Optimizing Retrieval Quality in Reproductive Health RAG Systems: A Comparative Study of Chunking Strategies and Embedding Models

CHRISTINA BRAUN, University of Twente, The Netherlands

Retrieval-augmented generation systems have shown potential in a healthcare context. However, response quality is crucial and must be safe, accurate, and reliable. A major factor for high quality responses is the retrieved context. Especially chunking strategies and embedding models are a key factor in this stage. The study examines which combinations of chunking strategies and embedding models provide the most reliable results through a factorial experiment. The considered chunking strategies are fixed-size chunking and sliding window chunking both with 256, 512 and 1024 tokens, as well as proposition chunking, semantic chunking, and document-structure-based chunking. The embedding models tested are a OpenAI, Qwen3, Ollama, BioMedBERT, SapBERT, MedEmbed, MedCPT and BioLORD. A basic RAG pipeline was built that asked the same 46 questions to all combinations of chunking strategies and embedding models. The questions are based on an ESHRE guideline paper concerning ovarian stimulation for IVF [11]. This was also the document that was chunked and embedded. The retrieval quality was evaluated with MRR, ndcg@k, recall@k, content overlap, contextual precision and contextual recall. The findings show that while the variables have no combined effect, they individually affect retrieval quality and that embedding model choice has a greater impact.

Additional Key Words and Phrases: IVF, Healthcare RAG, Chunking Strategy, Embedding Model, Retrieval

1 INTRODUCTION

For patients undergoing In Vitro Fertilization (IVF), a complex multistep procedure, access to accurate knowledge can directly affect health outcomes [18]. The steps involve suppressing natural hormone production, boosting egg supply, egg collection, and embryo transfer [16]. The recommendations given by clinical staff must be followed carefully, particularly regarding timing and using medications properly [16]. The procedure can be physically and mentally demanding for patients and their close social circle, and many questions can come up throughout the process [27].

Given these information needs, some patients may try to find answers via online platforms. However, a study by John et al. (2025) found that there is an increasing amount of unreliable information circulating about women's reproductive health online on topics including fertility, contraception, and chronic conditions. Often the misleading content contains contradicting advice and incorrect recommendations such as claiming higher risk for certain medical interventions without medical data, recommendations that contradict standard medical practice or exaggerating known risks [18]. Other patients turn directly to medical staff via email or phone to find trustworthy answers [2]. However, not all questions posed are urgent and could often easily be answered with a single follow-up meeting [2]. In clinics the amount of direct contact via phone or email translates into a significant workload that consumes time that would otherwise be allocated to direct patient care [2].

A potential solution to relieve staff workload caused by patient questions, while still addressing patient needs with reliable information, are large language models (LLMs). While LLMs have gained considerable attention in recent years for answering complex questions, there are risks that come with it especially in healthcare. As the paper by Sedrakyan et al. (2025) discusses, there are risks concerning data security and algorithmic issues such as hallucinations, provider related risks such as deskilling of staff, patient related risk such as overconfidence in AI answers, and systematic risks such as reduced innovation [28]. Additionally, LLMs do not always clearly indicate what their answers are based on [1]. This limits their suitability and readiness in serious healthcare environments. At the same time, if responsibly designed within a human-centric framework such as the HiCARE framework, AI systems in healthcare have the potential to reduce clinical workload and simultaneously aid patients with trustworthy information [28]. In contrast to traditional pretrained models, retrieval-augmented generation (RAG) systems have become more interesting in healthcare, as they can answer questions based on a given set of documents such as clinical guidelines and can access external knowledge directly [24]. The responses of such a system are based on the query and the retrieved content. RAG systems reduce inaccuracies, hallucinations, and improve transparency, making them more suitable for answering patient questions and compatible with the HiCARE framework compared to standard AI models [24]. Problems that remain in RAG systems are retrieval noise (irrelevant or low-quality retrieved data), domain shift issues (the system is asked about something it has no information on), and generation latency (the time it takes to answer) [24].

Two major factors that influence retrieval quality are the chunking strategy and the embedding model with which the chunk are embedded [14][4]. This is how source documents are segmented before being passed to the LLM and how they are represented as numerical vectors necessary for retrieval. Research has been conducted on optimizing embedding models and on comparing chunking strategies, but a gap remains regarding how they jointly influence retrieval quality, especially in a safety critical domain such as healthcare. This study attempts to address this gap by comparing combinations of chunking strategies and embedding models. The RAG pipeline is applied to an ESHRE guideline document concerning ovarian stimulation for IVF [11] and evaluated via seven retrieval metrics.

2 PROBLEM STATEMENT

While prior research has studied chunking strategies and embedding models in isolation, there remains a gap on how the two variables act in combination on retrieval quality for reproductive healthcare systems. Especially in women's reproductive health, it is important for patients to seek answers to health questions via reliable sources

TScIT45, July 3, 2026, Enschede, The Netherlands
© 2026 University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

as they are already exposed to misinformation online [18]. For clinics, a RAG-based system would mean a reduction in workload. For these reasons it is valuable to research how chunking strategies and embedding models act in combination on retrieval quality.

2.1 Research Question

Main Research Question

- How do different combinations of chunking strategies and embedding models affect RAG retrieval quality for a patient guidance system in reproductive healthcare?

Sub-questions

- To what extent do combinations of chunking strategy and embedding model affect retrieval quality in combination rather than through just one of the two factors?
- What is the relative impact of chunking strategy versus embedding model on retrieval quality?
- How do different chunk-model combinations perform across different evaluation metrics?
- How do chunk-model combinations perform across different categories of patient questions?

3 RELATED WORK

3.1 RAG Pipeline

Retrieval-augmented generation (RAG) is a method for generating text based on a specific set of documents using a large language model [22]. Documents such as PDFs, Word files, and HTML pages are provided to the RAG system. First, they need to be converted into plain text. Then they are split into segments called chunks [13]. A chunk can be created in multiple ways, for example dividing the text into sets of 100 words [22]. Chunking is necessary because LLMs can only process a certain amount of text at a time [13]. After the text is chunked, each section is embedded and stored in a vector database. When a user asks a question, it also is embedded using the same model as for the chunks. The similarity between the user vector and the chunk vectors is calculated and the top K chunks are retrieved [13]. The relevant chunks and the user question are combined, and the LLM is asked to answer the question based on the retrieved chunks [13].

3.2 Chunking Strategies

There are various chunking strategies that can be used for RAG systems. The simplest version, introduced by Lewis et al. [22] is fixed-size chunking which, segments a document into a specific number of tokens or words. This chunking strategy is relatively ineffective, as sentences or words can be divided in the middle [14]. Other strategies divide by paragraphs [29]. This approach has demonstrated strong performance across multiple use cases. In the context of healthcare, adaptive chunking has been found to perform at a high level [14]. Although adaptive chunking is mentioned in different papers, there are different approaches. While one computes similarity sentence by sentence [14], another divides by content density [29]. A study also found that proposition and semantic chunking strategies demonstrated strong performance in a clinical context [14]. Proposition chunking breaks content into atomic facts via an LLM so that sentences can stand by alone and do not need any more additional context [5]. Semantic chunking splits text by meaning by embedding sentences and comparing the similarity to the following sentence [14]. Lastly, another successful chunking strategy is dividing by document structure such as headings, figures, and paragraphs. This approach is named different across papers: some call it hierarchical chunking [29], while others call it element-based chunking [30]. Overall, there is no single highest chunking strategy, as it depends on what type of document is chunked, how long a text is, and what it is needed for.

3.3 Embedding models

Embedding models are neural networks that convert text or images into vectors [19]. These vectors have many dimensions, which range from a few hundred to thousands of dimensions, depending on model [19]. Vectors are calculated so that similar concepts have closer vectors than concepts that are less related [19]. This is necessary so when users pose a query to a system, the query is embedded, and closer vectors can be found. These vectors are more likely to contain text related to the question. Embedding models vary in size and training data. Some may be trained on general data [31], while others are trained only on medical documents [9], [17] making them highly specialized on one domain. Models may also vary in language depending on training corpus [31]. Some models use different encoders for embedding queries and for texts. An example of such a model is MedCPT [17]. This may be a model more adapt to bridge the gap between simple user language and complex medical guidelines.

3.4 Evaluation Metrics

Retrieval-Augmented Generation Assessment (RAGAS) is a widely used evaluation framework that works without human manual checking, allowing faster evaluation cycles of RAG architectures [10], [10]. RAGAS incorporates a functionality to generate questions-answer pairs to test RAG retrieval and generation. Other popular evaluation metrics included LLM-as-a-judge, where content is provided to an LLM and it is asked to score it with focus on a specific aspect [7]. DeepEval is a framework that provides LLM-as-a-judge metrics. For retrieval specifically there exists contextual recall, which evaluates how well the retrieved context aligns with the expected output[7], and contextual precision which measures how well relevant content is ranked compared to irrelevant one [6]. For evaluating retrieval quality specifically, there a four common metrics. Recall@K, Top-1 Precision (also called Precision@K), MRR, and NDCG@K. The first, is a metric that calculates the fraction of the relevant chunks contained in the retrieved chunks. [12]. R_C is the relevant chunk, and Top_k is the top-k retrieved chunks.

$$\text{Recall@K} = \frac{|R_C \cap Top_k|}{|R_C|} \tag{1}$$

Top-1 Precision measures if the top ranked chunk from the RAG system is also a relevant chunk [12]. The score is 1 if it is and 0 otherwise.

Mean Reciprocal Rank (MRR) measures how high the most relevant chunk is ranked on average [12]. If five chunks are retrieved for a query and only the third is relevant, then the score becomes

1/3. The mean score is then calculated over all queries. $|Q|$ is the number of queries and rank_i is the rank position of the first relevant document for the i-th query [12].

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{2}$$

NDCG@K is used when the order of retrieved chunks matter. It compares the ideal rank to the retrieved rank [12]. DCG@k scores the retrieved ranking at position k. IDCG@k is the DCG@k for the ideal ranking. rel_i is the relevance score at position i[12].

$$\text{NDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}} \tag{3}$$

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)} \tag{4}$$

All metrics give scores between 0 and 1 where 1 is the highest.

4 DESIGN AND IMPLEMENTATION

To evaluate retrieval quality for combinations of chunking strategies and embedding models, 46 questions were posed to all 72 combinations of nine chunking strategies and eight embedding models. Retrieval was tested on a single medical guideline by the European Society of Human Reproduction and Embryology (ESHRE) covering ovarian stimulation for IVF/ICSI [11].

The evaluated chunking strategies comprised fixed size chunking as a baseline model, with 256, 512 and 1024 tokens. These sizes are a common benchmark to assess small, medium and large chunks. Sliding window chunking with the same token sizes was included to address the boundary-splitting limitation of fixed-size chunking. Also, document-structure-based chunking was implemented. This strategy was programmed with the help of Claude AI, for detecting logical boundaries in the document. Claude AI was chosen because of the LLMs document understanding, allowing consistent and less time-consuming segmentation of the document structure. The code was then reviewed by the researcher. Propositional chunking was implemented by first creating 1000-token chunks, and then prompting GPT-4o-mini to extract atomic propositions. This strategy was evaluated to test whether small chunks with one fact, combine favourably with certain embedding models. Given that clinical documents such as the ESHRE guideline are densely packed in information, representing chunks in facts may facilitate precise embedding. The final tested strategy was semantic chunking, as text is split by meaning rather than a fixed size. In a medical context it appeared promising as different recommendations are given.

Embedding models were selected to cover a broad spectrum of general-purpose and medical models. The general models were OpenAI's text-embedding-3-small (hereafter OpenAI) [25], serving as a strong commercial baseline, Qwen3-embedding:0.6b (hereafter Qwen3)[31], an open-source model with high performance on the MTEB leaderboard [31], and mxbai-embed-large (hereafter Ollama) [21], a model used within the Ollama environment. The medical models included abhinand/MedEmbed-large-v0.1 (MedEmbed) [9], ncbi/MedCPT-Article-Encoder (MedCPT) [17], cambridgeltl/SapBERT-from-PubMedBERT-fulltext (SapBERT) [23], FremyCompany/ BioLORD-2023-C (BioLORD) [26] and microsoft/BiomedNLP-BioMedBERT-base-uncased-abstract-fulltext (BioMedBERT) [15]. These models were all pretrained or fine-tuned on medical and biomedical text, making them promising models to test on clinical guidelines.

The ESHRE document was selected as it comes from a reliable source, it covers an important aspect of reproduction, and it contains clear recommendations with underlying justifications, allowing the RAG system to base its answers on them. Further, the document has a significant length that allowed the creation of a data set of 46 question-answer pairs, with answers grounded in the text. The dataset can be found in the Appendix. For each question, one to two relevant text sections were added, to later evaluate if they can be found in the retrieved chunks. The length of the relevant text sections varied per question. The questions were divided into five categories to assess retrieval across varying question phrasing. The first category are abbreviation questions which ask for the full definition of an abbreviation. The second category are abbreviation-explained. These questions asked for the meaning of abbreviations with additionally an explanation of the concept in relation to ovarian stimulation. Fact questions aimed to test how retrieval performs on explicitly stated information related to ovarian stimulation. Patient questions are formulated to simulate patient phrasing. They often contain background information of the patient, such as "I am currently in the luteal phase of my menstrual [...]". Numerical questions aimed to evaluate how chunking strategies and embedding models capture numbers, and how it effects retrieval. For all question categories 12 questions-answer pairs were made with the exception for abbreviation and abbreviation-explained questions. Here only five were made due to their repetitive format. For the other categories, more complex questions could be asked with a greater variety in phrasing, hence more questions were made.

The RAG pipeline was developed by extending a simple baseline implementation [20] into three code files to meet the requirements of the experiment. The first file extracted all text from the source document, cleaned it by collapsing white space and rejoining hyphenated line breaks, and iterated over all chunking strategies to split the ESHRE guideline. The chunks were saved as txt files to allow reading them later. Then each set of chunks was embedded with every embedding model resulting in 72 combinations. ChromaDB was used as the vector database, as it is open source and designed for similarity search. One ChromaDB collection was created per strategy-embedding combination.

The second code file implemented the retrieval and generation phase of the experiment. On all 72 combinations stored in ChromaDB, the 46 questions were asked. The questions were embedded with the same model that the chunks of the tested combination were embedded with. Then the three most relevant chunks were retrieved from the collection. This number was chosen to test retrieval with minimal chunks. Retrieving more chunks increases the likelihood of relevant text appearing in the retrieved chunks by chance and inflating scores. Additional noise also decreases response quality later in the pipeline [8]. The chunks were retrieved with cosine similarity and passed to the LLM alongside the question and a system prompt, instructing to strictly answer on the provided context, to behave

empathetically and use medical terminology where appropriate. The LLM that was used was OpenAI based on its extensive documentation and prior research on its performance [3]. All information from the retrieval was saved by combination in JSON format. This included the query question, the ground truth, the answer from the LLM, the three retrieved chunks with their id, text, and metadata.

The third file applied both deterministic and LLM-based retrieval evaluation metrics on all combinations. Deterministic metrics comprised Recall@k, Top-1-Precision, MRR, nDCG and content overlap. The last metric was not derived from literature. It calculated the ratio of overlap between retrieved chunks and the relevant text, in order to assess the quantity of retrieved irrelevant text. The other metrics applied the formulas as mentioned in the literature. All deterministic metrics compared the retrieved chunks to the relevant text from the question-answer dataset from the ESHRE guideline. All metrics gave a score from 0 to 1 to simplify comparison. The LLM-based metrics were contextual recall and contextual precision from the DeepEval framework. This was chosen because of its ready to use implementation and to contrast the deterministic metrics with an alternative perspective. The deterministic metrics produced exact, reproducible scores while the LLM-based metrics captured semantic relevance. RAGAS was excluded as an evaluation framework as question-answer pairs are generated with an LLM may not understand medical concepts adequately to generate relevant questions. Further, the questions cannot be divided into categories as it was done in this experiment. Additionally, the study measures retrieval quality only while RAGAS focuses on both retrieval and generation, and excludes retrieval specific metrics such as the deterministic metrics ultimately used. After evaluation of retrieval quality, all scores were exported to a CSV file to facilitate analysis. The csv file with the experiment result was cleaned and various tables were created that can be found in the Appendix. Throughout the experiment, all variables such as the vector database, number of retrieved chunks, and retrieval method remained the same. Only the chunking strategies and embedding models were varied.

5 ANALYSIS

5.1 Chunking

Sliding window chunking with 256 and 512 tokens achieved the highest mean scores across all metrics (sliding 512 = 0.4404, sliding 256= 0.4400), while propositional chunking (0.259) and semantic chunking recorded the lowest (0.311) (Figure 1). The last two strategies show a larger performance gap to the next lowest mean which is fixed 1024 (0.3928). All sliding window chunking strategies had higher mean scores than all other strategies. The next highest score comes from structured chunking (0.423). It achieved a higher result than all fixed size chunking. Both sliding window chunking and fixed size chunking with 1024 tokens, attained lower mean scores than medium or small chunks. For both sliding window and fixed size chunking, 512 tokens recorded higher mean scores than 256 and 256 performed higher than 1024. Table 11 breaks this down further by visualizing how different chunk sizes performed per metric. For four out of seven metrics 512 token size outperformed the others. Although a size of 256 tokens recoded lower mean scores for four out of seven metrics, when compared with other token sizes, it nonetheless achieved the second highest mean across all metrics, due to its higher mean performance on contextual precision (0.729) and content overlap (0.091). 1024 tokens achieved a substantially greater mean on contextual recall than it did on other metrics yet performed poorly on metrics for which 256 token chunks recorded the highest scores. Overall, 1024-token chunks yielded the lowest mean performance. Sliding window chunking outperforms fixed size chunking on all metrics except for content overlap (Table 12). There it scored marginally lower than fixed size chunking (sliding = 0.926, fixed = 0.927).

Propositional chunking performed lowest on each metric except for content overlap (Table 1). Semantic chunking performed second lowest and showed a similar pattern to propositional chunking with its highest result for content overlap. The top 4 chunking strategies are constantly green for each metric with only a few exceptions. These are for content overlap where structured chunking scored lowest across all strategies and sliding 1024 was second lowest. Smaller chunking strategies such as sliding 256 and sliding 512 scored slightly higher but nonetheless in the yellow area. Also, for contextual recall smaller chunking strategies had a higher mean. A notable difference is the lower performance of sliding 256 for top 1 precision, indicating it performs lower on ranking the most important chunk as first position than the other highly scoring strategies.

Sliding 256 and fixed 256 are the highest performers on abbreviation questions (sliding 256 = 0.386, fixed 256 = 0.378) (Table 2). The mean score of fixed 256 drops slightly for explaining abbreviations (0.373) but for sliding 256 it increases (0.464) achieving one of the two highest means for this category alongside sliding 512 (0.447). All three sliding window strategies demonstrated strong performance on factual questions. Multiple chunking strategies perform similarly well on patient questions. However, sliding 256 which was a high performer on all other categories performs only the fifth highest strategy (0.418). The other sliding window strategies alongside structured chunking and fixed 512 had a higher mean score. Structured chunking and sliding 1024 had the highest mean for numerical questions (0.410,0.409). Propositional and semantic score lowest across all question categories. There is no chunking strategy that constantly performs the highest across all question types. However, when looking at the question categories, factual questions achieved the highest mean score across all chunking strategies except for propositional and semantic chunking. Semantic scored only marginally higher on explaining abbreviations (abbreviation explained = 0.355, fact = 0.354). Propositional chunking scored higher on patient questions (0.341) than it did on fact questions (0.278).

While factual questions are the category attaining the highest means, abbreviation questions record the lowest (3). Textually heavy questions such as fact and patient questions achieve higher mean scores than short exact questions such as abbreviation and numerical questions. For the further analysis questions categories will be divided in two types. Exact value questions comprise numerical questions and abbreviation questions. Textual questions comprise factual questions with patient questions. Abbreviation explained is excluded from both categories, as it contains characteristics of both. Figure 3 shows that from the tested combinations of embedding models and chunking strategies, the system retrieves chunks most

effectively for textual questions, followed by abbreviation explained questions, and least effectively for exact value questions.

Sliding 256 ranks highest for exact questions but is the third highest for textual questions (Table 3). Propositional and semantic chunking rank eighth and ninth for both sides. Small chunking strategies such as sliding 256 and fixed 256 rank higher on exact questions compared to textual questions. Big chunking strategies such as structured and sliding 1024 rank higher on textual questions. For textual questions, sliding window strategies rank higher than all other strategies. While all strategies had higher absolute values for textual questions, when looking at their overall ranking sliding 512, sliding 1024 and structured chunking ranked higher for exact value questions. The biggest difference in mean comes from propositional chunking (exact = 0.152, textual = 0.310, difference = 0.158). The smallest difference comes from fixed 1024 (exact = 0.361, textual = 0.414, difference = 0.053).

5.2 Embedding

The most notable finding across embedding models is that all general-purpose embedding models outperform medically specialized models (Figure 2). The highest performing model is OpenAI (mean = 0.489) followed by Qwen3 (0.436) and Ollama (0.411). The lowest mean score is recorded for BioMedBERT (0.226), which has a substantial gap of 0.126 points to the second lowest mean from BioLORD (0.352). The exact mean scores per metric for each model can be found in Table 4. OpenAI is the leading embedding model across each metric with the exception of content overlap, for which BioMedBERT and BioLORD demonstrate higher means. The medical model that achieved the highest mean performance is MedEmbed (0.423). It surpasses Ollama on all metrics except relevancy overlap and contextual recall, resulting in Ollama ranking higher in the overall mean.

BioMedBERT consistently records the lowest mean scores across all question categories (Table 5). OpenAI achieves the highest means for each, except for abbreviation questions. Here, MedCPT attains the highest mean. All models achieved higher mean scores on factual questions and then patient questions. Six out of the eight models recorded the lowest means for abbreviation questions. Table 6 presents the ranking from highest mean (1) to lowest mean (8) on exact value versus textual questions per model. All models achieve higher means on textual questions. OpenAI ranks first on both categories and BioMedBERT ranks lowest. MedCPT has the biggest difference in rank (third on exact, sixth on textual), but the smallest difference in mean between the two categories (exact = 0.370, textual = 0.403). Also, Qwen3 has a jump between the categories. It ranks more favourably on textual question (second place) compared to exact questions (fourth place), yet exhibits the smallest difference in mean between the categories (exact = 0.361, textual = 0.505).

5.3 Combinations

This section analyses how combinations of chunking strategies and embedding models performed across different metrics (RQ3). In Table 16 the combinations are ranked from overall highest to lowest mean scores across all metrics. Three heatmaps were created to analyse the interaction between the two variables.

Table 7 is coloured by overall highest means. The top three combinations of the whole table are sliding 256 embedded with Qwen3, sliding 256 embedded with OpenAI, and structured chunking embedded with OpenAI. A clear outstanding observation is the low performance of BioMedBERT, which is dark red across all chunking strategies. There is no equivalent always high performing embedding model or chunking strategy. Green fields are not only found at the top or only the left of the table but are rather ordered in a diagonal line across the table. However, they are skewed more to the top rather than the left.

Table 8 is coloured by column and visualizes which embedding model is highest given a chunking strategy. OpenAI ranks among the top options regardless of chunking strategy. Qwen3 similarly achieves high mean scores, with the exception when combined with propositional chunking and semantic chunking. BioMedBERT consistently records the lowest mean scores across all chunking strategies. Ollama attains relatively high mean scores on six out of the nine chunking strategies, but mean scores are reduced when paired with larger strategies such as fixed 1024 and structured chunking.

Table 9, is coloured by row, and shows which chunking strategies perform best per embedding model. General models (OpenAI, Qwen3 and Ollama) achieve the highest scores with sliding window chunking, specifically sliding 256 and sliding 512. Medical models perform better with chunking strategies that produce larger chunks such as sliding 1024 and structured chunking. MedEmbed is the only medical model that achieves its highest performance with sliding 256 rather than structured chunking or sliding 1024. Sliding 512 demonstrates relatively high mean scores for both medical and general models. Propositional and semantic chunking are poor choices no matter what model, as are small fixed size chunking compared to larger fixed size chunks. The combination of fixed 1024 embedded with Ollama is a notably low-performing combination.

The relationship between embedding models and chunking strategies and their overall influence on mean scores is visualized in Table 17. This table shows the ranking from highest mean to lowest mean per combination. Both columns contain the same ranking, but they are coloured differently. The left column is coloured by embedding model and the right by chunking strategy. While we can see that there is more dark purple on the top of the right column and more brown and red at the bottom, there is not such a clear divide as there is in the left table. High performing models such as OpenAI and Qwen3 make up a big section of the top performing combinations and BioMedBERT creates a visible red block at the bottom of the left column.

Tables 13–15 list mean scores of all combinations per metric, question type, and exact versus textual questions respectively, including performance differences. To assess interaction effects statistically, multiple two-way ANOVA tests were conducted, which provided the results found in Table 18.

For all metrics except content overlap, both chunking strategies and embedding models had a statistically significant effect but embedding models always had a greater influence. This excludes impact on contextual recall, because no ANOVA test could be done for this metric due to the lack of variance in the results. For the metric content overlap, embedding models had no significant impact but chunking strategies had a very strong effect. The ANOVA test for

the impact of chunking strategies and embedding models on the overall mean shows that the embedding model choice has a medium effect and the chunking strategy a small effect on outcomes on retrieval quality overall. No combined effect was found between chunking strategies and embedding models on any of the metrics or the overall mean.

This is different when looking at how embedding models and question categories interact. A significant but small combined effect was found, which shows that certain combinations of embedding models and question category produce more favourable outcomes than others. Both embedding models and question type individually affect the overall outcome. Embedding model choice had a medium strong effect while question category had a small effect. When examining the effect of embedding models and chunking strategies on mean performance of abbreviation questions, both had a medium strong individual effect but chunking strategies had a greater influence than embedding models. No combined effect was found for the two variables on abbreviation questions. For all other question categories, the model matters more than the chunking strategy. No significant combined effect was found for any combination.

6 DISCUSSION

6.1 Key findings

The key finding from the results is that while both embedding models and chunking strategies have an individual effect on retrieval performance, they do not act in combination (RQ1). Their effects from both chunking strategies and embedding models add up but there is no significant clash or synergy between the two variables. Because sliding 256, sliding 512, OpenAI and Qwen3 are chunking strategies and embedding models that achieve the highest mean scores across all metrics, it is expected that their combinations rank among the highest scoring ones overall. As the ANOVA tests have shown, the choice of embedding model has greater impact on retrieval quality than the choice of chunking strategy (RQ2). An explanation is that even well-formed chunks that do not split sentences across chunks and contain all relevant content to answer a question will not be retrieved if the embedding model fails to effectively encode the text. This finding is consistent with Table 17, where the colour gradient is more pronounced for embedding models than for chunking strategies. Table 7 also confirms this finding as the green highlighted cells form a diagonal pattern concentrated in the upper-left quadrant and are slightly skewed to the top, suggesting that both variables influence retrieval quality but embedding models have a greater impact.

6.2 Embedding

All general-purpose models outperform all medical models, suggesting that the latter are less effective at understanding and representing the ESHRE guideline. The superior performance of general models likely comes from their broader training data. Medical models may be over-specialised for terminology-dense text. Among medical models, MedEmbed achieved the overall highest mean score. Similar to all general-purpose models, it produces the highest mean scores on retrieval when combined with small and medium sliding window chunks. The relatively stronger performance may be attributable to its training on more general medical text than the other medical models. All other medical models demonstrated higher scores with larger chunking strategies, possibly because of their training on terminology-dense data. A short chunk may not provide sufficient context for these models to understand and encode the text effectively. General models understand all types of text well. A shorter chunk contains focused information. For larger chunks, general embedding models need to compress more text into the same vector space which can dilute the information, ultimately negatively affecting retrieval quality. Especially Ollama shows a weak performance with larger chunks. A possible explanation is that among general-purpose models, it is a less effective model.

6.3 Chunking

Although embedding models impacts retrieval quality more than chunking strategies, chunking strategies remains an important factor. While no single strategy constantly achieved the highest mean scores across all question types and metrics, sliding window strategies show relatively higher means. This likely is due to the overlap preventing unfavorably context splits across chunks. Semantic chunking, which was expected to be a high performing strategy because chunks are made by semantic similarity, had the second lowest mean scores across all chunking strategies, and yielded even lower means than all fixed size chunking. The comparatively poor performance of semantic chunking may be attributable to lacking topic diversity in the source document, which is restricted to a single subject: ovarian stimulation. Consequently, large chunks are created. As shown in Table 10, the chunk size variance of semantic chunking is extremely large. The smallest chunks are two tokens long while the largest are over 7000. When embedding chunk of this size, much content is diluted, which explains the lower performance of semantic chunking. A more effective alternative is structured chunking. This strategy splits by logical boundaries in the document such as headers and paragraphs. It ranked highest after all sliding window strategies. Table 10 shows that structured chunking created the largest chunks on average and thus also the least. As noted before, the highest performing models preferred smaller to medium large chunks. This may account for structured chunking ranking fourth overall. A consideration that should be made is that structured chunking was specifically designed for the ESHRE guideline. If structured chunking is implemented on any other document the function must be changed as documents may have a different layout. If structured chunking is considered, it should be implemented with sufficient flexibility to detect boundaries across diverse medical document types.

Propositional chunking recorded the lowest mean scores relative to the other strategies, despite having demonstrated high performance in prior literature. The results can be explained through the experimental set up. All metrics except the LLM as a judge ones, evaluated if the relevant sentences were found in the retrieved chunks. Because propositional chunking rephrases text into standalone factual statements, exact sentences never appear verbatim in the chunks, systematically penalising this strategy and therefore the true retrieval quality of propositional chunking is not accurately reflect in the results. When looking at the performance of

propositional chunking on LLM as a judge metrics it nonetheless performs poorly indicating that this strategy may not be preferred in a medical context.

A key factor that appears to influence retrieval, is chunk size. While chunking strategies with 512 tokens achieve higher mean scores than 256 tokens and 1024 performs lowest, there are metric specific differences. One is the lower mean score of sliding 256 on top 1 precision. All other four leading strategies performed well here, suggesting sliding 256 is less effective at ranking the most relevant chunk at first position compared to the other high-scoring strategies. A possible reason for this finding is that sliding 256 produces many small overlapping chunks. The relevant information may be spread across chunks, thereby reducing the system's capacity to rank them accurately. Larger chunks showed higher mean performance, because relevant information is more likely to be contained in one chunk and not spread across multiple chunks.

The higher scores from medium and small chunks for both sliding window chunking and fixed size chunking are because content must be diluted in larger chunks. The results show that strategies with 1024 tokens scored high on contextual recall. This is intuitive, as larger chunks contain more text, making the chunk more likely to contain the answer. The lower score on contextual precision is because larger chunks contain more noise and the chunk is less focused compared to smaller chunks. This also explains the opposite effect for 256-token chunks. Larger chunks achieved higher mean scores on textual questions than on exact value questions. An explanation may be that numbers and abbreviations become diluted in larger chunks rather than in smaller chunks while embedding. Textual questions also require more explanation than exact value questions. This may account for why 1024 token chunks achieved higher mean scores with textual questions. The inverse pattern is seen for 256 token chunks, which were recorded higher means on exact questions, suggesting that precise numerical values are more faithfully represented when embedding smaller chunks. The results show that metrics on which small chunks score higher, large chunks score worse and vice versa. Overall, 512 is the highest size as it has advantages of both sides.

The metric content overlap shows opposite results across all tables. As the ANOVA test shows, chunking strategies have a strong effect on the metric and embedding models do not. Together with the findings from Table 1, it is observed that smaller chunking strategies such as propositional chunking, fixed 256 and sliding 256 substantially outperform larger chunking strategies, such as structured chunking or sliding 1024. This is expected, given that small chunks wrap closely around a relevant sentence rather than large chunks, which will contain more noise. Since the score is determined by the ratio of relevant phrase length to the total chunk length, the embedding model has no influence. Propositional chunking substantially outperforms all other strategies on content overlap because it also creates short standalone chunks. The comparatively higher scores of BioMedBERT and BioLORD on this metric are likely attributable to random variation. Because content overlap measures chunk size rather than retrieval quality it is not an adequate metric for answering the research question. However, chunk noise is a factor that influences response quality later in the pipeline [8]. For this reason, the results retain relevance.

6.4 Question Categories

Across all chunking strategies, there is no strategy that performs highest on all question categories. This is because each question type has different requirements from chunking strategies and embedding models. The following paragraphs examine each question category and explain the likely causes of their performance. The paragraphs answer the fourth research question (RQ4) which asks how combinations of chunking strategies and embedding model perform across different categories of patient questions.

Factual questions require coherent paragraphs without diluted context, which explains the high performance of sliding 512 and sliding 256. Factual questions achieved the overall highest mean across all combinations relative to all other question categories, likely because the query language and the chunk language are more aligned than on other question types. Both use medical concepts in similar terms leading to higher semantic similarity which then facilitates retrieval.

The primary challenge of patient questions lies in the linguistic gap between patient language and technical source document language. Patient questions often contain conversational text such as ""My doctor told me…" or "I have a medical condition …", which is perceived as noise from the embedding models perspective. When embedding the query, a high semantic understanding of the question is required which is the probable cause why general-purpose models outperform medical models on this category.

Explaining abbreviations had higher mean scores for all combinations than questions asking only for the meaning of an abbreviation or than to answer numerical questions. However, the mean scores were lower than those from factual or patient questions. The lower means likely arise from the system having to retrieve two relevant text sections. The first with the meaning of the abbreviation and the second to how it relates to ovarian stimulation. As a hybrid category combining elements of abbreviation and factual questions, its performance reflects the characteristics of both.

Abbreviation questions were affected more by the chunking strategy than by embedding models according to the ANOVA test. The results showed that small chunks achieved the highest mean scores for this category. An explanation for this is that when the meaning of the abbreviation is contained in a larger chunk, during embedding the definition is inadequately represented during embedding because of irrelevant noise from surrounding text. Overall, abbreviation questions had the lowest mean score across all combinations. This may result from chunking strategies splitting definitions, embedding models diluting the meaning because of noise, or because of little semantic similarity between the query which only contains the acronym and the chunk which contains both the acronym and the expanded version. A notable finding is that MedCPT achieves the highest mean scores on abbreviation questions. This may be attributable to MedCPT's training on biomedical data, which makes more adept at representing medical terminology compared to general-purpose models.

The question category with the second lowest mean overall are numerical questions. The primary difficulty of this category is that retrieval must yield both the precise numerical value and sufficient contextual information. Numerical questions likely achieved higher

mean scores than abbreviation questions because there is more semantic similarity between the query and the chunks. For numerical questions, structured chunking and sliding 1024 had the highest means. These chunks are more likely to contain the exact number and also the context to it. Smaller chunks may cut numbers in half or miss context.

7 RECOMMENDATIONS

All recommendations are based solely on the results from this experiment which evaluates retrieval performance on only one ESHRE guideline. The recommendation may differ when evaluating retrieval on a wider set of medical documents and with more advanced RAG pipelines. These limitation should be considered accordingly. The primary recommendation arising from the findings is to prioritize optimization of embedding models before optimizing chunking strategies. While both have an individual effect, embedding models have a greater impact on retrieval quality. The absence of an interaction effect between the two variables implies that they can be optimized simultaneously without having to take combinations into account. When selecting an embedding model for a healthcare RAG system, general-purpose embedding models are recommended over medical models because they have a stronger understanding of language and can bridge the gap between patient query language and medical document language more effectively. Following this, more research in optimizing embedding models is recommended. From the tested models OpenAI showed the highest overall performance. BioMedBERT and BioLORD are not recommended as they demonstrated the lowest mean scores on average. In case a medical model is preferred, MedEmbed represents the most suitable option. For any other medical model, the chunk size should be adapted as they performed highest with chunking strategies that created larger chunks.

For chunking strategies, sliding window chunking with small to medium sized chunk (256 token to 512 tokens) as they balance contextual coverage because of the overlap and semantic focus because of their size. Structured chunking is a promising strategy but may have performed lower than sliding window chunking because of the large chunks it created. Future research should evaluate the performance of structured chunking with smaller chunk sizes to see if it outperforms sliding window chunking. Semantic chunking is not recommended on medical documents given the high degree of semantic homogeneity, which creates large chunks that cannot be embedded properly. Also, propositional chunking is not recommended as it reformulates text. Medical terms and processes might not be captured optimally. These results should, however, be viewed critically because as mentioned above, the evaluation methodology systematically penalised propositional chunking.

Additionally, cleaning the input document before chunking is recommended. The greater the ratio of noise is in the source document, the more irrelevant chunks will be created. In the ESHRE document the noise specifically came from references but also from document structure elements such as headers, titles page numbers. If structured chunking is chosen as a chunking strategy, these elements are necessary for chunk boundary detection.

A further recommendation is to test rephrasing patient questions before embedding. Given the reduced semantic similarity between patient questions and medical text, a potential solution is the elimination of conversational noise prior to embedding. This may facilitate more precise embedding, thereby improving retrieval quality and ultimately response quality.

Although certain combinations demonstrate comparatively higher mean scores, no combination exceeded a mean score of 0.527. When looking at recall@k in isolation in Table 13, the highest-performing combination is fixed-size chunking 1024 with OpenAI, yielding a mean score of 0.554. This indicates that for the highest scoring combination for this metric, relevant chunks are retrieved in only 55.4% of cases. This indicates that retrieval performance in this experiment remains insufficient for reliable use in a clinical context. Because retrieval quality directly affects answer quality, this limitation needs to be addressed before deploying RAG-systems in healthcare environments. This also has implications on how RAG systems should be implemented in clinical workflows. Humans cannot yet be removed from the loop and should validate AI-generated responses before they reach patients, as proposed in the HiCARE framework by Sedrakyan et al. (2025) [28]. With the current quality of retrieval, hybrid models combining humans and AI are recommended.

To further develop reliable and safe RAG-systems for reproductive health, it is strongly recommended to test systems on a wide range of medical documents and assess whether performance differs. The study does not provide enough evidence on retrieval performance on all medical guidelines. Further, the system should be tested with real patient questions and not only constructed by the author as done in this study. A recommendation for clinics is to provide real patient questions and how professionals respond to them. A factor that has the potential to significantly improve retrieval quality are more advanced RAG architectures. These are recommended to be researched in more depth. Finally, also the generation side of the RAG system should be studied in a clinical setting, as it is equally as important for optimal performance. This study exclusively gives insights on the retrieval side.

8 LIMITATIONS

The primary limitation of this study is the use of a single document. For a system used in practice, a broader set of medical guidelines should be tested. A further limitation is that for the analysis the metric “content overlap” was considered. For reasons discussed above, the metric may not provide accurate insights on retrieval quality. Also for propositional chunking the results may not be representative of actual performance as the metrics penalized this strategy. As OpenAI served both as an embedding model and as the LLM-as-a-judge, there exists a potential evaluation bias. For application in a clinical setting, it should be noted that this study solely provides insight on the retrieval side of medical RAG systems. Answer generation is also a major factor in the overall quality, and needs to be researched as well. This study evaluates how well relevant content is retrieved, but the connection this has to actual patient guidance remains assumed rather than demonstrated. Finally, this experiment tested a simple RAG pipeline. When applying more advanced retrieval techniques such as reranking, results might change.

9 CONCLUSION

The research tested effects of different embedding models and chunking strategies on retrieval quality for a healthcare RAG system focused on ovarian stimulation. The results show that while both chunking strategies and embedding models influence retrieval quality individually, they have no combined effect. Overall, the choice of embedding model matters more than of chunking strategy. Specifically general-purpose models such as OpenAI or Qwen3 performed highly on chunk retrieval of the ESHRE document. Medical models are not recommended. For chunking strategies, sliding window chunking had the highest mean scores. Especially small and medium size chunks of 256 or 512 tokens perform highly. Textually heavy questions achieved higher mean results across all combination compared to numerical questions or questions asking about the meaning of abbreviations. The results show a strong aversion to propositional chunking and semantic chunking as well as BioMedBERT as an embedding model. Before implementation in a clinical setting more research is recommended specifically with a larger variety of medical documents. Also more advanced RAG pipelines should be considered, such as reranking or hybrid RAG as results may change. With the limitations taken into account, the research provides valuable insight on how chunking strategies and embedding models perform on retrieval quality, which clinics can utilize to build a reliable and high performing RAG system focused on reproductive health.

10 AI STATEMENT

The artificial intelligence model Claude was used in this research to review code and to write logging code to find errors faster. During experiment implementation it was used to create a first layout code to allow faster implementation. All generated code was reviewed by the researcher and edited as needed. All text in the paper was written by the author. However, AI was utilized to scan for spelling and grammar mistakes. The researcher takes full responsibility for the content of the work.

11 ACKNOWLEDGEMENTS

The author would like to acknowledge the support and inspiration provided by the HiCARE framework developed within the Living Models Lab under the leadership of G. Sedrakyan. The case, concepts, discussions, and methodological principles emerging from these initiative contributed to shaping the direction of this work, particularly regarding human-centred AI, responsible integration approach to learning, design and development within healthcare.

REFERENCES

[1] Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health* 4, 6 (06 2025), 1–33. https://doi.org/10.1371/journal.pdig.0000877

[2] ASRM Practice Committee. 2021. Guidance for using text, email, and video communication in practices devoted to reproductive medicine. *Fertility and Sterility* 115, 5 (2021), 1156–1158. https://doi.org/10.1016/j.fertnstert.2021.02.006

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45

[4] Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. 2024. Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems. arXiv:2407.08275 [cs.IR] https://arxiv.org/abs/2407.08275

[5] Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X Retrieval: What Retrieval Granularity Should We Use?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15159–15177. https://doi.org/10.18653/v1/2024.emnlp-main.845

[6] Confident AI. 2026. *Contextual Precision DeepEval Documentation.* https://deepeval.com/docs/metrics-contextual-precision

[7] Confident AI. 2026. *Contextual Recall DeepEval Documentation.* https://deepeval.com/docs/metrics-contextual-recall

[8] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 719–729. https://doi.org/10.1145/3626772.3657834

[9] dtaivpp. 2024. MedEmbed-large-v0.1-onnx: ONNX Export of MedEmbed-Large-v0.1. https://huggingface.co/dtaivpp/MedEmbed-large-v0.1-onnx.

[10] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

[11] ESHRE Ovarian Stimulation Guideline Group. 2025. *Ovarian Stimulation for IVF/ICSI: Update 2025.* Technical Report. European Society of Human Reproduction and Embryology.

[12] Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. 2025. Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey. arXiv:2504.14891 [cs.CL] https://arxiv.org/abs/2504.14891

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2, 1 (2023), 32.

[14] Cesar Abraham Gomez-Cabello, Srinivasagam Prabha, Syed Ali Haider, Ariana Genovese, Bernardo G. Collaco, Nadia G. Wood, Sanjay Bagaria, and Antonio Jorge Forte. 2025. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. *Bioengineering* 12, 11 (2025). https://doi.org/10.3390/bioengineering12111194

[15] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare* 3, 1 (Oct. 2021), 1–23. https://doi.org/10.1145/3458754

[16] Human Fertilisation and Embryology Authority. 2016. In vitro fertilisation (IVF). https://www.hfea.gov.uk/treatments/explore-all-treatments/in-vitro-fertilisation-ivf/ Accessed: 24 April 2026.

[17] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* 39, 11 (2023), btad651.

[18] Jennifer N. John, Sara Gorman, David Scales, and Jack Gorman. 2025. Online Misleading Information About Women's Reproductive Health: A Narrative Review. *Journal of General Internal Medicine* 40, 5 (2025), 1123–1131. https://doi.org/10.1007/s11606-024-09118-6

[19] S. Joshua Johnson, M. Ramakrishna Murty, and I. Navakanth. 2024. A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications* 83, 13 (2024), 37979–38007. https://doi.org/10.1007/s11042-023-17007-z

[20] Dr Julija. 2024. How I Built a Simple Retrieval-Augmented Generation (RAG) Pipeline? https://medium.com/@drjulija/what-is-retrieval-augmented-generation-rag-938e4f6e03d1.

[21] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. *Open Source Strikes Bread - New Fluffy Embedding Model.* https://www.mixedbread.com/blog/mxbai-embed-large-v1

[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel,

Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[23] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4228–4238. https://www.aclweb.org/anthology/2021.naacl-main.334

[24] Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. 2025. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI* 6, 9 (2025). https://doi.org/10.3390/ai6090226

[25] OpenAI. 2024. Vector Embeddings. https://developers.openai.com/api/docs/guides/embeddings.

[26] François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association* (02 2024), ocae029. https://doi.org/10.1093/jamia/ocae029 arXiv:https://academic.oup.com/jamia/advance-article-pdf/doi/10.1093/jamia/ocae029/56772025/ocae029.pdf

[27] Martin Alexander Schaller, Georg Griesinger, and Constanze Banz-Jansen. 2016. Women show a higher level of anxiety during IVF treatment than men and hold different concerns: a cohort study. *Archives of Gynecology and Obstetrics* 293, 5 (2016), 1137–1145. https://doi.org/10.1007/s00404-016-4033-x

[28] Gayane Sedrakyan, Gayane Arustamyan, Simone Borsci, Valeria {Resendez Gomez}, and Caroline Figueroa. 2025. HiCARE Framework: Rethinking Reproductive Healthcare with AI: Balancing Opportunities and Risks with Human-Centric AI Recommenders for Responsible Innovations. In *International Conference on Innovation in Medicine and Healthcare*. Springer, Germany. 13 International KES Conference on Innovation in Medicine and Healthcare (KES-InMed-25), KES-InMed-25 ; Conference date: 25-06-2025 Through 27-06-2025.

[29] Muhammad Arslan Shaukat, Muntasir Adnan, and Carlos C. N. Kuhn. 2026. A Systematic Investigation of Document Chunking Strategies and Embedding Sensitivity. arXiv:2603.06976 [cs.CL] https://arxiv.org/abs/2603.06976

[30] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial Report Chunking for Effective Retrieval Augmented Generation. arXiv:2402.05131 [cs.CL] https://arxiv.org/abs/2402.05131

[31] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

A APPENDIX

A.1 Chunking strategies

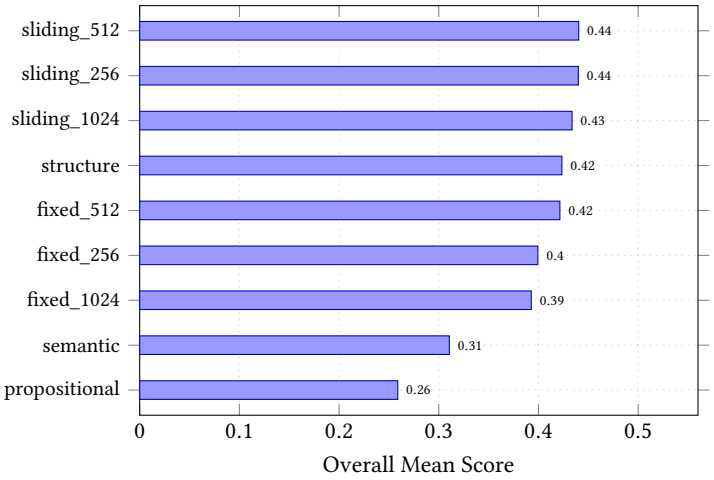


Fig. 1. Mean performance of chunking strategies, ranked best to worst.

Table 1. Mean scores of chunking strategies per metric, coloured by column .

Strategy	R@K	Top-1 P	MRR	nDCG@K	C.Ovlp	Ctx.Rec	Ctx.Pre	Mean
sliding_512	0.352	0.258	0.319	0.339	0.052	0.727	0.352	0.440
sliding_256	0.362	0.231	0.306	0.331	0.091	0.750	0.362	0.440
sliding_1024	0.357	0.272	0.329	0.345	0.022	0.656	0.357	0.434
structure	0.353	0.269	0.325	0.343	0.017	0.628	0.353	0.423
fixed_512	0.328	0.231	0.293	0.311	0.047	0.686	0.328	0.421
fixed_256	0.284	0.182	0.241	0.260	0.092	0.709	0.284	0.399
fixed_1024	0.311	0.207	0.271	0.291	0.024	0.582	0.311	0.393
semantic	0.223	0.128	0.183	0.201	0.037	0.493	0.223	0.311
propositional	0.154	0.103	0.135	0.146	0.672	0.508	0.154	0.259

Table 2. Mean scores of chunking strategies per question category, coloured by column.

Strategy	Abbrev	Abbr. Expl	Fact	Patient	Num
sliding_512	0.361	0.447	0.522	0.462	0.368
sliding_256	0.386	0.464	0.522	0.418	0.393
sliding_1024	0.313	0.398	0.507	0.450	0.409
structure	0.307	0.354	0.484	0.454	0.410
fixed_512	0.339	0.368	0.490	0.435	0.396
fixed_256	0.378	0.373	0.454	0.404	0.360
fixed_1024	0.345	0.374	0.462	0.366	0.378
semantic	0.237	0.355	0.354	0.321	0.269
propositional	0.063	0.245	0.278	0.341	0.244

Table 3. Ranking of chunking strategies on exact-value vs. textual questions.

Strategy	Exact	Rk	Textual	Rk	Avg	ΔRk	ΔMean	Better
sliding_256	0.389	1	0.470	3	0.429	2	0.081	textual
sliding_512	0.364	4	0.492	1	0.428	3	0.128	textual
sliding_1024	0.361	6	0.479	2	0.420	4	0.118	textual
fixed_512	0.367	3	0.463	5	0.415	2	0.096	textual
fixed_256	0.367	2	0.429	6	0.398	4	0.062	textual
structure	0.358	7	0.469	4	0.414	3	0.111	textual
fixed_1024	0.361	5	0.414	7	0.388	2	0.053	textual
semantic	0.253	8	0.337	8	0.295	0	0.084	textual
propositional	0.152	9	0.310	9	0.231	0	0.158	textual

A.2 Embedding models

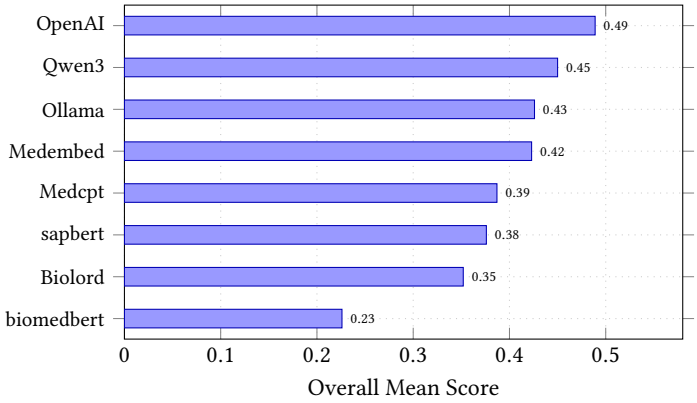


Fig. 2. Mean performance of embedding models, ranked best to worst.

Table 4. Mean scores of embedding models per metric, coloured by column.

Model	R@K	Top-1 P	MRR	nDCG@K	C.Ovlp	Ctx.Rec	Ctx.Pre	Mean
OpenAI	0.443	0.302	0.384	0.409	0.082	0.773	0.443	0.489
Qwen3	0.373	0.273	0.337	0.357	0.076	0.731	0.373	0.450
Ollama	0.322	0.246	0.295	0.310	0.084	0.721	0.322	0.426
Medembed	0.342	0.234	0.299	0.320	0.078	0.695	0.342	0.423
Medcpt	0.297	0.188	0.253	0.275	0.076	0.631	0.297	0.387
sapbert	0.277	0.179	0.241	0.261	0.086	0.626	0.277	0.376
Biolord	0.257	0.186	0.234	0.249	0.100	0.560	0.257	0.352
biomedbert	0.113	0.063	0.093	0.102	0.103	0.365	0.113	0.226

Table 5. Mean scores of embedding models per question category, coloured by column.

Model	Abbrev	Abbr. Expl	Fact	Patient	Num
OpenAI	0.358	0.544	0.546	0.519	0.435
Qwen3	0.309	0.418	0.530	0.480	0.414
Ollama	0.346	0.373	0.486	0.451	0.396
Medembed	0.305	0.433	0.484	0.450	0.379
Medcpt	0.381	0.383	0.438	0.368	0.360
sapbert	0.343	0.297	0.475	0.401	0.297
Biolord	0.279	0.394	0.430	0.332	0.309
biomedbert	0.105	0.160	0.232	0.245	0.278

Table 6. Ranking of embedding models on exact-value vs. textual questions.

Model	Exact	Rk	Textual	Rk	Avg	ΔRk	ΔMean	Better
OpenAI	0.395	1	0.532	1	0.464	0	0.137	textual
Ollama	0.370	2	0.469	3	0.420	1	0.098	textual
Qwen3	0.361	4	0.505	2	0.433	2	0.144	textual
Medembed	0.341	5	0.467	4	0.404	1	0.126	textual
Medcpt	0.370	3	0.403	6	0.386	3	0.033	textual
sapbert	0.320	6	0.438	5	0.379	1	0.118	textual
Biolord	0.294	7	0.381	7	0.337	0	0.087	textual
biomedbert	0.191	8	0.238	8	0.215	0	0.048	textual

A.3 Heatmaps

Table 7. Mean scores for all combinations, coloured by global value.

Model / Chunk	sl.256	sl.512	sl.1024	struct	fx.512	fx.256	fx.1024	prop	sem
OpenAI	0.524	0.500	0.490	0.503	0.501	0.455	0.501	0.422	0.417
Qwen3	0.527	0.496	0.477	0.459	0.491	0.448	0.460	0.328	0.315
Ollama	0.498	0.499	0.459	0.405	0.452	0.469	0.302	0.350	0.303
Medembed	0.487	0.453	0.421	0.453	0.450	0.437	0.387	0.333	0.302
Medcpt	0.400	0.403	0.442	0.388	0.396	0.347	0.397	0.310	0.350
sapbert	0.381	0.406	0.403	0.395	0.350	0.338	0.367	0.309	0.336
Biolord	0.382	0.413	0.369	0.378	0.371	0.357	0.353	0.319	0.196
biomedbert	0.234	0.213	0.263	0.264	0.218	0.223	0.222	0.212	0.162

Table 8. Combinations coloured by column: best embedding per chunking strategy.

Model / Chunk	sl.256	sl.512	sl.1024	struct	fx.512	fx.256	fx.1024	prop	sem
OpenAI	0.524	0.500	0.490	0.503	0.501	0.455	0.501	0.422	0.417
Qwen3	0.527	0.496	0.477	0.459	0.491	0.448	0.460	0.328	0.315
Ollama	0.498	0.499	0.459	0.405	0.452	0.469	0.302	0.350	0.303
Medembed	0.487	0.453	0.421	0.453	0.450	0.437	0.387	0.333	0.302
Medcpt	0.400	0.403	0.442	0.388	0.396	0.347	0.397	0.310	0.350
sapbert	0.381	0.406	0.403	0.395	0.350	0.338	0.367	0.309	0.336
Biolord	0.382	0.413	0.369	0.378	0.371	0.357	0.353	0.319	0.196
biomedbert	0.234	0.213	0.263	0.264	0.218	0.223	0.222	0.212	0.162

Table 9. Combinations coloured by row: best chunking strategy per embedding model.

Model / Chunk	sl.256	sl.512	sl.1024	struct	fx.512	fx.256	fx.1024	prop	sem
OpenAI	0.524	0.500	0.490	0.503	0.501	0.455	0.501	0.422	0.417
Qwen3	0.527	0.496	0.477	0.459	0.491	0.448	0.460	0.328	0.315
Ollama	0.498	0.499	0.459	0.405	0.452	0.469	0.302	0.350	0.303
Medembed	0.487	0.453	0.421	0.453	0.450	0.437	0.387	0.333	0.302
Medcpt	0.400	0.403	0.442	0.388	0.396	0.347	0.397	0.310	0.350
sapbert	0.381	0.406	0.403	0.395	0.350	0.338	0.367	0.309	0.336
Biolord	0.382	0.413	0.369	0.378	0.371	0.357	0.353	0.319	0.196
biomedbert	0.234	0.213	0.263	0.264	0.218	0.223	0.222	0.212	0.162

A.4 General

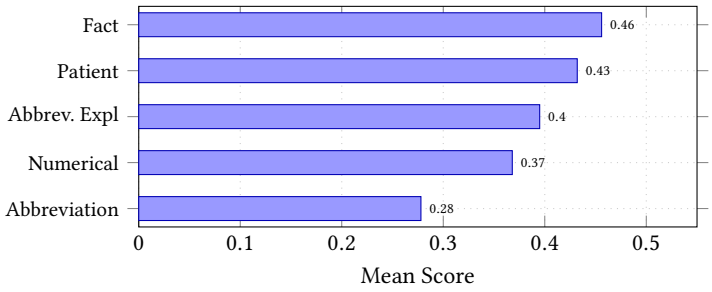


Fig. 3. Mean performance per question category across all combinations.

Table 10. Token count statistics per chunking strategy.

Strategy	Count	Mean	Var	SD	Min	Max
structure	97	1393.2	4,436	66.6	744	1400
fixed_1024	132	1023.8	4	2.1	1000	1024
sliding_1024	166	1016.7	6,423	80.1	9	1024
fixed_512	264	511.9	2	1.5	488	512
semantic	323	418.6	1,082,959	1025.9	2	7476
sliding_512	331	511.0	345	18.6	174	512
fixed_256	528	255.9	1	1.0	232	256
sliding_256	663	255.8	39	6.2	96	256
propositional	1014	55.2	13,860	117.7	5	870

Table 11. Average performance by chunk size across metrics, coloured by column.

Size	R@K	Top-1 P	MRR	nDCG@K	C.Ovlp	Ctx.Rec	Ctx.Pre	Mean
512	0.338	0.245	0.306	0.325	0.049	0.923	0.706	0.413
256	0.321	0.207	0.274	0.296	0.091	0.926	0.729	0.406
1024	0.333	0.239	0.300	0.318	0.023	0.930	0.619	0.395

Table 12. Average performance: sliding window vs. fixed size chunking, coloured by column.

Strategy	R@K	Top-1 P	MRR	nDCG@K	C.Ovlp	Ctx.Rec	Ctx.Pre	Mean
Sliding	0.356	0.254	0.318	0.339	0.055	0.926	0.711	0.423
Fixed	0.305	0.207	0.268	0.287	0.054	0.927	0.659	0.387

A.5 Combinations

Table 13. Mean scores for all combinations per metric.

Combination	R@K	Top-1 P	MRR	nDCG@K	C.Ovlp	Ctx.Rec	Ctx.Pre	Mean
fixed_1024_Biolord	0.250	0.196	0.250	0.262	0.032	0.950	0.529	0.353
fixed_1024_biomedbert	0.120	0.043	0.083	0.095	0.031	0.819	0.362	0.222
fixed_1024_Medcpt	0.348	0.217	0.297	0.320	0.021	1.000	0.576	0.397
fixed_1024_Medembed	0.337	0.196	0.264	0.291	0.017	0.979	0.627	0.387
fixed_1024_Ollama	0.185	0.152	0.181	0.192	0.019	0.862	0.522	0.302
fixed_1024_OpenAI	0.554	0.326	0.453	0.490	0.023	0.954	0.703	0.501
fixed_1024_Qwen3	0.413	0.304	0.373	0.394	0.024	0.970	0.743	0.460
fixed_1024_sapbert	0.272	0.217	0.264	0.280	0.026	0.913	0.598	0.367
fixed_256_Biolord	0.228	0.196	0.228	0.233	0.088	0.901	0.623	0.357
fixed_256_biomedbert	0.065	0.065	0.065	0.065	0.162	0.728	0.409	0.223
fixed_256_Medcpt	0.272	0.087	0.178	0.208	0.089	0.921	0.672	0.347
fixed_256_Medembed	0.337	0.239	0.308	0.328	0.091	0.967	0.790	0.437
fixed_256_Ollama	0.380	0.261	0.330	0.352	0.089	0.989	0.880	0.469
fixed_256_OpenAI	0.370	0.239	0.326	0.354	0.094	0.989	0.815	0.455
fixed_256_Qwen3	0.359	0.239	0.315	0.341	0.093	0.971	0.815	0.448
fixed_256_sapbert	0.228	0.130	0.181	0.201	0.083	0.880	0.663	0.338
fixed_512_Biolord	0.304	0.196	0.257	0.279	0.045	0.938	0.580	0.371
fixed_512_biomedbert	0.065	0.043	0.054	0.057	0.061	0.804	0.442	0.218
fixed_512_Medcpt	0.293	0.239	0.293	0.308	0.050	0.948	0.641	0.396
fixed_512_Medembed	0.391	0.283	0.348	0.370	0.043	0.974	0.743	0.450
fixed_512_Ollama	0.402	0.283	0.355	0.377	0.042	0.936	0.772	0.452
fixed_512_OpenAI	0.457	0.348	0.417	0.436	0.050	0.996	0.801	0.501
fixed_512_Qwen3	0.424	0.304	0.388	0.408	0.046	0.978	0.888	0.491
fixed_512_sapbert	0.272	0.152	0.232	0.254	0.052	0.870	0.620	0.350
propositional_Biolord	0.152	0.109	0.145	0.158	0.722	0.493	0.455	0.319
propositional_biomedbert	0.076	0.043	0.065	0.071	0.782	0.183	0.264	0.212
propositional_Medcpt	0.141	0.065	0.098	0.112	0.728	0.560	0.466	0.310
propositional_Medembed	0.152	0.130	0.149	0.153	0.697	0.487	0.562	0.333
propositional_Ollama	0.163	0.109	0.138	0.145	0.628	0.601	0.663	0.350
propositional_OpenAI	0.228	0.196	0.232	0.245	0.607	0.768	0.678	0.422
propositional_Qwen3	0.120	0.087	0.105	0.112	0.832	0.563	0.476	0.328
propositional_sapbert	0.185	0.087	0.152	0.169	0.547	0.520	0.504	0.309
semantic_Biolord	0.120	0.065	0.105	0.114	0.079	0.563	0.322	0.196
semantic_biomedbert	0.076	0.022	0.054	0.063	0.012	0.724	0.185	0.162
semantic_Medcpt	0.261	0.174	0.217	0.231	0.016	0.949	0.604	0.350
semantic_Medembed	0.250	0.065	0.167	0.202	0.019	0.880	0.529	0.302
semantic_Ollama	0.163	0.152	0.174	0.178	0.073	0.854	0.529	0.303
semantic_OpenAI	0.391	0.239	0.315	0.340	0.023	0.886	0.725	0.417
semantic_Qwen3	0.217	0.174	0.214	0.226	0.034	0.822	0.522	0.315
semantic_sapbert	0.304	0.130	0.217	0.251	0.058	0.848	0.545	0.336
sliding_1024_Biolord	0.261	0.261	0.275	0.283	0.020	0.911	0.572	0.369
sliding_1024_biomedbert	0.163	0.130	0.159	0.169	0.022	0.813	0.388	0.263
sliding_1024_Medcpt	0.424	0.239	0.351	0.388	0.023	0.972	0.696	0.442
sliding_1024_Medembed	0.370	0.261	0.330	0.353	0.019	0.928	0.688	0.421
sliding_1024_Ollama	0.370	0.348	0.377	0.377	0.021	0.957	0.764	0.459
sliding_1024_OpenAI	0.467	0.348	0.424	0.446	0.025	0.935	0.786	0.490
sliding_1024_Qwen3	0.467	0.326	0.406	0.427	0.021	0.974	0.717	0.477
sliding_1024_sapbert	0.326	0.261	0.308	0.320	0.023	0.947	0.634	0.403
sliding_256_Biolord	0.315	0.196	0.254	0.277	0.086	0.881	0.663	0.382
sliding_256_biomedbert	0.120	0.065	0.091	0.101	0.113	0.739	0.409	0.234
sliding_256_Medcpt	0.293	0.196	0.261	0.283	0.079	0.978	0.708	0.400
sliding_256_Medembed	0.446	0.283	0.377	0.407	0.092	0.972	0.837	0.487
sliding_256_Ollama	0.457	0.283	0.373	0.405	0.094	0.991	0.884	0.498
sliding_256_OpenAI	0.489	0.348	0.424	0.452	0.095	0.979	0.884	0.524
sliding_256_Qwen3	0.500	0.326	0.431	0.466	0.094	0.996	0.873	0.527
sliding_256_sapbert	0.272	0.152	0.236	0.260	0.080	0.929	0.736	0.381
sliding_512_Biolord	0.348	0.217	0.301	0.326	0.058	0.897	0.746	0.413
sliding_512_biomedbert	0.120	0.043	0.080	0.093	0.048	0.703	0.406	0.213
sliding_512_Medcpt	0.315	0.239	0.301	0.324	0.050	0.929	0.663	0.403
sliding_512_Medembed	0.391	0.304	0.359	0.379	0.045	0.928	0.764	0.453
sliding_512_Ollama	0.446	0.348	0.420	0.444	0.049	0.978	0.808	0.499
sliding_512_OpenAI	0.457	0.326	0.406	0.429	0.049	0.981	0.855	0.500
sliding_512_Qwen3	0.424	0.370	0.420	0.434	0.049	0.967	0.808	0.496
sliding_512_sapbert	0.304	0.217	0.264	0.282	0.065	0.945	0.763	0.406
structure_Biolord	0.326	0.239	0.290	0.308	0.016	0.902	0.562	0.378
structure_biomedbert	0.207	0.109	0.181	0.202	0.022	0.713	0.418	0.264
structure_Medcpt	0.315	0.239	0.283	0.298	0.015	0.920	0.645	0.388
structure_Medembed	0.380	0.348	0.388	0.400	0.016	0.924	0.717	0.453
structure_Ollama	0.315	0.283	0.308	0.318	0.018	0.929	0.663	0.405
structure_OpenAI	0.533	0.348	0.457	0.487	0.018	0.972	0.707	0.503
structure_Qwen3	0.413	0.326	0.384	0.404	0.017	0.930	0.736	0.459
structure_sapbert	0.326	0.261	0.312	0.327	0.015	0.951	0.576	0.395

Table 14. Mean scores for all combinations per question category.

Combination	Abbrev	Abbr. Expl	Num	Patient	Fact	Mean
fixed_1024_Biolord	0.374	0.312	0.302	0.359	0.482	0.366
fixed_1024_biomedbert	0.123	0.285	0.333	0.171	0.248	0.232
fixed_1024_Medcpt	0.372	0.505	0.454	0.299	0.475	0.421
fixed_1024_Medembed	0.472	0.254	0.373	0.412	0.477	0.398
fixed_1024_Ollama	0.158	0.225	0.357	0.332	0.392	0.293
fixed_1024_OpenAI	0.499	0.565	0.429	0.529	0.569	0.518
fixed_1024_Qwen3	0.316	0.496	0.466	0.456	0.575	0.462
fixed_1024_sapbert	0.448	0.350	0.305	0.369	0.478	0.390
fixed_256_Biolord	0.409	0.411	0.309	0.346	0.431	0.381
fixed_256_biomedbert	0.067	0.117	0.321	0.236	0.231	0.194
fixed_256_Medcpt	0.502	0.410	0.258	0.392	0.359	0.384
fixed_256_Medembed	0.326	0.532	0.402	0.444	0.533	0.447
fixed_256_Ollama	0.587	0.443	0.420	0.466	0.551	0.493
fixed_256_OpenAI	0.430	0.518	0.436	0.453	0.519	0.471
fixed_256_Qwen3	0.409	0.330	0.429	0.478	0.561	0.442
fixed_256_sapbert	0.296	0.221	0.277	0.415	0.451	0.332
fixed_512_Biolord	0.401	0.234	0.373	0.353	0.497	0.371
fixed_512_biomedbert	0.067	0.167	0.316	0.258	0.236	0.209
fixed_512_Medcpt	0.402	0.428	0.392	0.359	0.489	0.414
fixed_512_Medembed	0.276	0.407	0.385	0.585	0.539	0.438
fixed_512_Ollama	0.550	0.369	0.392	0.506	0.520	0.467
fixed_512_OpenAI	0.361	0.495	0.515	0.562	0.548	0.496
fixed_512_Qwen3	0.308	0.500	0.517	0.492	0.608	0.485
fixed_512_sapbert	0.351	0.345	0.266	0.369	0.483	0.363
propositional_Biolord	0.022	0.491	0.234	0.334	0.182	0.253
propositional_biomedbert	0.025	0.065	0.027	0.223	0.163	0.101
propositional_Medcpt	0.092	0.230	0.285	0.323	0.180	0.222
propositional_Medembed	0.025	0.274	0.283	0.309	0.310	0.240
propositional_Ollama	0.069	0.100	0.380	0.399	0.292	0.248
propositional_OpenAI	0.086	0.506	0.294	0.456	0.466	0.362
propositional_Qwen3	0.058	0.137	0.231	0.304	0.316	0.209
propositional_sapbert	0.127	0.155	0.195	0.381	0.318	0.235
semantic_Biolord	0.173	0.402	0.118	0.199	0.215	0.222
semantic_biomedbert	0.131	0.133	0.230	0.205	0.139	0.168
semantic_Medcpt	0.344	0.348	0.329	0.300	0.416	0.347
semantic_Medembed	0.267	0.339	0.310	0.271	0.392	0.316
semantic_Ollama	0.253	0.501	0.201	0.342	0.367	0.333
semantic_OpenAI	0.171	0.614	0.373	0.492	0.478	0.425
semantic_Qwen3	0.253	0.219	0.273	0.392	0.413	0.310
semantic_sapbert	0.307	0.286	0.318	0.365	0.412	0.338
sliding_1024_Biolord	0.243	0.282	0.391	0.376	0.505	0.359
sliding_1024_biomedbert	0.181	0.144	0.373	0.309	0.264	0.254
sliding_1024_Medcpt	0.436	0.311	0.430	0.436	0.575	0.438
sliding_1024_Medembed	0.333	0.350	0.364	0.512	0.526	0.417
sliding_1024_Ollama	0.241	0.566	0.477	0.470	0.556	0.462
sliding_1024_OpenAI	0.348	0.652	0.410	0.544	0.569	0.505
sliding_1024_Qwen3	0.311	0.542	0.467	0.542	0.526	0.478
sliding_1024_sapbert	0.409	0.335	0.355	0.412	0.539	0.410
sliding_256_Biolord	0.292	0.566	0.291	0.201	0.577	0.385
sliding_256_biomedbert	0.033	0.167	0.343	0.260	0.237	0.208
sliding_256_Medcpt	0.521	0.399	0.324	0.435	0.454	0.427
sliding_256_Medembed	0.320	0.693	0.454	0.487	0.559	0.503
sliding_256_Ollama	0.605	0.404	0.462	0.505	0.578	0.511
sliding_256_OpenAI	0.452	0.501	0.516	0.509	0.640	0.524
sliding_256_Qwen3	0.390	0.600	0.498	0.539	0.616	0.529
sliding_256_sapbert	0.473	0.379	0.252	0.403	0.514	0.404
sliding_512_Biolord	0.335	0.507	0.354	0.432	0.515	0.428
sliding_512_biomedbert	0.153	0.167	0.227	0.244	0.273	0.212
sliding_512_Medcpt	0.428	0.457	0.305	0.420	0.516	0.425
sliding_512_Medembed	0.404	0.571	0.382	0.493	0.520	0.474
sliding_512_Ollama	0.379	0.443	0.475	0.533	0.621	0.491
sliding_512_OpenAI	0.435	0.543	0.479	0.536	0.565	0.511
sliding_512_Qwen3	0.415	0.515	0.409	0.565	0.606	0.502
sliding_512_sapbert	0.343	0.377	0.300	0.472	0.559	0.410
structure_Biolord	0.263	0.344	0.404	0.384	0.465	0.372
structure_biomedbert	0.166	0.198	0.319	0.297	0.299	0.256
structure_Medcpt	0.334	0.364	0.454	0.346	0.475	0.394
structure_Medembed	0.323	0.476	0.439	0.534	0.504	0.455
structure_Ollama	0.276	0.302	0.385	0.508	0.500	0.394
structure_OpenAI	0.441	0.504	0.441	0.589	0.559	0.507
structure_Qwen3	0.324	0.418	0.422	0.549	0.547	0.452
structure_sapbert	0.330	0.227	0.408	0.426	0.523	0.383

Table 15. Ranking of all combinations on exact-value vs. textual questions.

Combination	Exact	Rk	Textual	Rk	Avg	ΔRk	ΔPerf	Higher Rk	Higher Perf
fixed_512_Medembed	0.331	44	0.562	6	0.446	38	0.231	exact	textual
fixed_1024_Medcpt	0.413	16	0.387	51	0.400	35	0.026	textual	exact
propositional_OpenAI	0.190	62	0.461	34	0.326	28	0.272	exact	textual
structure_Medcpt	0.394	20	0.411	46	0.402	26	0.017	textual	textual
sliding_512_sapbert	0.321	45	0.516	19	0.418	26	0.194	exact	textual
fixed_512_Medcpt	0.397	19	0.424	43	0.410	24	0.027	textual	textual
fixed_256_Medcpt	0.380	28	0.376	52	0.378	24	0.004	textual	exact
semantic_OpenAI	0.272	53	0.485	29	0.378	24	0.213	exact	textual
sliding_256_Medcpt	0.422	13	0.445	36	0.434	23	0.022	textual	textual
fixed_1024_Medembed	0.422	14	0.445	37	0.433	23	0.022	textual	textual
sliding_1024_OpenAI	0.379	29	0.556	7	0.468	22	0.177	exact	textual
fixed_256_Ollama	0.504	2	0.509	23	0.506	21	0.005	textual	textual
sliding_1024_Medembed	0.349	38	0.519	17	0.434	21	0.170	exact	textual
structure_Qwen3	0.373	31	0.548	12	0.461	19	0.176	exact	textual
fixed_512_Ollama	0.471	4	0.513	22	0.492	18	0.042	textual	textual
fixed_256_OpenAI	0.433	11	0.486	28	0.460	17	0.053	textual	textual
sliding_512_Qwen3	0.412	18	0.585	1	0.499	17	0.173	exact	textual
structure_Ollama	0.331	43	0.504	26	0.417	17	0.173	exact	textual
fixed_512_Biolord	0.387	25	0.425	41	0.406	16	0.039	textual	textual
sliding_1024_Ollama	0.359	37	0.513	21	0.436	16	0.154	exact	textual
sliding_1024_Medcpt	0.433	10	0.506	25	0.469	15	0.072	textual	textual
fixed_1024_sapbert	0.377	30	0.424	44	0.400	14	0.047	textual	textual
fixed_256_Biolord	0.359	36	0.388	50	0.373	14	0.029	textual	textual
semantic_Medcpt	0.336	41	0.358	54	0.347	13	0.022	textual	textual
sliding_256_Ollama	0.533	1	0.541	13	0.537	12	0.008	textual	textual
fixed_256_sapbert	0.287	51	0.433	39	0.360	12	0.146	exact	textual
fixed_1024_biomedbert	0.228	57	0.209	69	0.219	12	0.019	textual	exact
propositional_sapbert	0.161	67	0.350	56	0.255	11	0.189	exact	textual
propositional_Qwen3	0.145	70	0.310	59	0.227	11	0.165	exact	textual
sliding_1024_biomedbert	0.277	52	0.286	62	0.282	10	0.009	textual	textual
sliding_512_Ollama	0.427	12	0.577	3	0.502	9	0.150	exact	textual
sliding_1024_Qwen3	0.389	23	0.534	14	0.462	9	0.145	exact	textual
sliding_256_Medembed	0.387	24	0.523	15	0.455	9	0.136	exact	textual
structure_Medembed	0.381	27	0.519	18	0.450	9	0.138	exact	textual
sliding_1024_Biolord	0.317	46	0.440	38	0.379	8	0.123	exact	textual
fixed_512_sapbert	0.308	48	0.426	40	0.367	8	0.118	exact	textual
semantic_Medembed	0.289	50	0.331	58	0.310	8	0.043	textual	textual
fixed_256_biomedbert	0.194	60	0.233	68	0.213	8	0.040	textual	textual
propositional_Medembed	0.154	68	0.310	60	0.232	8	0.156	exact	textual
fixed_512_Qwen3	0.413	17	0.550	10	0.481	7	0.137	exact	textual
fixed_256_Medembed	0.364	34	0.489	27	0.426	7	0.125	exact	textual
sliding_512_Biolord	0.344	39	0.473	32	0.409	7	0.129	exact	textual
semantic_Qwen3	0.263	54	0.402	47	0.333	7	0.139	exact	textual
propositional_Biolord	0.128	71	0.258	64	0.193	7	0.130	exact	textual
fixed_1024_OpenAI	0.464	5	0.549	11	0.506	6	0.085	textual	textual
fixed_512_biomedbert	0.191	61	0.247	67	0.219	6	0.055	textual	textual
semantic_biomedbert	0.180	66	0.172	72	0.176	6	0.008	textual	exact
sliding_256_Qwen3	0.444	7	0.578	2	0.511	5	0.134	exact	textual
fixed_1024_Biolord	0.338	40	0.421	45	0.379	5	0.083	textual	textual
structure_biomedbert	0.243	56	0.298	61	0.270	5	0.056	textual	textual
sliding_1024_sapbert	0.382	26	0.475	30	0.429	4	0.094	textual	textual
sliding_512_OpenAI	0.457	6	0.550	9	0.504	3	0.094	textual	textual
structure_OpenAI	0.441	8	0.574	5	0.507	3	0.133	exact	textual
sliding_512_Medembed	0.393	21	0.506	24	0.450	3	0.114	textual	textual
semantic_Ollama	0.227	58	0.354	55	0.291	3	0.127	exact	textual
fixed_1024_Qwen3	0.391	22	0.515	20	0.453	2	0.124	exact	textual
semantic_sapbert	0.313	47	0.389	49	0.351	2	0.076	textual	textual
fixed_1024_Ollama	0.257	55	0.362	53	0.310	2	0.105	exact	textual
propositional_Ollama	0.225	59	0.346	57	0.285	2	0.121	exact	textual
sliding_256_OpenAI	0.484	3	0.574	4	0.529	1	0.091	textual	textual
fixed_512_OpenAI	0.438	9	0.555	8	0.497	1	0.117	exact	textual
fixed_256_Qwen3	0.419	15	0.519	16	0.469	1	0.100	textual	textual
structure_sapbert	0.369	32	0.475	31	0.422	1	0.106	exact	textual
sliding_256_Biolord	0.291	49	0.389	48	0.340	1	0.098	exact	textual
propositional_Medcpt	0.188	64	0.251	65	0.220	1	0.063	textual	textual
sliding_256_biomedbert	0.188	65	0.249	66	0.219	1	0.060	textual	textual
semantic_Biolord	0.146	69	0.207	70	0.176	1	0.062	textual	textual
propositional_biomedbert	0.026	72	0.193	71	0.109	1	0.167	exact	textual
sliding_512_Medcpt	0.367	33	0.468	33	0.417	0	0.101	same	textual
sliding_256_sapbert	0.363	35	0.459	35	0.411	0	0.096	same	textual
structure_Biolord	0.333	42	0.424	42	0.379	0	0.091	same	textual
sliding_512_biomedbert	0.190	63	0.258	63	0.224	0	0.069	same	textual

Table 16. All combinations ranked by overall mean.

Combination	Mean
sliding_256_Qwen3	0.527
sliding_256_OpenAI	0.524
structure_OpenAI	0.503
fixed_512_OpenAI	0.501
fixed_1024_OpenAI	0.501
sliding_512_OpenAI	0.500
sliding_512_Ollama	0.499
sliding_256_Ollama	0.498
sliding_512_Qwen3	0.496
fixed_512_Qwen3	0.491
sliding_1024_OpenAI	0.490
sliding_256_Medembed	0.487
sliding_1024_Qwen3	0.477
fixed_256_Ollama	0.469
fixed_1024_Qwen3	0.460
sliding_1024_Ollama	0.459
structure_Qwen3	0.459
fixed_256_OpenAI	0.455
sliding_512_Medembed	0.453
structure_Medembed	0.453
fixed_512_Ollama	0.452
fixed_512_Medembed	0.450
fixed_256_Qwen3	0.448
sliding_1024_Medcpt	0.442
fixed_256_Medembed	0.437
propositional_OpenAI	0.422
sliding_1024_Medembed	0.421
semantic_OpenAI	0.417
sliding_512_Biolord	0.413
sliding_512_sapbert	0.406
structure_Ollama	0.405
sliding_512_Medcpt	0.403
sliding_1024_sapbert	0.403
sliding_256_Medcpt	0.400
fixed_1024_Medcpt	0.397
fixed_512_Medcpt	0.396
structure_sapbert	0.395
structure_Medcpt	0.388
fixed_1024_Medembed	0.387
sliding_256_Biolord	0.382
sliding_256_sapbert	0.381
structure_Biolord	0.378
fixed_512_Biolord	0.371
sliding_1024_Biolord	0.369
fixed_1024_sapbert	0.367
fixed_256_Biolord	0.357
fixed_1024_Biolord	0.353
fixed_512_sapbert	0.350
propositional_Ollama	0.350
semantic_Medcpt	0.350
fixed_256_Medcpt	0.347
fixed_256_sapbert	0.338

Table 16 *(continued)*

Combination	Mean
semantic_sapbert	0.336
propositional_Medembed	0.333
propositional_Qwen3	0.328
propositional_Biolord	0.319
semantic_Qwen3	0.315
propositional_Medcpt	0.310
propositional_sapbert	0.309
semantic_Ollama	0.303
fixed_1024_Ollama	0.302
semantic_Medembed	0.302
structure_biomedbert	0.264
sliding_1024_biomedbert	0.263
sliding_256_biomedbert	0.234
fixed_256_biomedbert	0.223
fixed_1024_biomedbert	0.222
fixed_512_biomedbert	0.218
sliding_512_biomedbert	0.213
propositional_biomedbert	0.212
semantic_Biolord	0.196
semantic_biomedbert	0.162

Table 17. All combinations: left coloured by embedding model, right by chunking strategy.

By Embedding Model		By Chunking Strategy	
Combination	Mean	Combination	Mean
sliding_256_Qwen3	0.527	sliding_256_Qwen3	
sliding_256_OpenAI	0.524	sliding_256_OpenAI	
structure_OpenAI	0.503	structure_OpenAI	
fixed_512_OpenAI	0.501	fixed_512_OpenAI	
fixed_1024_OpenAI	0.501	fixed_1024_OpenAI	
sliding_512_OpenAI	0.500	sliding_512_OpenAI	
sliding_512_Ollama	0.499	sliding_512_Ollama	
sliding_256_Ollama	0.498	sliding_256_Ollama	
sliding_512_Qwen3	0.496	sliding_512_Qwen3	
fixed_512_Qwen3	0.491	fixed_512_Qwen3	
sliding_1024_OpenAI	0.490	sliding_1024_OpenAI	
sliding_256_Medembed	0.487	sliding_256_Medembed	
sliding_1024_Qwen3	0.477	sliding_1024_Qwen3	
fixed_256_Ollama	0.469	fixed_256_Ollama	
fixed_1024_Qwen3	0.460	fixed_1024_Qwen3	
sliding_1024_Ollama	0.459	sliding_1024_Ollama	
structure_Qwen3	0.459	structure_Qwen3	
fixed_256_OpenAI	0.455	fixed_256_OpenAI	
structure_Medembed	0.453	structure_Medembed	
sliding_512_Medembed	0.453	sliding_512_Medembed	
fixed_512_Ollama	0.452	fixed_512_Ollama	
fixed_512_Medembed	0.450	fixed_512_Medembed	
fixed_256_Qwen3	0.448	fixed_256_Qwen3	
sliding_1024_Medcpt	0.442	sliding_1024_Medcpt	
fixed_256_Medembed	0.437	fixed_256_Medembed	
propositional_OpenAI	0.422	propositional_OpenAI	
sliding_1024_Medembed	0.421	sliding_1024_Medembed	

Table 17 *(continued)*

By Embedding Model		By Chunking Strategy	
Combination	**Mean**	**Combination**	**Mean**
semantic_OpenAI	0.417	semantic_OpenAI	
sliding_512_Biolord	0.413	sliding_512_Biolord	
sliding_512_sapbert	0.406	sliding_512_sapbert	
structure_Ollama	0.405	structure_Ollama	
sliding_512_Medcpt	0.403	sliding_512_Medcpt	
sliding_1024_sapbert	0.403	sliding_1024_sapbert	
sliding_256_Medcpt	0.400	sliding_256_Medcpt	
fixed_1024_Medcpt	0.397	fixed_1024_Medcpt	
fixed_512_Medcpt	0.396	fixed_512_Medcpt	
structure_sapbert	0.395	structure_sapbert	
structure_Medcpt	0.388	structure_Medcpt	
fixed_1024_Medembed	0.387	fixed_1024_Medembed	
sliding_256_Biolord	0.382	sliding_256_Biolord	
sliding_256_sapbert	0.381	sliding_256_sapbert	
structure_Biolord	0.378	structure_Biolord	
fixed_512_Biolord	0.371	fixed_512_Biolord	
sliding_1024_Biolord	0.369	sliding_1024_Biolord	
fixed_1024_sapbert	0.367	fixed_1024_sapbert	
fixed_256_Biolord	0.357	fixed_256_Biolord	
fixed_1024_Biolord	0.353	fixed_1024_Biolord	
fixed_512_sapbert	0.350	fixed_512_sapbert	
propositional_Ollama	0.350	propositional_Ollama	
semantic_Medcpt	0.350	semantic_Medcpt	
fixed_256_Medcpt	0.347	fixed_256_Medcpt	
fixed_256_sapbert	0.338	fixed_256_sapbert	
semantic_sapbert	0.336	semantic_sapbert	
propositional_Medembed	0.333	propositional_Medembed	
propositional_Qwen3	0.328	propositional_Qwen3	
propositional_Biolord	0.319	propositional_Biolord	
semantic_Qwen3	0.315	semantic_Qwen3	
propositional_Medcpt	0.310	propositional_Medcpt	
propositional_sapbert	0.309	propositional_sapbert	
semantic_Ollama	0.303	semantic_Ollama	
fixed_1024_Ollama	0.302	fixed_1024_Ollama	
semantic_Medembed	0.302	semantic_Medembed	
structure_biomedbert	0.264	structure_biomedbert	
sliding_1024_biomedbert	0.263	sliding_1024_biomedbert	
sliding_256_biomedbert	0.234	sliding_256_biomedbert	
fixed_256_biomedbert	0.223	fixed_256_biomedbert	
fixed_1024_biomedbert	0.222	fixed_1024_biomedbert	
fixed_512_biomedbert	0.218	fixed_512_biomedbert	
sliding_512_biomedbert	0.213	sliding_512_biomedbert	
propositional_biomedbert	0.212	propositional_biomedbert	
semantic_Biolord	0.196	semantic_Biolord	
semantic_biomedbert	0.162	semantic_biomedbert	

A.6 Statistics

Table 18. Two-way ANOVA test results.

Independent Var.	**Dependent Var.**	***p*-value**	η^2	**Strength**	**Significant**
Embedding Model	Recall@K	1.11×10^{-16}	0.042	small	Yes
Chunking Strategy	Recall@K	9.44×10^{-15}	0.024	small	Yes
Combination effect	Recall@K	0.829	0.013	small	No
Embedding Model	Top-1 Precision	1.11×10^{-16}	0.029	small	Yes
Chunking Strategy	Top-1 Precision	0.169	0.020	small	Yes
Combination effect	Top-1 Precision	1.000	0.008	small	No
Embedding Model	MRR	1.11×10^{-16}	0.040	small	Yes
Chunking Strategy	MRR	1.78×10^{-15}	0.025	small	Yes
Combination effect	MRR	0.990	0.010	small	No
Embedding Model	nDCG@K	1.11×10^{-16}	0.042	small	Yes
Chunking Strategy	nDCG@K	4.44×10^{-16}	0.026	small	Yes
Combination effect	nDCG@K	0.972	0.011	small	No
Embedding Model	Content Overlap	0.522	0.002	small	No
Chunking Strategy	Content Overlap	1.11×10^{-16}	0.636	very strong	Yes
Combination effect	Content Overlap	0.795	0.015	small	No
Embedding Model	Contextual Recall	*Insufficient variance - not applicable*			
Chunking Strategy	Contextual Recall				
Combination effect	Contextual Recall				
Embedding Model	Contextual Precision	1.11×10^{-16}	0.073	medium	Yes
Chunking Strategy	Contextual Precision	1.11×10^{-16}	0.038	small	Yes
Combination effect	Contextual Precision	0.882	0.012	small	No
Embedding Model	Overall Mean	1.11×10^{-16}	0.074	medium	Yes
Chunking Strategy	Overall Mean	1.11×10^{-16}	0.048	small	Yes
Combination effect	Overall Mean	0.880	0.012	small	No
Embedding Model	Overall Mean	1.11×10^{-16}	0.074	medium	Yes
Question Category	Overall Mean	1.11×10^{-16}	0.029	small	Yes
Combination effect	Overall Mean	0.003	0.014	small	Yes
Embedding Model	Abbreviation Q	9.15×10^{-5}	0.081	medium	Yes
Chunking Strategy	Abbreviation Q	2.45×10^{-6}	0.111	medium	Yes
Combination effect	Abbreviation Q	1.000	0.066	medium	No

A.7 Question Dataset


```
[
  {
    "id": 1,
    "category": "ABBREVIATION",
    "question": "What does PCOM stand for?",
    "ground_truth": "PCOM stands for polycystic ovary morphology",
    "relevant_chunk_1": "Polycystic ovary morphology",
    "relevant_chunk_2": "",
    "relevant_page_1": [180],
    "relevant_page_2": null
  },
  {
    "id": 2,
    "category": "ABBREVIATION",
    "question": "What does hp-FSH stand for?",
    "ground_truth": "hp-FSH stands for highly purified follicle stimulating hormone",
    "relevant_chunk_1": "Highly purified follicle stimulating hormone",
    "relevant_chunk_2": "",
    "relevant_page_1": [180],
    "relevant_page_2": null
  },
  {
    "id": 3,
    "category": "ABBREVIATION",
    "question": "What does MD stand for?",
    "ground_truth": "MD stands for mean difference",
    "relevant_chunk_1": "Mean difference",
    "relevant_chunk_2": "",
    "relevant_page_1": [180],
    "relevant_page_2": null
  },
   {
    "id": 4,
    "category": "ABBREVIATION",
    "question": "What does DHEA stand for?",
    "ground_truth": "DHEA stands for dehydroepiandrosterone",
    "relevant_chunk_1": "Dehydroepiandrosterone",
    "relevant_chunk_2": "",
    "relevant_page_1": [180,92],
    "relevant_page_2": null
  },
  {
    "id": 5,
    "category": "ABBREVIATION",
    "question": "What does LBR stand for?",
    "ground_truth": "LBR stands for Live birth rate",
    "relevant_chunk_1": "Live birth rate",
    "relevant_chunk_2": "",
    "relevant_page_1": [180,9],
    "relevant_page_2": null
  },



  {
    "id": 6,
    "category": "ABBREVIATION_EXPLAINED",
    "question": "What does hCG stand for and what is its role in ovarian stimulation?",
```

```
  "ground_truth": "hCG stands for human chorionic gonadotrophin and is used for triggering final oocyte maturation in ovarian
    stimulation.",
  "relevant_chunk_1": "human chorionic gonadotrophin",
  "relevant_chunk_2": "The use of recombinant hCG and urinary hCG is equally recommended for triggering final oocyte
   maturation in ovarian stimulation protocols.",
  "relevant_page_1": [180, 13, 82, 115, 135, 142, 160],
  "relevant_page_2": [148]
},
{
  "id": 7,
  "category": "ABBREVIATION_EXPLAINED",
  "question": "What does Duostim stand for what is its role in ovarian stimulation?",
  "ground_truth": "Duostim stands for double stimulation, ovarian stimulation during the follicular and luteal phase of the
   same cycle. It is a strategy in IVF where the ovaries are stimulated twice in the same menstrual cycle allowing to recover
    more oocytes in a short time.",
  "relevant_chunk_1": "Double stimulation, ovarian stimulation during the follicular and luteal phase of the same cycle",
  "relevant_chunk_2": "2 stimulation protocols within the same menstrual cycle: first in the follicular phase then second,
   immediately after the oocyte pick up, in the luteal phase of the same cycle.",
  "relevant_page_1": [180],
  "relevant_page_2": [101]
},
{
  "id": 8,
  "category": "ABBREVIATION_EXPLAINED",
  "question": "What does MNC stand for what is its role in ovarian stimulation?",
  "ground_truth": "MNC stands for modified natural cycle and is a procedure in which one or more oocytes are collected from
   the ovaries during a spontaneous menstrual cycle",
  "relevant_chunk_1": "Modified natural cycle",
  "relevant_chunk_2": "A procedure in which one or more oocytes are collected from the ovaries during a spontaneous menstrual
    cycle",
  "relevant_page_1": [180],
  "relevant_page_2": [49, 175]
},
{
  "id": 9,
  "category": "ABBREVIATION_EXPLAINED",
  "question": "What does CC stand for and what is its role in ovarian stimulation?",
  "ground_truth": "CC stands for clomiphene citrate and is an oral medication used for ovarian stimulation.",
  "relevant_chunk_1": "Clomiphene citrate",
  "relevant_chunk_2": "ovarian stimulation with a combination of CC and reduced dose gonadotropins compared to conventional
   stimulation without oral medication",
  "relevant_page_1": [180, 78, 79] ,
  "relevant_page_2": [78]
},
{
  "id": 10,
  "category": "ABBREVIATION_EXPLAINED",
  "question": "What does GnRH stand for and what is its role in ovarian stimulation?",
  "ground_truth": "GnRH stands for gonadotropin-releasing hormone and signals to the pituitary to release LH which causes
   ovulation. It can happen that follicles have not grown sufficiently large before ovulation resulting in a failed oocyte
   pick up. In IVF and ICSI GnRH can be controlled to counter this problem.",
  "relevant_chunk_1": "Gonadotropin-releasing hormone",
  "relevant_chunk_2": "When stimulating the ovaries to create multifollicular development, the fast-rising oestradiol levels
   may elicit an untimely LH surge. Untimely, as follicles may not have grown sufficiently large to ensure the best quality
   oocytes, and when passed unnoticed, oocyte pick up may become a failed procedure. The use of agents that block the
   signalling by the GnRH pulse generator towards the pituitary, such as GnRH agonists, GnRH antagonists and progestins, have
    almost completely ruled such mishaps and have greatly contributed to the efficiency of ovarian stimulation for IVF/ICSI",
  "relevant_page_1": [180, 56, 137, 142, 145],
  "relevant_page_2": [13]
},
```

```
{
  "id": 11,
  "category": "SIMPLE_FACT",
  "question": "Which works better for ovarian stimulation: FSH or hMG?",
  "ground_truth": "Both FSH and hMG are equally recommended",
  "relevant_chunk_1": "The use of recombinant human FSH (r-hFSH) and human menopausal gonadotropin (hMG) for ovarian
   stimulation is equally recommended.",
  "relevant_chunk_2": "",
  "relevant_page_1": [16, 65],
  "relevant_page_2": null
},
 {
  "id": 12,
  "category": "SIMPLE_FACT",
  "question": "What is more recommended for low responders, clomiphene citrate or gonadotropin stimulation?",
  "ground_truth": "Clomiphene citrate alone or in combination with gonadotrophins, and gonadotropin stimulation alone are
   probably equally recommended for predicted low responders",
  "relevant_chunk_1": "Clomiphene citrate alone or in combination with gonadotrophins, and gonadotropin stimulation alone are
    probably equally recommended for predicted low responders",
  "relevant_chunk_2": "",
  "relevant_page_1": [18, 80],
  "relevant_page_2": null
},
{
  "id": 13,
  "category": "SIMPLE_FACT",
  "question": "Is it possible to have a too high response to a standard dosage of FSH?",
  "ground_truth": "Yes, a high response to a standard dosage of 150 IU is a potential source for the development of the
   Ovarian Hyperstimulation Syndrome (OHSS) which can potentially be a life-threatening condition.",
  "relevant_chunk_1": "a high response to a standard dosage of 150 IU may be undesirable as it is a potential source for the
   development of the Ovarian Hyperstimulation Syndrome (OHSS), even today a potentially life-threatening condition.",
  "relevant_chunk_2": "",
  "relevant_page_1": [13],
  "relevant_page_2": null
},
{
  "id": 14,
  "category": "SIMPLE_FACT",
  "question": "Is a high AMH a predictor for a successful pregnancy?",
  "ground_truth": "AMH alone is not a predictor of the outcome pregnancy.",
  "relevant_chunk_1": "AMH alone is not a predictor of the outcome pregnancy.",
  "relevant_chunk_2": "",
  "relevant_page_1": [35],
  "relevant_page_2": null
},
{
  "id": 15,
  "category": "SIMPLE_FACT",
  "question": "Can BMI predict the outcomes of pregnancy?",
  "ground_truth": "Yes, BMI alone is a predictor of the outcome pregnancy.",
  "relevant_chunk_1": "BMI alone is a predictor of the outcome pregnancy.",
  "relevant_chunk_2": "",
  "relevant_page_1": [38],
  "relevant_page_2": null
},
{
  "id": 16,
  "category": "SIMPLE_FACT",
  "question": "What is hormone pre-treatment used for?",
```

```
  "ground_truth": "Hormone pre-treatment therapies suppress or reduce LH and or FSH hormones and allow control over multiple
   IVF steps. It is used to ensure follicles grow at the same time and that no follicle becomes too large early on. It is
   also used to hinder cyst formation and also allows better scheduling of IVF cycles",
  "relevant_chunk_1": "Pre-treatment therapies aim to suppress or to reduce LH and/or FSH secretion prior to gonadotrophin
   stimulation in IVF cycles. They are used by clinicians for different purposes such as synchronisation of follicular
   development, prevention of occurrence of early large follicle or spontaneous LH-surge, reduction of cyst formation. Pre-
   treatment is also used for scheduling IVF cycles",
  "relevant_chunk_2": "",
  "relevant_page_1": [41],
  "relevant_page_2": null
},
{
  "id": 17,
  "category": "SIMPLE_FACT",
  "question": "What methods are used to monitor ovarian stimulation cycles?",
  "ground_truth": "Ultrasound scans are a common method to monitoring hormones during ovarian stimulation for IVF",
  "relevant_chunk_1": "Most respondents (98.9%) used ultrasound for monitoring ovarian stimulation cycles.",
  "relevant_chunk_2": "",
  "relevant_page_1": [126],
  "relevant_page_2": null
},
{
  "id": 18,
  "category": "SIMPLE_FACT",
  "question": "What is IVF?",
  "ground_truth": "IVF stands for in vitro fertilisation and is a tool for treating infertility. It involves stimulating the
   ovaries to obtain several oocytes, which are then fertilised in a laboratory. The resulting embryos are then placed in the
    uterine cavity within the window of implantation in order to achieve a pregnancy",
  "relevant_chunk_1": "IVF as a tool for treating infertility",
  "relevant_chunk_2": "Ovarian stimulation for IVF/ICSI aims at obtaining several oocytes that will be turned into embryo's
   through the laboratory process of in vitro fertilisation. These embryos can then be placed in the uterine cavity within
   the window of implantation in order to achieve a pregnancy leading to live birth.",
  "relevant_page_1": [11],
  "relevant_page_2": [48]
},
{
  "id": 19,
  "category": "SIMPLE_FACT",
  "question": "What is ovarian stimulation?",
  "ground_truth": "Ovarian stimulation is a treatment that stimulates the development of ovarian follicles and trigger the
   ovulation process. It is used for IVF treatment and for timed intercourse or insemination. In IVF ovarian stimulation aims
    at obtaining several oocytes that will be turned into embryo's through laboratory processes.",
  "relevant_chunk_1": "Ovarian stimulation is defined as pharmacological treatment with the intention of inducing the
   development of ovarian follicles and trigger the ovulation process of these follicles. It can be used for two purposes: 1)
    for timed intercourse or insemination; 2) for IVF/ICSI, to obtain multiple oocytes at follicular aspiration.",
  "relevant_chunk_2": "Ovarian stimulation for IVF/ICSI aims at obtaining several oocytes that will be turned into embryo's
   through the laboratory process of in vitro fertilisation",
  "relevant_page_1": [9],
  "relevant_page_2": [48]
},
{
  "id": 20,
  "category": "SIMPLE_FACT",
  "question": "What is OHSS?",
  "ground_truth": "OHSS stands for Ovarian Hyperstimulation Syndrome. It is an exaggerated systemic response to ovarian
   stimulation characterized by a wide spectrum of clinical and laboratory manifestations. It may be classified as mild,
   moderate or severe according to the degree of abdominal distention, ovarian enlargement and respiratory, hemodynamic and
   metabolic complications. OHSS can be a potentially life-threatening condition.",
  "relevant_chunk_1": "An exaggerated systemic response to ovarian stimulation characterized by a wide spectrum of clinical
   and laboratory manifestations. It may be classified as mild, moderate or severe according to the degree of abdominal
   distention, ovarian enlargement and respiratory, hemodynamic and metabolic complications.",
```

```
  "relevant_chunk_2": "(OHSS), even today a potentially life-threatening condition",
  "relevant_page_1": [175],
  "relevant_page_2": [13]
},
{
  "id": 21,
  "category": "SIMPLE_FACT",
  "question": "When is it considered a high ovarian response?",
  "ground_truth": "A high ovarian response is characterized by the presence of more follicles and/or oocytes than intended.
   There is no universally accepted definitions of high ovarian response but the GDG suggests defining it as more than 18
   follicles >= 11 mm in size on day of oocyte maturation trigger and/or 18 oocytes collected",
  "relevant_chunk_1": "Generally, more than 18 follicles >= 11 mm in size on day of oocyte maturation trigger and/or 18
   oocytes collected",
  "relevant_chunk_2": "lack of universally accepted definitions of high and low ovarian response",
  "relevant_page_1": [9],
  "relevant_page_2": [9]
},
{
  "id": 22,
  "category": "SIMPLE_FACT",
  "question": "What can AFC predict?",
  "ground_truth": "AFC stands for antral follicle count and can predict ovarian response to ovarian stimulation",
  "relevant_chunk_1": "The prediction of ovarian response categories by AFC alone is reliable.",
  "relevant_chunk_2": "ANTRAL FOLLICLE COUNT (AFC)",
  "relevant_page_1": [25],
  "relevant_page_2": [13, 15, 24, 31, 34, 180]
},



{
  "id": 23,
  "category": "PATIENT_TYPE",
  "question": "I am currently in the luteal phase of my menstrual cycle. Could I start ovarian stimulation for elective
   oocyte preservation next week?",
  "ground_truth": "Yes you can. It does not matter in what menstrual cycle phase you are for starting ovarian stimulation for
    elective oocyte preservation.",
  "relevant_chunk_1": "Ovarian stimulation for elective oocyte preservation can be started irrespective of the menstrual
   cycle phase.",
  "relevant_chunk_2": "",
  "relevant_page_1": [19, 111],
  "relevant_page_2": null
},
{
  "id": 24,
  "category": "PATIENT_TYPE",
  "question": "I have a very low AMH count. Will a higher FSH dosage result in more oocytes?",
  "ground_truth": "When AMH test results are below a certain level, a higher FSH dosage may not produce more oocytes.",
  "relevant_chunk_1": "This is certainly much dependant on the of Antral Follicle Count or AMH result. With test results
   below a certain level, the so called predicted low responder may not produce more oocytes with a higher FSH dosage.",
  "relevant_chunk_2": "",
  "relevant_page_1": [13],
  "relevant_page_2": null
},
{
  "id": 25,
  "category": "PATIENT_TYPE",
  "question": "I would want to have a child one day but not right now. Does my age predict how likely a successful pregnancy
   will be?",
  "ground_truth": "Yes female age is a predictor of the outcome pregnancy.",
  "relevant_chunk_1": "Female age alone is a predictor of the outcome pregnancy.",
```

```
  "relevant_chunk_2": "",
  "relevant_page_1": [38],
  "relevant_page_2": null
},
{
  "id": 26,
  "category": "PATIENT_TYPE",
  "question": "Why does my doctor recommend a reduced gonadotropin dose?",
  "ground_truth": "You most likely are a high responder meaning you produce more follicles and/or oocytes than intended. A
   reduced gonadotropin dose is recommended to high responders decrease the risk of OHSS",
  "relevant_chunk_1": "A reduced gonadotropin dose (100 to <150 IU) is probably recommended to decrease the risk of OHSS in
   predicted high responders",
  "relevant_chunk_2": "",
  "relevant_page_1": [16, 50],
  "relevant_page_2": null
},
{
  "id": 27,
  "category": "PATIENT_TYPE",
  "question": "I have PCOS and a low AMH count. I am afraid that this will reduce my chances of becoming pregnant. What are
   my perspectives?",
  "ground_truth": "For women with PCOS, AMH alone is not a predictor of the outcome pregnancy",
  "relevant_chunk_1": "women with PCOS undergoing their first IVF treatment cycle (Xi et al., 2012). The study population was
    divided into low AMH (4.85 ng/mL; n=41), average AMH (4.85–8.82 ng/mL; n=82) and high AMH (8.82 ng/mL; n=41). No
   significant difference in clinical pregnancy rate was observed with low, average, or high AMH",
  "relevant_chunk_2": "",
  "relevant_page_1": [34],
  "relevant_page_2": null
},
{
  "id": 28,
  "category": "PATIENT_TYPE",
  "question": "My doctor told me I have a high risk of OHSS. What is this?",
  "ground_truth": "OHSS is a response to ovarian stimulation that is potentially life threatening. Symptoms include abdominal
    distention, ovarian enlargement and respiratory, hemodynamic and metabolic complications.",
  "relevant_chunk_1": "An exaggerated systemic response to ovarian stimulation characterized by a wide spectrum of clinical
   and laboratory manifestations. It may be classified as mild, moderate or severe according to the degree of abdominal
   distention, ovarian enlargement and respiratory, hemodynamic and metabolic complications.",
  "relevant_chunk_2": "Ovarian Hyperstimulation Syndrome (OHSS), even today a potentially life-threatening condition.",
  "relevant_page_1": [175],
  "relevant_page_2": null
},
{
  "id": 29,
  "category": "PATIENT_TYPE",
  "question": "I have previously had a poor response to IVF stimulation and only got 2 eggs. Should I try a higher dose of
   FSH next time?",
  "ground_truth": "No, there is no evidence of any benefit of higher FSH dosing for live birth/ongoing pregnancy rates",
  "relevant_chunk_1": "there is no evidence of any benefit of higher FSH dosing for live birth/ongoing pregnancy rates",
  "relevant_chunk_2": "",
  "relevant_page_1": [55],
  "relevant_page_2": null
},
{
  "id": 30,
  "category": "PATIENT_TYPE",
  "question": "I was diagnosed with cancer, and I do not have much time to wait until my new cycle for ovarian stimulation.
   Is it somehow possible to do it before?",
  "ground_truth": "Ovarian stimulation can also be done during the luteal phase.",
  "relevant_chunk_1": "Luteal phase stimulation can be regarded as an extension to urgent oncologic fertility preservation",
  "relevant_chunk_2": "",
```

```
  "relevant_page_1": [99],
  "relevant_page_2": null
},
{
  "id": 31,
  "category": "PATIENT_TYPE",
  "question": "I am a low responder. Is testosterone before ovarian stimulation recomended?",
  "ground_truth": "Use of testosterone before ovarian stimulation is probably not recommended for low responders",
  "relevant_chunk_1": "Use of testosterone before ovarian stimulation is probably not recommended for low responders.",
  "relevant_chunk_2": "",
  "relevant_page_1": [18, 91],
  "relevant_page_2": null
},
{
  "id": 32,
  "category": "PATIENT_TYPE",
  "question": "I know I have a thin endometrium. Will this in any way affect my chances of becoming pregnant?",
  "ground_truth": "A thin endometrium can be an indicator of lower pregnancy chances.",
  "relevant_chunk_1": "There are indications that thin endometrium is related to lower ongoing/clinical pregnancy chances as
   an independent factor",
  "relevant_chunk_2": "",
  "relevant_page_1": [131],
  "relevant_page_2": null
},
{
  "id": 33,
  "category": "PATIENT_TYPE",
  "question": "I have a BMI of 30 and I would like to become pregnant via IVF. What effects will this have on my chance of
   becoming pregnant?",
  "ground_truth": "A higher BMI is negatively associated to pregnancy success. This means that with a BMI of 30 if could
   become more difficult to get a child.",
  "relevant_chunk_1": "A negative association was found between obese women and live birth rate after IVF",
  "relevant_chunk_2": "highest BMI included in the IPD by Broer et al. was 30",
  "relevant_page_1": [39],
  "relevant_page_2": [39]
},
{
  "id": 34,
  "category": "PATIENT_TYPE",
  "question": "My progesterone level was elevated on the day of trigger. Does this mean my embryo transfer will fail?",
  "ground_truth": "Higher progesterone levels on the day of trigger injection are associated with lower pregnancy chances if
   you do a fresh transfer in the same cycle. This is because the womb lining matures too quickly, putting it out of sync
   with the embryo and not because there are any problem with the eggs themselves. However, studies suggest that if your
   embryos are frozen and transferred in a later cycle, when the lining has had time to reset, your chances of pregnancy are
   similar to women who had normal progesterone levels.",
  "relevant_chunk_1": "higher progesterone levels are associated with lower ongoing pregnancy/live birth rates",
  "relevant_chunk_2": "effect of elevated progesterone levels is mediated through endometrial advancement and resultant
   embryo – endometrium asynchrony, not through a perceivable effect on oocyte developmental potential (Racca et al., 2021).
   Thus, deferring embryo transfer to a frozen embryo transfer cycle without endometrial advancement seems to provide similar
    live birth rates with non-progesterone elevated cycles",
  "relevant_page_1": [137],
  "relevant_page_2": [137]
},



{
  "id": 35,
  "category": "NUMERICAL",
  "question": "What is the percentage of couples that visit infertility centres that have an unsuccessful treatment?",
  "ground_truth": "roughly 35-40% will not achieve the so desired goal",
```

```
  "relevant_chunk_1": "Still, of all couples visiting infertility centres, roughly 35-40% will not achieve the so desired
   goal, in spite of lengthy efforts, including IVF,",
  "relevant_chunk_2": "",
  "relevant_page_1": [11],
  "relevant_page_2": null
},
{
  "id": 36,
  "category": "NUMERICAL",
  "question": "What is a standard daily dosage level of FSH for the optimal response level in terms of oocytes",
  "ground_truth": "in terms of oocytes a daily dosage of 150 to 225 IU is mostly considered as standard",
  "relevant_chunk_1": "in terms of oocytes a daily dosage of 150 to 225 IU is mostly considered as standard",
  "relevant_chunk_2": "",
  "relevant_page_1": [13],
  "relevant_page_2": null
},
{
  "id": 37,
  "category": "NUMERICAL",
  "question": "Around what age do most women not have any eggs left?",
  "ground_truth": "At around 50 years old the ovarian primordial follicle pools will have become depleted.",
  "relevant_chunk_1": "age of 50 years, when the ovarian primordial follicle pools will have become depleted",
  "relevant_chunk_2": "",
  "relevant_page_1": [12],
  "relevant_page_2": null
},
{
  "id": 38,
  "category": "NUMERICAL",
  "question": "What is the recommended gonadotropin dose for high responders?",
  "ground_truth": "A gonadotropin dose of 100 to <150 IU is recommended to decrease the risk of OHSS in predicted high
   responders",
  "relevant_chunk_1": "A reduced gonadotropin dose (100 to <150 IU) is probably recommended to decrease the risk of OHSS in
   predicted high responders",
  "relevant_chunk_2": "",
  "relevant_page_1": [16, 50],
  "relevant_page_2": null
},
{
  "id": 39,
  "category": "NUMERICAL",
  "question": "At what number of follicles of  11 mm on ultrasound is there an increased risk of OHSS in a GnRH agonist cycle
   ?",
  "ground_truth": "In GnRH agonist cycles, when there is an ovarian response of 19 follicles of 11 mm, there is an increased
   risk of OHSS and preventative measures are recommended, which should primarily include cancelling final oocyte maturation
   trigger.",
  "relevant_chunk_1": "In GnRH agonist cycles with an ovarian response of 19 follicles of 11 mm, there is an increased risk
   of OHSS and preventative measures are recommended, which should include primarily cancelling final oocyte maturation
   trigger",
  "relevant_chunk_2": "",
  "relevant_page_1": [21, 145],
  "relevant_page_2": null
},
{
  "id": 40,
  "category": "NUMERICAL",
  "question": "At what size is final oocyte maturation is triggered?",
  "ground_truth": "Most often, final oocyte maturation is triggered at sizes of several of the leading follicles between
   16-22 mm",
  "relevant_chunk_1": "Most often, final oocyte maturation is triggered at sizes of several of the leading follicles between
   16-22 mm",
```

```
    "relevant_chunk_2": "",
    "relevant_page_1": [20, 133],
    "relevant_page_2": null
  },
  {
    "id": 41,
    "category": "NUMERICAL",
    "question": "What are recommended typical progesterone dosages?",
    "ground_truth": "Doses of natural progesterone are usualy: 50 mg once daily for intramuscular progesterone 25 mg once daily
      for subcutaneous progesterone 90 mg once daily for vaginal progesterone gel 200 mg three times daily for micronized
     vaginal progesterone in-oil capsules 100 mg two or three times daily for micronized vaginal progesterone in starch
     suppositories 400 mg two times daily for vaginal pessary",
    "relevant_chunk_1": "The dosing of natural progesterone has evolved empirically, usually dosages used include: 50 mg once
     daily for intramuscular progesterone 25 mg once daily for subcutaneous progesterone 90 mg once daily for vaginal
     progesterone gel 200 mg three times daily for micronized vaginal progesterone in-oil capsules 100 mg two or three times
     daily for micronized vaginal progesterone in starch suppositories 400 mg two times daily for vaginal pessary",
    "relevant_chunk_2": "",
    "relevant_page_1": [22, 157],
    "relevant_page_2": null
  },
  {
    "id": 42,
    "category": "NUMERICAL",
    "question": "What reduced dose of urinary hCG is recommended over the standard 10,000 IU dose for triggering final oocyte
     maturation?",
    "ground_truth": "A reduced-dose of 5.000 IU urinary hCG for final oocyte maturation is probably recommended over a 10.000
     IU dose in GnRH agonist protocols, as it may improve safety",
    "relevant_chunk_1": "A reduced-dose of 5.000 IU urinary hCG for final oocyte maturation is probably recommended over a
     10.000 IU dose in GnRH agonist protocols, as it may improve safety.",
    "relevant_chunk_2": "",
    "relevant_page_1": [21, 148],
    "relevant_page_2": null
  },
  {
    "id": 43,
    "category": "NUMERICAL",
    "question": "What is the recommended dosage of GnRH agonist trigger with triptorelin",
    "ground_truth": "If the GnRH agonist trigger with triptorelin is applied, dosages ranging between 0.1-0.4 mg can be chosen
     .",
    "relevant_chunk_1": "If the GnRH agonist trigger with triptorelin is applied, dosages ranging between 0.1-0.4 mg can be
     chosen",
    "relevant_chunk_2": "",
    "relevant_page_1": [22, 150],
    "relevant_page_2": null
  },
  {
    "id": 44,
    "category": "NUMERICAL",
    "question": "During which days of the menstrual cycle can ovarian stimulation be done for luteal phase stimulation?",
    "ground_truth": "Between day 15 and 19 of the menstrual cycle ovarian stimulation can be done in the luteal phase.",
    "relevant_chunk_1": "ovarian stimulation in the luteal phase (day 15-19)",
    "relevant_chunk_2": "",
    "relevant_page_1": [99],
    "relevant_page_2": null
  },
  {
    "id": 45,
    "category": "NUMERICAL",
    "question": "A study on ovarian stimulation in predicted high responders compared 100 IU FSH with 150 IU FSH. What
     percentage of women had a live birth in their first cycle in the 100 IU FSH group?",
```

```
    "ground_truth": "36% of women had a live birth in their first cycle in the study that compared FSH dosage in high
     responders",
    "relevant_chunk_1": "1st cycle live birth (fresh and cryopreserved embryos) (36.0%",
    "relevant_chunk_2": "",
    "relevant_page_1": [50],
    "relevant_page_2": null
  },
  {
    "id": 46,
    "category": "NUMERICAL",
    "question": "A study tested the effect of postponing final oocyte maturation by 24–48 hours. For how many women could
     significantly more oocytes be retrieved in the late hCG group?",
    "ground_truth": "For 743 women significantly more oocytes could be retrieved with hCG.",
    "relevant_chunk_1": "significantly more oocytes were retrieved in late hCG group than in early hCG group (4 RCT, MD 1.2,
     95% CI 1.11–1.30, 743 women)",
    "relevant_chunk_2": "",
    "relevant_page_1": [133],
    "relevant_page_2": null
  }
]
```