redacted] and [redacted] The AI's interpretation of colors was positively received.

Feedback Experience :

- Tedious Process: The experts found the evaluation process to be repetitive and daunting, indicating that the task was inherently tedious.
- Platform Satisfaction: Despite the tedious nature of the task, the experts expressed satisfaction with the feedback platform compared to traditional methods. They appreciated the platform's design and found it facilitated a smoother evaluation experience.
- Prior Information: Experts noted the necessity of being informed about the post-evaluation interview earlier as they would have paid more attention to specific examples during the quantitative evaluation.

4.3.4 Overall Interpretation of the Results:

Quantitative & Qualitative Results Alignment:

The quantitative results for incorrectness indicate specific areas where the AI's performance is lacking, particularly in [redacted] and Style (Part (a) of Figure 4.21. This aligns with the qualitative feedback from experts, which also highlighted issues in these aspects. The low incompleteness in the quantitative evaluation aligns with the qualitative feedback stating that experts rarely found something incomplete. This could perhaps also be attributed to the tedious nature of the task as indicated in the interviews. It can be challenging to evaluate about 30 annotations with each 10 aspects of interior design insights. It was perhaps easier for experts to pinpoint what was wrong rather than identify all that was missing.

Subjective vs Objective Aspects:

As mentioned earlier, although there is not a clear distinction, certain aspects could be identified as more objective than others. The experts in the post-evaluation interview indicate that the AI performs better in generating objective insights. However, the quantitative results do not specifically show any differences in results between subjective and objective aspects. This opinion of the experts could then have been due to the fact that there were more insights generated for objective aspects than subjective ones. Because of this majority and having a clearer visual confirmation of the objective insights, experts may have remembered them more prominently. Further investigation is needed to confirm this interpretation.

Implications of Moderate Inter-rater Reliability :

The moderate inter-rater reliability indicates that while there was some consistency in the experts' evaluations, there were also areas of disagreement. This raises questions about the extent to which the quantitative evaluation can be considered as ground truth. Since this exploration aims to assess how well the AI's results align with those of any interior designer, the moderate inter-rater reliability does not directly undermine our analysis of the AI's competency in this task. However, if the evaluated results are to be used as a ground truths, for example for creating image-annotation datasets, this moderate reliability needs to be investigated to establish a clearer understanding of what is considered true or false by the experts. A better understanding of the application of the data might help experts align on how strict or lenient they should be during the evaluation. This insight was also reflected in the interviews, where experts indicated an uncertainty of the relevance of the insights generated by the VLM.

Another point to consider is that the moderate inter-rater reliability between the two highly experienced experts could suggest either that the domain itself is inherently subjective, leading to low agreement between any two experts, or that the differences in agreement are specific to the particular experts involved in this research. To explore this further, assessing inter-rater reliability with a larger group of experts would provide more insight into this speculation.

VLM's Interior Design Analysis Ability:

The experts perceived the AI's insights as being from a 'layman's perspective', with a mid-level understanding of the domain. They expressed a positive feedback towards the overall performance of the VLM in generating interior design insights, which aligns with the low incorrectness and incompleteness rates from the quantitative results. However, given their opinions about the emotional aspects of the terminology, it can be interpreted that the VLM's insights lacked a 'human-ness' (An example quote can be found below to support this interpretation). Although they could not recall many specific examples of the technical and complex terminology, the fact that they repeatedly mentioned this issue suggests a clear gap between the language and terminology the experts would have used and what the AI used in its insights. They did highlight that if an interior design expert performs such an annotation task for many images, they would inevitably tend to provide varied insights using more creative terms. This analysis gives us an understanding of the pre-trained interior design knowledge of the VLM, specifically GPT-4. We can interpret that that while the AI can generate broadly correct annotations, it lacks the finesse required for expert-level performance.

> *"Then I also miss maybe this soft feeling of home furnishing, the being more emotional part" — An evaluator of the AI's outputs.*

Stage C Summary

In this section, the process followed to evaluate the synthetic image annotations generated in the previous stage, has been explained. The criteria for evaluations were established - 'Incompleteness' and 'Incorrectness'. 2 experts were sought for this.
The process was two-fold:

1. Quantitative Evaluation : The evaluations on the set criteria was gathered from the experts via a feedback platform.

2. Qualitative Evaluation : A post-evaluation interview was conducted with the same experts to get a deeper understanding of their opinions about the performance of the AI model.

Results of both quantitative and qualitative evaluation were analysed and elaborated. The quantitative evaluation resulted in 9% incorrectness and 6% incompleteness for the insights overall. These results analysed in combination with the post-evaluation interview with the experts led to a richer understanding of the AI's competency in this task of interior design analysis. It was analysed that while the AI was able to produce overall correct insights the terminology and language used did not match the way the interior designers would have analysed the space with.

Recap of the Sub-Research Questions

SQ 1.1: How can a CF be derived for the annotations to ensure all necessary information is included in them?

By first gathering key interior design aspects from LLM generated taxonomies for the domain. Although the taxonomies fell short in organisation and mutual-exclusivity the concepts identified can be useful in forming a foundation set of key aspects for the CF. These concepts can then be validated by synthesizing an image analysis workshop with interior design experts, followed by a direct consultation with them. This way a CF containing 10 key aspects of interior design was defined.

SQ 1.2: How can prompts for a VLM be effectively designed to generate interior design image annotations?

Through a set of multi-stage prompts enhanced by insights derived from domain-experts' analysis of rooms (workshop).

SQ 1.3: How can the validity of the generated annotations be checked?

Given the absence of a ground truth dataset, this answer was obtained through human expert evaluations. The validity was determined using two criteria 'Incorrectness' and 'Incompleteness'. A combination of quantitative and qualitative evaluations was applied to incorporate a rich human-feedback into the loop.

Auxiliary SQ 1.1: How can a *human-in-the-loop* approach be effectively integrated throughout the process?

Stage C: User experience considerations on the feedback platform to ensure that the inherently tedious evaluation process is as easy as possible for the evaluators.

Chapter 5

Discussing Key Learnings

Although this research focused on leveraging the image annotation capabilities of GPT-4, the study structure facilitated learnings across various facets of AI development along the way.

The following section elaborated key topics and insights. They also include limitations and recommendations for future research.

5.1 GPT-4's Domain Expertise

5.1.1 GPT-4 as a taxonomy 'copilot'

Despite the varied prompt-design strategies tested for the generation of a taxonomy for the use case - bedroom interiors analysis, results stayed quite similar. The results were regularly assessed by an IKEA taxonomy expert who deemed them inadequate due to a lack of mutual-exclusivity between concepts/categories, and poor organisation in the hierarchical structuring. The concepts that most reoccurred in the results and most suitable for task were selected for the *conceptual framework* CF, to be collaboratively improved by the domain experts (interior designers) through a workshop and qualitative consultation session. The minimal modification to the initially GPT-4 identified set of concepts indicate its proficiency in identifying key concepts within a domain. Hence, as mentioned earlier in this document, further research could be valuable for improving the LLM's taxonomy outcomes by re-framing the LLM as an interactive, iterative, 'copilot'-style tool, to be used by taxonomists to help them with limited elements of the taxonomy-creation task, such as basic concept identification.

5.1.2 GPT-4V can be an Interior Design Image Annotator, but not an expert Interior Designer

The implementation of the proposed methodology for generating interior design image annotations led to an understanding of the pre-trained interior design knowledge of the models used (GPT4/GPT-4V)

Despite not producing a robust taxonomy, this work (by combining LLM + co-creation/evaluation workshop with interior designers), produced a set of 10 core ID concepts that according to the ID experts sufficiently encompass all the necessary interior design information of a room. This set of concepts could prove invaluable to IKEA in other endeavors (business planning, marketing planning, product categorization, budget allocation, content strategy, etc), where understanding IKEA's core subject-matter is key. Since most of the concepts identified were derived using GPT-4, we can infer that GPT-4 has a foundational understanding of the domain that can be leveraged for various domain-specific applications.

The scores from the quantitative evaluation show that the pre-trained model, when combined with sufficiently robust prompting, can be effective at interior design analysis, at least from the standpoint of describing pictures of rooms with a generic understanding of interior design concepts. At the same time, the experts expressed through the post-evaluation qualitative interview that they consider the AI to have a 'mid-level interior design' skill-set with a 'layman tone' in the outputs. Therefore, automated image annotation with interior-design insights using GPT-4V, in its current state-of-the-art, is applicable for use-cases that do not require the insights to closely exhibit the expertise of interior designers but require only true and detailed insights that cover all interior design aspects of a room image. Some examples of such use-cases are product recommendations personalised for customer rooms, domain-specific image metadata generation, creation of datasets of domain-specific image annotations that can be used as ground-truths or fine-tuning datasets other ML or gen AI models. It must however be considered that the annotations were validated for a sample size of 55 images due to the availability constraints of domain experts. Hence, future application of this process is recommended to involve regular validation routines. Another point to note was that resolution of the bedroom images from the dataset were lower than an average customer room image. It can be thus speculated that when applied to higher resolution images this process could produce better results. Despite these limitations, the results are an initial contribution to the understanding of GPT-4's image analysis ability specifically for interior design concepts.

5.1.3 Two-stage prompts for richer domain-specific image annotations

The prompting strategy of first instructing the VLM to interiors found in the given bedroom image for a specific ID aspect, then summarising this description into phrases of insights, gave more detailed hence richer output. It is well-known that decomposition of tasks to more manageable steps improved the outputs of LLMs and VLMs. Such a two-stage prompt therefore helped focus on one cognitive process per stage. If directly prompted for key 'bullet-points' like insights, there is the risk of the model focusing on the most prominent features or insights it can quickly derive, potentially overlooking subtleties. However, in the two-stage approach, the initial description gives the model a more thorough understanding of the context, allowing it to generate more nuanced and detailed insights during the summarisation phase. The very recent study by Sun et al. [83] justifies this approach of "prompt-chaining" over a single prompts containing all the steps which in this case would have been one single prompt containing both the first stage and the second stage. This prompting strategy could be applicable and beneficial for various image analysis tasks beyond the domain of interior design. It could also be a future direction for the LLM-driven taxonomy exploration with the chain-of-thought prompts broken down further into more discreet multi-stage prompts prompts.

5.2 Humans in the Loop

5.2.1 A Step Towards Better Collaboration with Non-prompt-designers

Prompt design, being the backbone of this exploration, was heavily informed by the collaboration with experts. As highlighted by Zamfirescu-Pereira et al. [73], the skill-set required for effective prompt design is not yet common enough to enable direct assistance from domain experts in crafting prompts. This study took a step towards utilizing expert collaboration for domain-specific prompting through qualitative methods such as interviews and workshops. The collaboration with interior design experts provided valuable insights that were incorporated into the prompt design, enhancing the specificity and relevance of the AI-generated annotations.

However, it is acknowledged that closer co-designing of prompts with experts might have been more efficient and effective. This level of collaboration would have been easier if the experts had a deeper understanding of prompting strategies. Hence this work promotes initiatives that offer domain-experts skill development in gen AI prompting techniques. Fostering a better understanding of prompt design among domain experts would make it more accessible and intuitive for them to contribute directly to AI training and application.

5.2.2 Better Feedback Platforms Make Happier Human Evaluators

Human evaluation is crucial in any AI development process. Given the novelty of applying a VLM to generate detailed interior design analyses and the absence of other ground truth sources for image annotations containing interior design insights, domain-expert evaluations played an especially pivotal role in this study. While the absence of a ground-truth dataset seemed like a limitation, it proved to be an opportunity to explore novel approaches to human-in-the-loop AI, which could be used broadly in AI projects by other researchers in the future.

The human evaluation process involved experts assessing AI-generated annotations for images, with each annotation covering ten distinct aspects of interior design, a potentially tedious and time-consuming task for human evaluators.

Recognizing the importance of user experience in gathering high-quality feedback [71, 72], a custom feedback platform was built to enhance user-friendliness and mitigate evaluator fatigue. Experts could easily view the AI-generated annotations alongside the corresponding images and they could pause and resume their assessment at their convenience. All leading to experts expressing a clear preference for this platform over a more manual methods of providing feedback which was inherently a tedious and mundane task.

5.2.3 Importance of Quantitative + Qualitative Evaluations of VLM Outputs

For assessing the capability of the VLM for generating interior design insights, two primary evaluation criteria were employed: *incorrectness* and *incompleteness*. These criteria were chosen due to their prevalent use in similar studies and their applicability for this use case. The quantitative evaluation yielded promising results, with GPT-4V demonstrating low rates of incorrectness ($\tilde{9}$%) and incompleteness ($\tilde{5}$%) in generating interior design insights. However, the qualitative feedback gathered from post-evaluation interviews with experts revealed a more nuanced perspective that the quantitative criteria alone failed to capture.

The quantitative results if looked at alone, indicate that the AI system was capable of producing relevant content that largely aligned with expert knowledge. However, in the post-evaluation interview, they expressed that the AI fell short in capturing professional interior designer terminology and creativity.

While the criteria of *incompleteness* and *incorrectness* were sufficient for assessing the basic capability of the VLM to generate accurate interior design insights, their results alone would have misled the researcher into concluding that the VLM had overall expert level interior design analysis capability. We have observed that this is, in fact, not the case as the interior designers expressed in the interviews that the output do not match "tone", the "creativity" and the "human-ness" with which they would have performed the interior design analysis.

Hence, if the use-case demands for a detailed understanding of a domain expertise of a gen AI model, future research should consider collaborating with domain experts to develop more nuanced, domain-specific evaluation criteria.

More importantly, this study highlights the value of integrating both quantitative criteria and qualitative insights in AI evaluation. As noted by Wang et al. [71], such a combined approach is particularly beneficial for complex feedback scenarios. The findings of the study provide empirical support for this recommendation, demonstrating how qualitative feedback can uncover critical insights that quantitative measures might overlook.

It must be noted that this research was limited to one iteration of evaluation. An ideal extension to this study would be incorporation of continuous feedback loops with domain-experts to refine the VLM generated data and the evaluation criteria.

Nonetheless, this study contributes to the growing body of research advocating for more comprehensive evaluation methods in AI [58, 60], particularly in domain-specific applications. While quantitatively measurable criteria provide a valuable baseline assessment, qualitative expert feedback is crucial for uncovering subtle yet significant aspects of AI performance. It is strongly recommended for future research in domain-specific synthetic data generation by generative AI to consider adopting this combined approach to evaluation, ensuring a more holistic understanding of AI capabilities and limitations.

5.3 Considerations Regarding Bias

5.3.1 AI Bias

Biases in AI is are inevitable. In fact, in some sense, pre-trained LLMs and other ML algorithms function by way of inductive biases [74, 75], which determine their output predictions. In this study, the *conceptual framework* and the constraints given in the prompts, attempted to steer the model to have an inductive bias towards providing desirable outcomes.

Since the main goal of this study was to evaluate the extent to which model aligned with the human domain expertise, the research did not focus on evaluating other the full gamut of potential counterproductive or harmful biases. However, the AI models, all from the GPT-4 series, had been pre-trained by OpenAI and had undergone extensive validation to control for common harmful bias [84, 85]. Nonetheless, if future research builds upon this study's findings, iterations of evaluations are recommended specifically to mitigate unwanted biases.

Another source from which biases might have emerged is the dataset of images used. Upon a cursory assessment of the images, they appeared to primarily depict bedrooms from Western countries. Therefore, future iterations of this research ought to make additional attempts to incorporate evaluations on a more measurably diverse image dataset.

5.3.2 Human Bias

The final evaluations in this study were performed by only 2 human experts, despite the sizeable volume of annotations that they were asked to review, such evaluations of AI outputs can be a tedious task that and could potentially induce an 'evaluation fatigue' [86]. As a result of this tricky effect, the evaluators in this study might have provided less critical feedback than necessary.

Additionally, every annotation for evaluations contained about 5 to 7 'bullet-points' of insights for each of the 10 aspects per image. Considering the second of the two evaluation criteria - *incompleteness*, there is a chance that the evaluators felt a 'false sense of completeness' upon seeing a visibly long list of bullet-points, which might have given the appearance of thoroughness simply due to its length.

For future research, a few recommendations to alleviate these factors can could be - by increasing the number of human evaluators, by providing more proactive evaluation guidelines (for example, guidelines that explicitly state that the length of a list is not necessarily correlated with its completeness), by providing extra blank spaces for the evaluators to modify values in the output or fill missing values, etc.

5.4 Answering the Research Question

It is evident that the answer to the research question **"How can LLMs and VLMs be applied to generate domain-specific synthetic image annotations for interior design with human-in-the-loop approach?"** is multi-layered.

The evaluation results indicate that the GPT-4 model can provide correct and thorough interior design insights for 10 different interior design aspects, with a small error margin. Therefore, the method of defining a *conceptual framework* for the annotations, generating domain-specific image annotations, and expert evaluations—all within a human-in-the-loop approach—can be considered a viable answer to the question. The emphasis on beginning the process by defining a *conceptual framework* for annotations is a key contribution, addressing the lack of guiding principles for generating domain-specific insights with gen AI without a reference dataset, and making this method adaptable beyond interior design.

This method can be implemented to produce automatic interior design insights for use cases that do not particularly require expert-level analysis. For example, within IKEA, this process could be used to generate metadata about customer room images, a task currently performed by manually by employees themselves. Another application is to generate domain-specific synthetic datasets of image annotations, which could be useful for further research and for fine-tuning smaller models.

However, this method has unveiled that while the model can generate correct interior design insights, it is unable to yet match the expertise of a human expert. Nonetheless, this research helped us realize that in order to get outputs that align closely with domain expertise, such as those needed for interior design consultations, a deeper understanding of the domain itself and its specific expectations is necessary among the researchers and developers. This highlights the importance of incorporating more experts in the loop.

This research utilized the GPT-4 suite of models, but the methodology can be replicated with other similarly performing LLMs and VLMs. Although similar performance cannot be guaranteed without proper testing, given that the study predominantly utilized the pre-trained capability of GPT-4, it can be inferred that a similarly performing LLM or VLM with comparable benchmark scores could potentially produce similar results. The same might also be said about adjacent domains such as art and fashion design. More importantly, seeking an answer to this research question has unlocked various insights and future considerations across the life-cycle of generative AI applications like collaborative prompt design, more qualitative evaluations.

Chapter 6

Conclusion

This work explored the potential of using an LLM and VLM, specifically *GPT-4*, to generate image annotations containing interior design insights. Through a structured three-stage methodology that included domain-expert collaborations, prompt designing, and expert evaluations, the research sought to understand how these advanced AI models could be applied to analyze and annotate images in a standardized way such that the annotations consistently contain the required information regarding various interior design aspects.

In pursuit of the CF(CF) for the annotations, a sub-exploration of LLM-driven taxonomy was conducted. This exploration of generating taxonomies for a specific use-case with prompt engineering, revealed that, although *GPT-4* is adept at generating key concepts within interior design with the pre-trained domain knowledge, it falls short in delineating and organising these categories into a satisfactory taxonomies. Nonetheless, this process helped form the base for the 10 key aspects of ID as the CF for the contents of the annotation. Such defining of a CF is a part of the methodology that can be adapted to other domain-specific image annotation tasks as well.

A two-stage prompt strategy (First stage - description based on an ID aspect; Second stage - summarising these descriptions into phrases) emerged most effective to capture detailed interior design information from an image with *GPT-4V*. These prompt were informed by insights derived from a workshop with interior designers performing the same task of room image analysis. This way of translating expert thought process for a task into prompts without them having any prompting knowledge opens possibilities to co-develop AI-systems with domain-experts via prompt engineering. Perhaps this approach would help start a conversation with Zamfirescu-Pereira et al. [73] about how ”Johny can help with prompt engineering even if he can't prompt himself”.

The generated sample set of annotations were evaluated with experts. The low *Incompleteness* of aspects and *Incorrectness* of specific insights for about 60 images, both under 10%, combined with the experts' qualitative feedback regarding the lack of finesse in the insights, indicated that GPT-4V can be utilised for automating interior design analysis for use-cases that only require a generic understanding of interior design. The quantitative and qualitative evaluation together led to an enriched understanding of the interior design image analysis capability of the AI model. The moderate (0.5) inter-rate reliability between the two domain-experts with over 10 years of ID expertise left us wondering about the subjective nature of the domain itself. This insight urges future research to define evaluation criteria of such gen AI tasks with improved understanding of the domain in close collaboration with experts.

Other recommendation for future research were made along the way including iterative ”co-pilot”-like usage of GPT-4 for taxonomy generation, closer co-designing prompts with domain-experts, testing the process with a diverse dataset and increased number of expert-evaluators, application and exploration of this process in other adjacent domains like fashion design, culinary arts, etc.

Acknowledgements

Gratitude

I would like to extend my gratitude to IKEA (Ingka Group, Amsterdam) for their collaboration on this project. A special thanks to my supervisor there, Dan Porder, whose extraordinary mentorship and enthusiasm was key to my consistency and drive throughout this journey. I am also grateful to the other team members at IKEA—the interior designers, taxonomists, and AI specialists, who provided great support at various stages of the process.

At the University, I am immensely thankful to my supervisor, dr. Lorenzo Gatti, for his support in navigating this collaboration and ensuring that my thesis remained academically relevant. His guidance consistently boosted my confidence at the same time provided actionable insights as next steps. I would also like to thank dr. Doina Bucur for joining the examination committee, chairing it effectively, and offering timely and insightful feedback.

Finally, I wish to express my heartfelt thanks to my incredible friends and family, both here in Enschede and back in India. During times when the thesis work got overwhelming, their warmth and care kept me going, helping me avoid burnout and complete this thesis with resilience.

Statement of Usage of AI for Writing

During the preparation of this document I used the Generative AI tools (e.g., ChatGPT) to organize my thoughts related to the information presented, as well as to improve the vocabulary and clarity of the content. The tools were in no way used to generate new information or perform any background research. I reviewed the content generated with the help of these tools and take full responsibility of the work.

Bibliography

[1] Pedro Cuenca. Pcuenq/lsun-bedrooms · datasets at hugging face. URL `https://huggingface.co/datasets/pcuenq/lsun-bedrooms`.

[2] PK Bhagat and Prakash Choudhary. Image annotation: Then and now. *Image and Vision Computing*, 80:1–23, 2018.

[3] Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

[4] Christos V Smailis and Dimitris K Iakovidis. Ontology-based automatic image annotation exploiting generalized qualitative spatial semantics. In *Hellenic Conference on Artificial Intelligence*, pages 299–306. Springer, 2012.

[5] Min Wang, Ata Mahjoubfar, and Anupama Joshi. FashionVQA: A Domain-Specific Visual Question Answering System. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3514–3519, 2023.

[6] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with GPT-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 262–271, 2023.

[7] Shinjini Ghosh and Sagnik Anupam. CapText: Large language model-based caption generation from image context and description. *arXiv preprint arXiv:2306.00301*, 2023.

[8] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

[9] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics, Volume 33*, pages 159–174, 1977.

[10] IKEAKreative. Inspired design for living spaces, in app or web - IKEA Kreativ. URL `https://www.ikea.com/nl/en/home-design/`. [Accessed 21-05-2024].

[11] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

[12] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of LMMs: Preliminary explorations with GPT-4V (ision). *arXiv preprint arXiv:2309.17421*, 9 (1):1, 2023.

[13] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[14] Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*, 2023.

[15] Vahid Ashrafimoghari, Necdet Gürkan, and Jordan W Suchow. Evaluating Large Language Models on the GMAT: Implications for the Future of Business Education. *arXiv preprint arXiv:2401.02985*, 2024.

[16] Yujuan Ding, Yunshan Ma, Wenqi Fan, Yige Yao, Tat-Seng Chua, and Qing Li. Fashionregen: Llm-empowered fashion report generation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 991–994, 2024.

[17] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? Or you think what you see? An image-context reasoning benchmark challenging for GPT-4V (ision), LLaVa-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023.

[18] Angie Lee. What Is a Pretrained AI Model? — blogs.nvidia.com, 2022. URL `https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/`. [Accessed 24-06-2024].

[19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[20] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatGPT. *arXiv preprint arXiv:2302.11382*, 2023.

[21] Xavier Amatriain. Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*, 2024.

[22] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[23] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

[24] OpenAI. OpenAI Models. URL `https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`. [Accessed 21-05-2024].

[25] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook.* sage, 1994.

[26] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[27] Chirag Shah. From prompt engineering to prompt science with human in the loop. *arXiv preprint arXiv:2401.04122*, 2024.

[28] Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*, 2024.

[29] OpenAI. Enterprise privacy at OpenAI, . URL `https://openai.com/enterprise-privacy`.

[30] Azure. Azure OpenAI service models - azure OpenAI. URL `https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models`.

[31] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[32] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4v? Closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[33] Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. GPT-3.5, GPT-4, or BARD? evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032, 2023.

[34] Aryan Rangapur and Aman Rangapur. The battle of LLMs: A comparative study in conversational QA tasks. *arXiv preprint arXiv:2405.18344*, 2024.

[35] Tianyu Han, Lisa C Adams, Keno Bressem, Felix Busch, Luisa Huck, Sven Nebelung, and Daniel Truhn. Comparative Analysis of GPT-4Vision, GPT-4 and Open Source LLMs in Clinical Diagnostic Accuracy: A Benchmark Against Human Expertise. *medRxiv*, pages 2023–11, 2023.

[36] OpenAI. Key concepts, . URL `https://platform.openai.com/docs/concepts`.

[37] Jillissa Lynn Moorman. *Interior design versus interior decoration: evaluating the public's perception of design television shows*. Iowa State University, 2008.

[38] Centre for Arts and Technology. What is the difference between interior design and interior decorating?, Nov 2023. URL `https://digitalartschool.com/news/what-difference-between-interior-design-and-interior-decorating/`.

[39] Leslie F. Sikos. *The Semantic Gap*, pages 51–66. Springer International Publishing, Cham, 2017. ISBN 978-3-319-54066-5. doi: 10.1007/978-3-319-54066-5_3. URL `https://doi.org/10.1007/978-3-319-54066-5_3`.

[40] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. ChatGPT asks, BLIP-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023.

[41] Hanieh Khorashadizadeh, Nandana Mihindukulasooriya, Sanju Tiwari, Jinghua Groppe, and Sven Groppe. Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text. *arXiv preprint arXiv:2305.08804*, 2023.

[42] Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. Iterative zero-shot LLM prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*, 2023.

[43] Rafael Roque de Souza, Thiago Luna Pinheiro, Julio Cesar Barbour Oliveira, and Julio Cesar dos Reis. Knowledge Graphs Extracted from Medical Appointment Transcriptions: Results Generating Triples Relying on LLMs. 2023.

[44] Ayato Kuwana, Atsushi Oba, Ranto Sawai, and Incheon Paik. Automatic Taxonomy Classification by Pretrained Language Model. *Electronics*, 10(21):2656, 2021.

[45] Chirag Shah, Ryen W White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, et al. Using large language models to generate, validate, and apply user intent taxonomies. *arXiv preprint arXiv:2309.13063*, 2023.

[46] Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. House-gan: Relational generative adversarial networks for graph-constrained house layout generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 162–177. Springer, 2020.

[47] Quang-Vinh Dang, Thai-Ha Pham, My-Linh Tran, Minh-Hung Dang, Quoc-Khanh Nguyen, and Anh-Nguyen Nguyen. Towards an automatic interior design system using GAN. In *2021 4th International Conference on Data Science and Information Technology*, pages 170–172, 2021.

[48] Hanan Tanasra, Tamar Rott Shaham, Tomer Michaeli, Guy Austern, and Shany Barath. Automation in Interior Space Planning: Utilizing Conditional Generative Adversarial Network Models to Create Furniture Layouts. *Buildings*, 13(7):1793, 2023.

[49] Jun-Hyeck Choi and Jae-Seung Lee. A study of interior style transformation with GAN model. *Journal of KIBIM*, 12(1):55–61, 2022.

[50] Liu Yanhua. Research on the Application of Artificial Intelligence in Interior Design. *International Journal of Science and Engineering Applications*, 2024.

[51] Yihan Hou, Manling Yang, Hao Cui, Lei Wang, Jie Xu, and Wei Zeng. C2Ideas: Supporting Creative Interior Color Design Ideation with Large Language Model. *arXiv preprint arXiv:2401.12586*, 2024.

[52] Rishabh Kabra, Loic Matthey, Alexander Lerchner, and Niloy Mitra. Leveraging VLM-Based Pipelines to Annotate 3D Objects. In *Forty-first International Conference on Machine Learning*, 2024.

[53] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a VLM. *arXiv preprint arXiv:2403.18406*, 2024.

[54] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. (A) I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469, 2024.

[55] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[56] Andrew Y Wang, Sherman Lin, Christopher Tran, Robert J Homer, Dan Wilsdon, Joanna C Walsh, Emily A Goebel, Irene Sansano, Snehal Sonawane, Vincent Cockenpot, et al. Assessment of pathology domain-specific knowledge of ChatGPT and comparison to human performance. *Archives of pathology & laboratory medicine*, 2024.

[57] Amara Tariq, Man Luo, Aisha Urooj, Avisha Das, Jiwoong Jeong, Shubham Trivedi, Bhavik Patel, and Imon Banerjee. Domain-specific llm development and evaluation–a case-study for prostate cancer. *medRxiv*, pages 2024–03, 2024.

[58] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.

[59] Baoyuan Wu, Weidong Chen, Peng Sun, Wei Liu, Bernard Ghanem, and Siwei Lyu. Tagging like humans: Diverse and distinct image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7967–7975, 2018.

[60] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*, 2021.

[61] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[63] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[64] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022.

[65] Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156, 2024.

[66] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021.

[67] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[68] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[69] Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18(193):1–46, 2018.

[70] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–35, 2019.

[71] Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. Putting humans in the natural language processing loop: A survey. *arXiv preprint arXiv:2103.04044*, 2021.

[72] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI magazine*, 35(4):105–120, 2014.

[73] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[74] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

[75] Samira Abnar, Mostafa Dehghani, and Willem Zuidema. Transferring inductive biases through knowledge distillation. *arXiv preprint arXiv:2006.00555*, 2020.

[76] Henrik Enghoff. What is taxonomy?–an overview with myriapodological examples. *Soil organisms*, 81 (3):441–451, 2009.

[77] Dumitru Dan Burdescu, Cristian Gabriel Mihai, Liana Stanescu, and Marius Brezovan. Automatic image annotation and semantic based image retrieval for medical domain. *Neurocomputing*, 109:33–48, 2013.

[78] IKEA. Bedroom tips and ideas. URL `https://www.ikea.com/us/en/rooms/bedroom/how-to/?itm_campaign=REVAMPHFH&itm_element=5_MODULE-Banner-&itm_content=PILL-SIDER-BEDROOM`.

[79] Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*, 2023.

[80] OpenAI. Text generation models, . URL `https://platform.openai.com/docs/guides/text-generation`.

[81] James W Drisko and Tina Maschi. *Content analysis*. Oxford University Press, USA, 2016.

[82] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[83] Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. Prompt Chaining or Stepwise Prompt? Refinement in Text Summarization. *arXiv preprint arXiv:2406.00507*, 2024.

[84] OpenAI. GPT-4 System Card, Mar 2023. URL `https://cdn.openai.com/papers/gpt-4-system-card.pdf`.

[85] OpenAI. GPT-4V(ision) system card, Sep 2023. URL `https://openai.com/index/gpt-4v-system-card/`.

[86] Mingming Fan, Xianyou Yang, TszTung Yu, Q Vera Liao, and Jian Zhao. Human-AI collaboration for UX evaluation: effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–32, 2022.

Appendix A

A.1 Full Outputs of Interior Design Insights Generated by VLM without a conceptual framework

Prompt : Generate interior design insights about this room. Ensure to include all the relevant aspects.

Room Image 1 :

Output from the VLM :

The interior design of the room depicted in the image features several noteworthy elements. Here are some insights:

1. **Wall Mural**: The most prominent feature is the large world map mural on the wall behind the bed. This mural serves as both an educational and decorative element, adding a sense of adventure and global awareness to the room.
2. **Color Scheme**: The room uses a combination of warm and earthy tones. The bedding features stripes in shades of green, orange, brown, and beige, which complement the colors in the map mural. This creates a cohesive and visually appealing color palette.
3. **Furniture**: The bed has a simple, modern design with a low profile. It is paired with a minimalist desk and chair, which are positioned near a window. The chair is made of transparent material, likely acrylic, which adds a contemporary touch to the room without making it feel cluttered.

[redacted]

Overall, the room is well-designed with a focus on functionality, comfort, and aesthetic appeal. The world map mural is a standout feature that sets the tone for the room, while the thoughtful use of color and light creates a warm and inviting atmosphere.

Figure A.1: Output of Interior Design Insights Generated by VLM without a conceptual framework - Image 1

Prompt: Generate interior design insights about this room. Ensure to include all the relevant aspects.

Room Image 2 :

Output from the VLM :

The image showcases a modern bedroom with several notable design elements and features. Here's an analysis of the interior design aspects of this room:

1. **Color Scheme**:
2. The room primarily uses neutral tones such as white, beige, and light grey. These colors contribute to a calm and relaxing atmosphere.
3. The accents of dark brown in the rugs add warmth and contrast to the otherwise light-colored room.

[redacted]

22. **Ambiance**:
23. The combination of neutral colors, soft lighting, and plush textures creates a cozy and inviting ambiance.
24. The room exudes a sense of tranquility, making it an ideal space for relaxation and rest.

In summary, this bedroom is a well-designed space that balances aesthetics with functionality. The use of neutral colors, varied textures, and strategic lighting creates a serene and comfortable environment.

Figure A.2: Output of Interior Design Insights Generated by VLM without a conceptual framework - Image 2

Appendix B

Defining a Conceptual Framework

B.1 Interview With Taxonomists

Using Gen AI to help organise Interior Design Knowledge Information Sheet - Interview

Hello,

I invite you to this interview for a master thesis study in collaboration with IKEA. I am Sandhiyaa Balasubramanian Yamuneswari pursuing the Interaction Technology Master's Program at the University of Twente. Please read the following details about the study carefully to make an informed decision on whether to take part in this interview. Please ask questions if you have any.

Overarching Research Aim

Exploring the potential of Generative AI, particularly Large Language Models (LLMs) and Vision-Language Models (VLMs), to support and enhance the room image analysis process in interior design. To achieve this, I will be exploring the potential of LLMs for transforming unstructured interior design knowledge, into structured taxonomies. Then I will explore the potential of VLMs in using them to derive structured interior design insights from room images. These insights could be useful for enhancing workflows related to the generation of structured insights for Interior Design.

Purpose of this Interview

The aim of this interview is to learn about taxonomies and your workflows for generating structured-content, to elicit best practices in the field in which the interviewee works, and to understand how professionals like you collaborate cross-functionally. Some of the information gathered may be used for designing automated processes for taxonomy and ontology generation, and some of the data collected may be utilized by generative AI.

The Interview Session

Participating in this interview involves approximately 1 hour of your time answering questions regarding taxonomies, and the process of a taxonomist for gathering subject knowledge, collaboration with domain experts, organizing information and validating the resulting taxonomies.

Possible Risks & Benefits

The interview topic should not cause any discomfort. Moreover, the research has been reviewed by the Ethics Committee of the University of Twente, ensuring that the study will be conducted ethically and with the utmost respect for participants. There is no compensation for participating in the study.

Data Privacy & Confidentiality

If consent is given, the interview will be recorded and transcribed. If not, the responses will be manually noted down. The transcription and notes will be analyzed as part of the research process. The transcriptions maybe be stored locally and securely until the research is complete, after which they will be promptly and permanently deleted from any form of my (the researcher's) storage. The recordings and the transcriptions, however, maybe be stored securely by INGKA. The responsibility of the artifacts of the workshop will then completely belong to INGKA. The transcription tool used will be locally run, ensuring that it will not be shared in any third-party platforms. To uphold confidentiality, all personal details will be anonymized. The final will not contain any personally identifiable information, and anonymous quotes may be utilized to support specific conclusions. Upon request, or as appendices in the final document, anonymized transcripts may be made available, trimmed of any proprietary company information. The final thesis report will be publicly accessible, and there is a possibility of its use in research publications within academic circles

Participation

Participation is completely voluntary. You have the right to refuse participation, the right to refuse any question and you can withdraw from the study at any point without any consequence and without providing any justification. After the interview, should you decide that you do not want your responses to be included in the study, you have the option to request the removal of your interview results and any

analyses derived from your answers any time before April 2024. The results will then be integrated into the research, after which it will not be feasible to retrieve the portion of results influenced by your interview and remove it.

Researcher Contact

If you have any further questions, feel free to ask me now, or contact me afterwards at ¡ikea_email_id¿

Additionally, you can contact the Secretary of the Ethics Committee of the Faculty of Electrical Engineering Mathematics and Computer Science at the University of Twente through ethicscommittee-cis@utwente.nl.

Once you have read the above information and have decided to take part in the study, please convey your consent verbally by answering to the following two questions:

Are you sufficiently informed and do you agree to take part in the study? Do you mind if I recorded this meeting?

Thank you

B.2 Interview with Taxonomist - Guide

Interviewing Taxonomists : (An interview guide with open-ended questions to get subject matter expertise of taxonomy making)

Note : A detailed information sheet is provided before beginning the interview and the interviewee's consent is asked for verbally

Brief : I would like to understand this field of knowledge structuring better. I am working on having a Generative AI model capture facts and categories from guides and other unstructured material, and have it come up with a taxonomy of these categories, specifically for interior design. I would like to understand Taxonomies making process so I could instruct the AI better for this task.

Are you sufficiently informed and do you agree to take part in the study?

Do you mind if I recorded this meeting?

Questions:

- Who is a taxonomist?
- Could you briefly talk about taxonomies? Just to get an expert perspective on this.

Understanding the domain

- What are your initial steps when you are presented with a subject to create a taxonomy for?
- What's your process for looking for relevant documents for the subject?
- When presented with unstructured information, how do you initially approach understanding the content?
- What criteria do you use to identify key concepts or entities within the information? Do you use any tools for this?

Conceptualization Categorizaton

- How do you determine the categories and hierarchical structure of concepts within a taxonomy?

Validation criteria and methods

- What do you call an effective or a complete taxonomy?
- What are some validation methods you use to check the effectiveness and completeness of a taxonomy
- What are the kinds of biases that could arise in a taxonomy? Is there a way to already know a true taxonomy for a specific subject?

Collaboration with stakeholders

- At what points do you collaborate with domain experts
- How do you incorporate feedback from various stakeholders into the taxonomy creation process?

Thanking for the participation and closure.

B.3 Interview with Taxonomy Expert - Semantic Codes

Quote	Semantic code
”The taxonomy is fundamentally a classification system ”	Definiton of taxonomy - classification system
”get pedantic about this.”	Pedantic Nature of Task
”get pedantic about this. Come up with the weird edge cases.”	Meticulous classification of concepts within domain.
”Look the domain what shape is it? Where is this edge?”	Identify shape and edge-cases of domain
”You basically start with all things.”	Begin with all the concepts within a domain
”And then you say well, what are sub subclasses of things? OK, these are subclass. What are subclasses of this one? Ohh those now.”	Continuous sub-classification
”coming up with a in a sense, a consistent set of rules for how you'll do it”	consistent logic for subdiving classes
”by stress testing the model . just just stress. Test it from every possible angle. Get others to stress test it.”	Stress testing the evolving model of taxonomy
”Are there actually any gaps between them?”	Revise to identify gaps and fill them
”that gives you is necessary every one of those block areas must be covered and it's sufficient because it's absolutely no overlap between them.”	Good taxonomy criteria : coverage and no overlap
”As much as you can and then some. Because there's there's only too much from the perspective of they get bored of it and fed up with it and don't want to participate.”	Importance of collaboration with SME

Table B.1: Semantic codes of quotes from the interview with a taxonomy expert from IKEA

Appendix C

Interior Design Insights Annotations Generation

C.1 Full Prompts and Outputs for examples of initial attempts

C.1.1 Direct Instruction to the VLM to generate insights for all the defined aspects

C.1.2 Question style prompting to the VLM to generate insights for all the defined aspects

PROMPT TO THE VLM:

Provided a list of aspects of interior design to analyse this space with, your task is to generate insights about the room image for each of the aspects.

Here is the list along with definitions and example values that these aspects could potentially have :

[redacted]

Style
Definition: The aesthetic and design theme of the room.
Examples: Modern, Minimalist, Rustic, Traditional, Eclectic.

Colors
Definition : The specific colour palette seen in the space and a description of the tone, colour-theme.
Examples : predominantly Yellow, Green accents, turquoise accessories, Beige, neutral tones, soft pastels, bold primary colours, etc.,

[redacted]

Storage
Definition: Solutions and systems for organizing and storing items within the room. This includes subcategories like Built-in Storage: Wardrobes, Shelves; Freestanding Storage: Chests, Cabinets; Innovative Storage: Under-bed drawers, Modular storage systems.
Examples: Modular storage units, Built-in wardrobe, Storage boxes under the bed.

[redacted]

RESPONSE FROM THE VLM:

[redacted]

Furniture
Definition: The movable objects intended to support various activities within the room.
1. Single bed
2. Desk
3. Transparent chair

Style
Definition: The aesthetic and design theme of the room.
1. Modern
2. Minimalist
3. Educational

Colors
The specific color palette seen in the space and a description of the tone, color-theme.
1. Neutral tones
2. Earthy hues
3. Bright accents

[redacted]

Storage
Definition: Solutions and systems for organizing and storing items within the room.
1. Under-bed storage
2. Desk drawers

[redacted]

Figure C.1: An initial attempt to generate interior design insights with VLM using direct instructions for all aspects in the conceptual framework. Definitions and examples were included.

PROMPT TO THE VLM:

Answer the following questions about the room :

[redacted]

Furniture
What are he movable objects intended to support various activities within the room?Examples: Upholstered bed with storage, Wooden desk, Metal chair.

Style
What are the aesthetic and design theme of the room?
Examples: Modern, Minimalist, Rustic, Traditional, Eclectic.

Colors
What is the specific colour palette seen in the space and the tone, colour-theme?
Examples : predominantly Yellow, Green accents, turquoise accessories, Beige, neutral tones, soft pastels, bold primary colours, etc.,

[redacted]

Storage
What are the solutions and systems for organizing and storing items within the room? This includes subcategories like Built-in Storage: Wardrobes, Shelves; Freestanding Storage: Chests, Cabinets; Innovative Storage: Under-bed drawers, Modular storage systems.
Examples: Modular storage units, Built-in wardrobe, Storage boxes under the bed.

[redacted]

RESPONSE FROM THE VLM:

Based on the provided image, here are the detailed observations of the room:

[redacted]

Furniture:
A bed with a simple frame and striped bedding.
A modern, transparent chair, likely made of acrylic or plastic, at the desk.
A small desk next to the bed, possibly used for studying or working.
Style:
The room has a modern and minimalistic style. The clean lines, simple furniture, and the world map mural contribute to a contemporary look.

Colors:
The color palette is neutral with pops of color. The bedding has shades of green, orange, and beige, while the world map mural adds a variety of colors.

[redacted]

Storage:
There is no visible storage furniture, but the bed might have under-bed storage.
The desk may have drawers for organizing smaller items.

[redacted]

Since the answers were descriptive, a following prompt to extract only the key insights out of the descriptions was sent as shown below :

Convert your answer into the following format?
{ Concept : List of strings }

Sure, here it is:

[redacted]

Furniture:
1. Bed
2. Acrylic chair
3. Desk

Style:
1. Modern
2. Minimalistic

Colors:
1. Neutral palette
2. Green accents
3. Orange accents
4. Beige

[redacted]

Storage:
1. Under-bed storage (implied)
2. Desk drawers (implied)

[redacted]

Figure C.2: An initial attempt to generate interior design insights with VLM using Question and Answer style of prompting for all aspects of the Conceptual Framework. Examples were included. This attempt also included a second stage to convert descriptive answers into key insights

Appendix D

Workshop With Experts

D.1 Information Sheet Using Gen AI to help organise Interior Design Knowledge

Hello, I invite you to this research activity as a part of a master thesis study in collaboration with IKEA. I am Sandhiyaa Balasubramanian Yamuneswari pursuing the Interaction Technology Master's Program at the University of Twente. Please read the following details about the study carefully to make an informed decision on whether to take part in this workshop. Please ask questions if you have any.

Overarching Research Aim Exploring the potential of Generative AI, particularly Large Language Models (LLMs) and Vision-Language Models (VLMs), to enhance workflows related to Interior Design

Purpose of this Activity
The aim of this workshop is to learn about Interior Design principles as practiced within IKEA, to elicit best practices in the field and the company, and to understand how professionals like you make interior design decisions.

The Session
Participating in this workshop involves approximately 1 hour of your time answering questions related to your interior design expertise and engaging in discussion regarding your answers with your co-participants of this workshop. An infinite canvas, Miro, will be used to facilitate this session.

Possible Risks & Benefits
The topic of this workshop should not cause any discomfort. Moreover, the research has been reviewed by the Ethics Committee of the University of Twente, ensuring that the study will be conducted ethically and with the utmost respect for participants. There is no compensation for participating in the study.

Data Privacy & Confidentiality
If consent is given, the workshop will be recorded and transcribed. If not, the responses will bemanually-noteddown. Thetranscriptions/noteswillbeanalyzedaspartoftheresearch process. The transcriptions maybe be stored locally and securely until the research is complete, after which they will be promptly and permanently deleted from any form of my (the researcher's) storage. The recordings and the transcriptions, however, maybe be stored securely by INGKA. The responsibility of the artifacts of the workshop will then completely belong to INGKA.
The transcription tool used will be locally run, ensuring that it will not be shared in any third-party platforms. To uphold confidentiality, all personal details will be anonymized. The final will not contain any personally identifiable information, and anonymous quotes may be utilized to support specific conclusions. Upon request, or as appendices in the final document, anonymized transcripts may be made available, trimmed of any proprietary company information. The final thesis report will be publicly accessible, and there is a possibility of its use in research publications within academic circles.

Participation
Participation is completely voluntary. You have the right to refuse participation, the right to refuse any

question and you can withdraw from the study at any point without any consequence and without providing any justification. After the workshop, should you decide that you do not want your responses to be included in the study, you have the option to request the removal of your answers and any analyses derived from your them any time before June 2024. The results will then be integrated into the research, after which it will not be feasible to retrieve the portion of results influenced by your answers and remove it.

Researcher Contact

If you have any further questions, feel free to ask me now, or contact me afterwards at
sandhiyaa.balasubramanian.yamuneswari@ingka.ikea.com

Additionally, you can contact the Secretary of the Ethics Committee of the Faculty of Electrical Engineering Mathematics and Computer Science at the University of Twente through
ethicscommittee-cis@utwente.nl.

Once you have read the above information and have decided to take part in the study, please convey your consent verbally by answering to the following two questions:
Are you sufficiently informed and do you agree to take part in the study? Do you mind if I recorded this meeting?
Thank you

D.2 Snippet from Workshop

Figure D.1: A snippet from the workshop with interior design experts. The experts performed the task of analysing images to give their interior design insights. This snippet is for one image. The experts performed this task for 3 images

Appendix E

Image Annotations

E.1 Sample data of image and corresponding annotation

Bedroom Image	[redacted]	Style	Colours	<other 6 aspects>	Storage
	[redacted]	'warm', 'inviting', 'traditional', 'modern', 'neutral'	'Various shades of brown', 'Light beige on the floor', 'Medium and dark browns on surfaces', 'Accents of green present', 'Earthy and neutral color scheme', 'Potential to balance with muted blues', 'Potential for subtle oranges'	... <insights of other 6 aspects > ...	'Drawers in nightstands', 'Multiple drawers in dresser', 'Additional drawers in tall chest', 'Tray on dresser for organization', 'Potential under-bed storage', 'Potential storage bench at foot of bed', 'No visible closet or wardrobe'

Table E.1: Sample data of image annotation for 4 aspects of interior design

Appendix F

Post Evaluation Interview with the Experts

F.1 Interview Guide

Introduction
Thank you for taking the time to review and evaluate the interior design insights generated by our AI technology. Your expert evaluations have been instrumental in understanding this technology's capabilities.

Based on your evaluations, I have analyzed the results and compiled the data. (Describe the quantitative evaluation results)

Now, I'd like to ask you a few questions regarding your experience evaluating the AI-generated insights. This interview is part of my thesis, so I need your official consent to record this meeting and use the insights from this meeting in my report.

I would like to point out that, to uphold confidentiality, all personal details will be anonymized. The final will not contain any personally identifiable information, and anonymous quotes may be utilized to support specific conclusions.

Could you please verbally confirm if you consent to this?

QUESTIONS:
General Experience

- How was your experience evaluating? Did you encounter any challenges?

Incorrectness

- In order to understand if I conveyed the criteria well, how did you make the decision to mark something incorrect?
- The data shows that aspects like [redacted] and 'Style' had a higher percentage of incorrect values. Can you elaborate on your observations regarding these aspects?
- Can you recall any patterns or commonalities among the incorrect values?

Incompleteness

- In order to understand if I conveyed the criteria well, how did you make the decision to mark something incomplete?
- Asking the same for Incompleteness, the charts indicate high frequency of incompleteness in aspects like [redacted] Can you recall why that is the case? Do you remember any observations regarding the incompleteness of these aspects?
- Similar to incorrectness, can you recall any patterns or commonalities among the incorrect values?
- Did you face any challenges in making decisions about an aspect's incompleteness?

Aspect Specific Insights

- How did the AI handle subjective aspects like 'Style' and [redacted]? Were there any instances where the AI's interpretation matched or clashed with your professional judgment?
- And the Objective aspects like [redacted] etc?
- Do you recall any interesting observations about any particular aspect?

Other Insights

- Imagine if a human interior designer performed this same task of Interior Design analysis of a room, for they key aspects, providing insights in the same way of bullet points, how do you think their insights would differ from the AI generated insights?
- What do you think about the AI's knowledge of the field of Interior design?

Open Feedback

- Are there any other observations you would like to highlight regarding the AI's performance or the evaluation process?

Closing:

- Thank you so much for your time and valuable feedback. This is the last collaboration for my thesis specifically, but if we learned anything from the results it is that the collaboration with domain experts is a crucial on-going part of the AI development life cycle. So, I believe the need for your feedback and insights is only going to increase henceforth. Hopefully this process is interesting to you as well. Thanks for all your input so far and have a lovely rest of the day!

F.2 Post Evaluation Interview Semantic Codes

Participant_ID	Quote	Semantic Code
Participant 1	"The overall experience to me, I evaluated the images was a bit daunting to be honest. It felt like very repetitive.","Repetitive and Daunting Experience"	Repetitive and Daunting Experience
Pariticipant 1	"It hadn't like a lot of items in each subsection, and a lot of the times it felt like very technical."	Technical Language
Participant 1	"It would have been easier to perhaps look at the image and actually write things down."	Enthusiasm to offer richer Feedback
Participant 1	"For example, I remember the part about sound proofing or something like that, but that was like way, way too technical for what we are normally used to considering in the work we do."	Overly Technical Aspects
Participant 1	"When I don't see the things in the picture, I marked it incorrect."	Visual Confirmation for Correctness
Participant 1	"If it said something like traditional or cosy, and then it was, the picture was more modern like the style was modern, then we marked it as incorrect."	Insight mismatch for incorrectness
Participant 1	"If we said sloping ceiling or window, but I didn't see anything, then I said it, there is no sloping ceiling or window."	Obvious omissions for incompleteness
Participant 1	"There was that those were marked right, like there was a it it was always written ohm."	Correctness in Improvement Suggestions
Participant 1	"In the case of incompleteness, I can't really remember, but it could have been, for example, this that one I remember very well where there was like a sloping ceiling."	Recalling difficulty for specific examples
Participant 1	"I think it felt a bit more like a customer or someone with little interior design knowledge was evaluating a space and saying this room is cosy or this room is if I would ask my aunt for example, how would you describe this room?"	Layman Perspective
Participant 1	"I think it performed better in objective elements than in subjective."	Better Performance in Objectivity

Participant_ID	Quote	Semantic Code
Participant 1	"I think AI complicating things a bit more than a person have done."	AI Over-complication
Participant 1	"I think it would have probably cooked about more emotional aspects and subjective things rather than saying, oh, there's a bed here because I think certain things are obvious."	Hallucinations related to emotional and subjective aspects
Participant 1	"I think it would also be nice to have, you know, like the feedback session, it it can be like 5 days afterwards, that's fine."	Indication of post evaluation interview required
Participant 1	"If it said something like traditional or cosy, and then it was, the picture was more modern like the style was modern, then we marked it as incorrect."	Style contradiction
Participant 1	"I'm not sure. Yeah, how that would affect. The outcome of this, but otherwise. Yeah."	Uncertainty about application of the results
Participant 1	"I would say maybe also five or six, yeah."	Mid-level domain knowledge rating
Participant 1	"No, not excellent. Excellent. That's super happy with this."	Satisfaction with feedback platform in comparison to traditional methods
Participant 1	"Maybe that's the only thing that maybe, as I said, I don't, I don't know how relevant that is."	Uncertainty about relevance of details
Participant 1	"It doesn't make so many mistakes or you know, "	Overall good performance
Participant 1	"I was very good at quite good at [redacted] even if it was a bit low number here, but actually understood that it was lacking [redacted] or those kind of things."	Positive Performance on [redacted]
Participant 1	"It performed poorly when it came to the time lower the country or sometimes they would say there's a there's a picture feature fixture in the ceiling and there wasn't anything because you could actually not see this thing."	Incorrect Identification of Architectural Elements
Participant 1	"And I also felt with [redacted] although it was very, very technical, it was also quite good as well marking or, you know"	Positive Performance on [redacted]
Participant 1	"It sounds more like a task when you're at school, and maybe you're a 10 or 8 years old and you have to analyze things and then you're just writing bullet points and then you repeat a lot of things because maybe your brain is not so evolved your your rational thinking or whatever or analytical thinking."	Repetitive terms used in annotations
Participant 1	"if you have a set of 20 spaces to analyze, you will get bored of saying the same thing with all the time"	Lack of creativity of an expert

Participant_ID	Quote	Semantic Code
Participant 2	”I think also the language is not language that we are, you know, connecting home furnishing to its much more technical language.	Technical Language
Participant 2	”Then I also miss maybe this soft feeling of the home furnishing. Yeah. Yeah, being more emotional part.”	Lack of emotional touch to the language.
Participant 2	”Also a lot of words that not IKEA, you know more this simple two things to express things I think that was very complicated in some way actually even if you're quite good at English anyway.”	Non-IKEA Terminology
Participant 2	”Lighting was very more technical. I could say there was a lot of phrases and things around that, was also not maybe relevant.”	Irrelevant technical details
Participant 2	”I can remember a lot of times where, for example, window solutions I would say that there was some pattern or some roller blind and there wasn't anything.”	Incorrect [redacted]
Participant 2	”I think also this [redacted] and those kind of things also [redacted] was very more technical.”	Insights for [redacted] too technical
Participant 2	”I felt that I very not very often actually felt this is not completed.”	Rarely felt incompleteness
Participant 2	”I think the objective elements it performed better, even though sometimes it miss certain things or maybe the message was a bit blurry.”	Better performance in objective elements
Participant 2	”I felt it was quite good at color then I think there was a very many similar so to say, color, yeah, it was very the only pictures was very similar in the way.”	Good Color Interpretation
Participant 2	”But I would agree.Maybe six or something like it performs well in my technical very technical things.”	Moderate domain knowledge
Participant 2	”So I think it was not about, you know, sockets and technical all this stuff. They are both in the roof, both in the wall. It said there was, but there was nothing.”	Incorrect Technical Details - [redacted]
Participant 2	”I really like this picture and this simple things.”	Positive Feedback on the platform

Table F.1: Semantic codes of quotes from the post-evaluation interview with interior designers from IKEA