A Zero-Shot Comparison of Large Language Models for Efficient Screening in Periodontal Regeneration Research

Carlo Galli *, Maria Teresa Colangelo, Stefano Guizzardi, Marco Meleti, Elena Calciolari *

Posted Date: 27 January 2025

doi: 10.20944/preprints202501.2029.v1

Keywords: Periodontal regeneration; Systematic review; Large Language Models; EMD; Bone graft; Artificial intelligence

Article

A Zero-Shot Comparison of Large Language Models for Efficient Screening in Periodontal Regeneration Research

Carlo Galli [1,*], Maria Teresa Colangelo [1], Stefano Guizzardi [1], Marco Meleti [2] and Elena Calciolari [2,3]

[1] Histology and Embryology Laboratory, Department of Medicine and Surgery, University of Parma, Via Volturno 39, 43126 Parma, Italy

[2] Department of Medicine and Surgery, Dental School, University of Parma, 43126 Parma, Italy

[3] Centre for Oral Clinical Research, Institute of Dentistry, Faculty of Medicine and Dentistry, Queen Mary University of London, London E1 2AD, UK

* Correspondence: carlo.galli@unipr.it

Abstract: The relentless growth of biomedical literature poses a significant challenge for systematic reviews, where manually screening thousands of studies can be very laborious. In this study, we explored the performance of large language models (LLMs) to assist in identifying randomized controlled trials (RCTs) pertaining to periodontal regeneration interventions. Starting from a recent review on Emdogain and bone grafts, we identified a target set of RCTs fulfilling its key inclusion criteria and then retrieved a larger pool of thematically related PubMed articles using the query "periodontal regeneration." Each article was embedded using all-mpnet-base-v2, and the resulting vectors were compared against the mean embedding of the target set to group articles into quartiles of descending relevance. Various LLMs (Open Hermes, Flan T5, GPT-2, GPT-3.5 turbo and GPT-4o) were then tasked with classifying each article in the 4 datasets as "Accepted" or "Rejected." By comparing their predictions, we calculated classification performance metrics. Overall, the results show that different LLM respond differently to prompt strategies and can effectively perform this task (or sub-tasks), confirming the potential role of these tools to expedite systematic reviews.

Keywords: Periodontal regeneration; Systematic review; Large Language Models; EMD; Bone graft; Artificial intelligence

1. Introduction

Systematic reviews and meta-analyses have become indispensable in biomedical research, offering robust insights into the efficacy and safety of interventions through the synthesis of multiple studies [1]. Yet one of the most resource-intensive aspects of conducting a systematic review is the initial screening of abstracts, which can number in the thousands when database searches are designed to maximize sensitivity [2]. As the volume of published research continues to expand exponentially, efficient strategies for article screening are increasingly critical for evidence-based decision-making [3,4].

Conducting such a review typically involves searching multiple databases—such as PubMed, Scopus, or Embase—to capture all potentially relevant articles based on a clearly defined criteria for population, intervention, comparison, outcomes, and time frame (PICOT) [5]. While these broad searches aim to avoid missing important studies, they also generate large numbers of irrelevant or duplicate results. Manual title and abstract screening are not only laborious but also subject to human error, fatigue, and inconsistencies among reviewers [6].

Recent advances in artificial intelligence (AI) and natural language processing (NLP)—particularly through large language models (LLMs)—offer the potential to automate or semi-

 doi:10.20944/preprints202501.2029.v1

automate this screening step [7–9]. Modern LLMs such as GPT can be prompted or fine-tuned to classify abstracts based on predefined inclusion criteria [10]. These automated approaches promise to accelerate systematic reviews, enhance consistency, and allow researchers to concentrate on critical appraisal and data synthesis [11].

This can be extremely advantageous in those fields where technical and scientific advances require scholars to continuously curate and update systematic reviews [12], such as periodontal regeneration. In periodontology, for instance, the regenerative treatment of bone defects has gained prominence, with interventions such as guided tissue regeneration, bone graft materials, and adjuncts bioactive molecules studied extensively [13–16]. However, evidence regarding these purported benefits remains mixed, and rigorous systematic reviews to integrate diverse findings from randomized controlled trials (RCTs) are still sorely needed.

In the present study, we developed a pipeline based on different LLMs to screen articles on periodontal regeneration and test their performance. We first chose a recent systematic review in periodontal regeneration by Fidan et al. [17], collected titles and abstracts from the set of RCTs identified in this review paper and designated these as our "target set". A broad PubMed query using the term "periodontal regeneration" was then performed, yielding a large pool of variably related articles. To better organize and prioritize the screening process, we employed sentence embeddings (all-mpnet-base-v2) on the titles and abstracts of all articles [18,19], calculating each study's cosine similarity to the mean embedding vector of the target set [20]. This approach stratified articles in the pool into quartiles of descending similarity, with the highest quartile expected to contain the most similar studies, and thus representing the most challenging task for LLMs. We then tested different LLMs (OpenHermes, Flan T5, GPT-2, GPT-3.5 Turbo, and GPT-4o) and let them classify the papers based on the focused question and inclusion criteria of Fidan et al.'s review, prompting them to output "ACCEPT" or "REJECT". We generated confusion matrices and standard classification metrics (accuracy, precision, recall, and F1 score) to evaluate how closely each model's performance matched expert judgments [21]. High recall was particularly important to avoid discarding studies that truly met the inclusion criteria [22].

This study thus serves as a proof of concept for how advanced NLP methods can streamline the taxing task of article screening, allowing more focused human input on final selection and data interpretation.

2. Materials and Methods

2.1. Study Design and Rationale

We designed this methodological study to test whether large language models could effectively screen articles for a systematic review. To do that, we arbitrarily chose the recent review "Combination of Enamel Matrix Derivatives with Bone Graft vs Bone Graft Alone in the Treatment of Periodontal Intrabony and Furcation Defects: A Systematic Review and Meta-Analysis" by Fidan et al. [17], which investigated whether combining enamel matrix derivative (EMD) with bone grafts (BG) offers additional clinical benefits compared to BG alone in patients with periodontal intrabony or furcation defects. The focused question of this review paper was:

"Does the combination of EMD + BG provide additional clinical benefits compared with BG alone in terms of CAL gain, PD reduction, pocket closure, composite outcome of treatment success, gingival recession (REC) and bone gain in periodontal intrabony and furcation defects?"

From the text of this systematic review, we extracted a PICOT framework [23]:

- Population (P): Adult periodontitis patients (≥18 years old) with at least one intrabony or furcation defect.
- Intervention (I): Periodontal regenerative surgical procedures using EMD combined with bone grafts (EMD+BG).
- Comparison (C): Periodontal regenerative surgical procedures using bone grafts alone (BG).
- Outcomes (O): CAL gain, PD reduction (primary); secondary outcomes included pocket closure, wound healing, gingival recession (REC), tooth loss, PROMs, and adverse events.

The inclusion criteria explicitly mentioned in the review were:

 doi:10.20944/preprints202501.2029.v1

1. Study Design: Randomized controlled trials (RCTs), parallel or split-mouth, with ≥10 patients per arm.
2. Follow-up: Minimum 6 months after the surgical procedure.
3. Population: Adult periodontitis patients (≥18 years) with intrabony or furcation defects.
4. Intervention: EMD + BG (i.e., Emdogain combined with any bone graft material).
5. Comparison: BG alone.
6. Outcomes: At least CAL gain and PD reduction.

We used these inclusion criteria to draft a set of explicit exclusion criteria:

- Studies focusing exclusively on children (<18 years).
- Studies without a clear mention of EMD or bone grafts.
- Follow-up period <6 months or uncertain.
- Non-randomized studies or fewer than 10 patients per arm.

Based on these criteria, the authors of this review had performed a literature search on Medline, Embase, Web of Science and conducted a manual search on several relevant publications in the field, eventually identifying nine target articles (Table 1). These target articles formed the core reference set for the evaluation of our models.

Table 1. List of the target articles identified in the systematic review.

PMID	Title	Journal	Year	Reference
11990444	A clinical comparison of a bovine-derived xenograft used alone and in combination with enamel matrix derivative for the treatment of periodontal osseous defects in humans.	Journal of periodontology	2002	[24]
12186348	Clinical evaluation of an enamel matrix protein derivative (Emdogain) combined with a bovine-derived xenograft (Bio-Oss) for the treatment of intrabony periodontal defects in humans.	The International journal of periodontics & restorative dentistry	2002	[25]
11990441	Clinical evaluation of an enamel matrix protein derivative combined with a bioactive glass for the treatment of intrabony periodontal defects in humans.	Journal of periodontology	2002	[26]
19053917	Clinical evaluation of demineralized freeze-dried bone allograft with and without enamel matrix derivative for the treatment of periodontal osseous defects in humans.	Journal of periodontology	2008	[27]
20054593	Comparative study of DFDBA in combination with enamel matrix derivative versus DFDBA alone for treatment of periodontal intrabony defects at 12 months post-surgery.	Clinical oral investigations	2011	[28]
23484181	Evaluation of the effectiveness of enamel matrix derivative, bone grafts, and membrane in the	The International journal of periodontics &	2013	[29]

 doi:10.20944/preprints202501.2029.v1

	treatment of mandibular Class II furcation defects.	restorative dentistry		
23379539	Hydroxyapatite/beta-tricalcium phosphate and enamel matrix derivative for treatment of proximal class II furcation defects: a randomized clinical trial.	Journal of clinical periodontology	2013	[30]
26556577	Enamel matrix protein derivative and/or synthetic bone substitute for the treatment of mandibular class II buccal furcation defects. A 12-month randomized clinical trial.	Clinical oral investigations	2016	[31]
31811645	Adjunctive use of enamel matrix derivatives to porcine-derived xenograft for the treatment of one-wall intrabony defects: Two-year longitudinal results of a randomized controlled clinical trial.	Journal of periodontology	2020	[32]

2.2. Data Acquisition

First, we generated a pool of thematically related articles by searching Medline [33] using the query "periodontal regeneration", which was arbitrarily chosen to comprise articles about similar topics to the systematic review. Due to PubMed's limit of retrieving up to 10,000 records in a single pass, we segmented our search by publication date ranges, ultimately retrieving over 16,000 abstracts. We used Biopython's Entrez.efetch [34,35] to parse and store the title and abstract texts within Python DataFrames [36]. Duplicates overlapping with the target articles were removed to ensure a clean distinction between relevant (Target=1) and presumed irrelevant (Target=0) papers.

2.3. Data Pre-Processing

We cleaned all dataframes by removing or replacing invalid or missing fields in Title and Abstract columns with empty strings. Titles and abstracts were concatenated to form a single text input per article for embeddings. We used the sentence-transformer *all-mpnet-base-v2* model to encode each article into a 768-dimensional embedding vector [19]. We computed the mean embedding of the nine target articles and calculated the cosine similarity between each pool article's embedding and this mean vector [20]. Articles in the pool were then ranked and split into four quartiles (Q1–Q4) according to similarity scores:

- Q1: Lowest similarity, with mean = 0.47
- Q2: Lower-mid similarity, with mean = 0.60
- Q3: Higher-mid similarity, with mean = 0.66
- Q4: Highest similarity, with mean = 0.76

Each quartile contained several thousand articles (~4,000–5,000), which we anticipated would require a long time to be processed by LLMs. For feasibility purposes, we then decided to randomly sample 200 articles from each quartile. The final sub-sets ("reduced quartiles", n=200) that were used for testing were therefore thus composed:

- Q1: Mean similarity = 0.43 (range= 0.02-0.60)
- Q2: Mean similarity = 0.55 (range=0.42-0.70)
- Q3: Mean similarity = 0.61 (range=0.48-0.75)
- Q4: Mean similarity = 0.76 (range=0.53-0.94)

 doi:10.20944/preprints202501.2029.v1

The target articles were then added to each dataset. These datasets were designed to pose increasingly difficult challenges for the LLMs to correctly discriminate between relevant and non-relevant articles.

2.4. LLM-Based Classification

We tested several models to classify each article in the reduced quartile sets as either "Accepted" or "Rejected" according to the inclusion criteria above.

1. OpenHermes: OpenHermes is an instruction-tuned language model based on the Mistral 7B architecture (7 billion parameters), designed for effective natural language understanding and generation across a wide range of tasks [37]. For this study, we employed the quantized version of OpenHermes-2.5-Mistral-7B-GGUF (openhermes-2.5-mistral-7b.Q4_K_M.gguf), freely available on Huggingface.com. Quantization reduced the model's 32-bit parameters to 4-bit values, significantly improving computational efficiency while maintaining high performance.
2. Flan T5: Flan-T5 is an instruction-tuned language model developed by Google, designed for general-purpose natural language understanding and generation tasks [38]. Flan-T5 was fine-tuned on a wide array of instruction-following datasets, and optimized for handling tasks such as classification, summarization, and question answering with high accuracy and contextual awareness.
3. GPT-2: GPT-2, developed by OpenAI, lacks the instruction-tuning and domain-specific optimization of more advanced models, but it remains a valuable baseline for understanding the capabilities of earlier-generation language models [39].
4. GPT-3.5 Turbo: GPT-3.5 Turbo, also developed by OpenAI, is an optimized and cost-efficient version of the GPT-3.5 model, providing robust natural language understanding and generation capabilities [40]. With significantly improved contextual reasoning and instruction-following compared to GPT-2, GPT-3.5 Turbo performs better in structured classification tasks. In this study, GPT-3.5 Turbo was utilized via OpenAI's API.
5. GPT-4o: GPT-4o is the optimized version of GPT-4, and it combines enhanced instruction-following capabilities with improved contextual understanding [41]. GPT-4o performs better in complex decision-making and classification scenarios than its predecessors. GPT-4o was accessed via OpenAI's API too.

LLMs require adequate prompting to function, i.e. a set of instructions that LLMs follow to perform their task [42]. We designed the prompt based on the Focused questions and the PICOT outlined in the paper. We drafted a general prompt as follows:

"

You are assisting in a systematic review on periodontal regeneration comparing Emdogain (EMD) + bone graft (BG) versus BG alone.

*Your task is to decide whether the following article should be **ACCEPTED** or **REJECTED** based on the following "soft approach" criteria:*

***Inclusion Criteria**:*
*1. **Population (P)**: Adult periodontitis patients (≥18 years old) with at least one intrabony or furcation defect.*
*2. **Intervention (I)**: Regenerative surgical procedures involving EMD combined with any type of bone graft material (EMD+BG).*
*3. **Comparison (C)**: Regenerative surgical procedures involving BG alone.*
*4. **Outcomes (O)**:*
- Primary: CAL (Clinical Attachment Level) gain, PD (Probing Depth) reduction.
- Secondary: Pocket closure, wound healing, gingival recession, tooth loss, patient-reported outcome measures (PROMs), adverse events.
*5. **Study Design**:*
- Randomized controlled trial (RCT), parallel or split-mouth design.
- ≥10 patients per arm.

 doi:10.20944/preprints202501.2029.v1

- ≥6 months follow-up.

***Decision Approach**:*
*- If **at least one** of the above criteria is explicitly met or strongly implied,*
***AND** none of the criteria are explicitly contradicted, then **ACCEPT**.*
*- If **any** criterion is clearly violated (e.g., population is exclusively children,*
*follow-up is 3 months, or design is not an RCT), then **REJECT**.*
*- If **no** criterion is clearly met, **REJECT**.*

Below is the article's title and abstract. Decide if it should be ACCEPTED or REJECTED according to the "soft approach" described.

Title: {title}
Abstract: {abstract}

If the article is acceptable, respond with exactly:
ACCEPT

Otherwise, respond with exactly:
REJECT
"

This prompt can be conceived as a soft prompt, because it emphasizes recall: if any criterion might be satisfied, the article is likely ACCEPTED unless explicitly contradicted. This prompt was tested with all the models. This prompt is also quite verbose.

While testing the LLMs, it soon became apparent that the less advanced models (OpenHermes, Flan T5, GPT-2) struggled with such a structured prompt, and we decided to also test their performance with a double prompt approach, i.e. by splitting the task into two sub tasks, each with its own dedicated prompt. The initial prompt was very simple and as follows:

"

Initial filter to reject articles that do not mention Bone Graft (BG) or Emdogain (EMD).

"

This prompt was designed to short-list the articles and reject all the reports that did not directly focus on Bone grafts or Emdogain and reduce the number of articles to screen in a more detailed way. The second prompt was then applied only to screen the articles selected by the first screening and was the same as the standard prompt.

We also decided to test a more concise prompt that REJECTS papers unless they definitively match inclusion criteria, only for GPT 3.5 Turbo and GPT 4o. This approach aimed to reduce false positives but risked higher false negatives.

The more concise prompt was as follows:

"

You are an expert periodontology assistant. You are assisting in a systematic review on periodontal regeneration comparing

Emdogain (EMD) + bone graft (BG) versus bone graft alone. Evaluate this article step by step:

1. Population: If the text states adult patients with intrabony/furcation defects, or is silent about age/defect type, it's not violated.

2. Intervention: If Emdogain + bone graft is mentioned or strongly implied, we consider this met.

3. Comparison: If a group uses bone graft alone, or there's at least a control lacking Emdogain, consider it met.

4. Outcomes: If they mention CAL gain or PD reduction, or are silent, do not penalize.
Only reject if they clearly never measure any clinical outcomes.

 doi:10.20944/preprints202501.2029.v1

5. Study design: If they claim RCT or strongly imply it, accept. If they mention a different design (case series, pilot with fewer than 10 patients, or <6-month follow-up), reject.

If at least one criterion is explicitly met and none are clearly violated, answer ACCEPT. Otherwise, REJECT.

If you are unsure, default to ACCEPT unless a contradiction is stated.

Article Title: {title}
Abstract: {abstract}

Respond with ONLY 'ACCEPT' or 'REJECT' accordingly.
"

2.5. Performance Evaluation

We compared the LLM-derived labels (Accepted=1, otherwise 0) against the ground truth label (Target). For the target articles, Target=1 indicated a truly relevant study. For a subset of pool articles, we performed a manual or semi-manual check to confirm Target=0 or Target=1 classification.

Using scikit-learn, we computed confusion matrices for each model, listing:

- True Positives (TP): Correctly accepted relevant articles.
- False Negatives (FN): Relevant articles the model rejected.
- False Positives (FP): Irrelevant articles the model accepted.
- True Negatives (TN): Correctly rejected irrelevant articles.

From these, we calculated accuracy, precision, recall (sensitivity), and F1 score.

Accuracy is the proportion of all correctly classified articles, and is calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is the proportion of accepted articles that are truly relevant, and is calculated as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall (or sensitivity) is the proportion of relevant articles correctly identified, as follows:

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score is the harmonic mean of precision and recall, which means:

$$\text{F1 score} = 2\,x\,\frac{(\text{Precision x Recall})}{(\text{Precision}+\text{Recall})}$$

In the context of systematic reviews, high recall (low FN) is often preferred to avoid missing key articles. However, precision is also valuable to reduce the volume of articles that proceed to manual full-text screening.

2.6. Software and Hardware

The entire pipeline was run in Google Colab notebooks [43]. We mounted Google Drive for data storage and used a Tesla T4 GPU environment to accelerate the embedding steps (MPNet) and the local inference for Open Hermes. For GPT-3.5 Turbo and GPT-4o, classification calls were made via OpenAI's API. The following major libraries and packages were employed:

- pandas version 2.2.2 for data handling [36]
- Biopython version 1.85 for retrieving PubMed records [35]
- sentence-transformers version 3.3.1 for sentence embedding [44]
- scikit-learn version 1.6 for confusion matrices and classification metrics [45]

3. Results

We successfully generated four reduced datasets, each containing 200 articles sampled from quartiles Q1 to Q4 based on their similarity to the target articles. All 9 previously identified target RCTs were added to each subset. The distribution of cosine similarity scores between the articles in each quartile and the target set follows a clear gradient of relevance. As expected, Quartile 1 (Q1)

 doi:10.20944/preprints202501.2029.v1

exhibited the lowest similarity, with a mean score of 0.43, while Quartile 4 (Q4) showed the highest similarity, with a mean of 0.76. Intermediate quartiles, Q2 and Q3, had mean similarities of 0.55 and 0.61, respectively.

A box-and-whisker plot was generated to visualize the distribution of similarity scores within each quartile (Fig. 1).

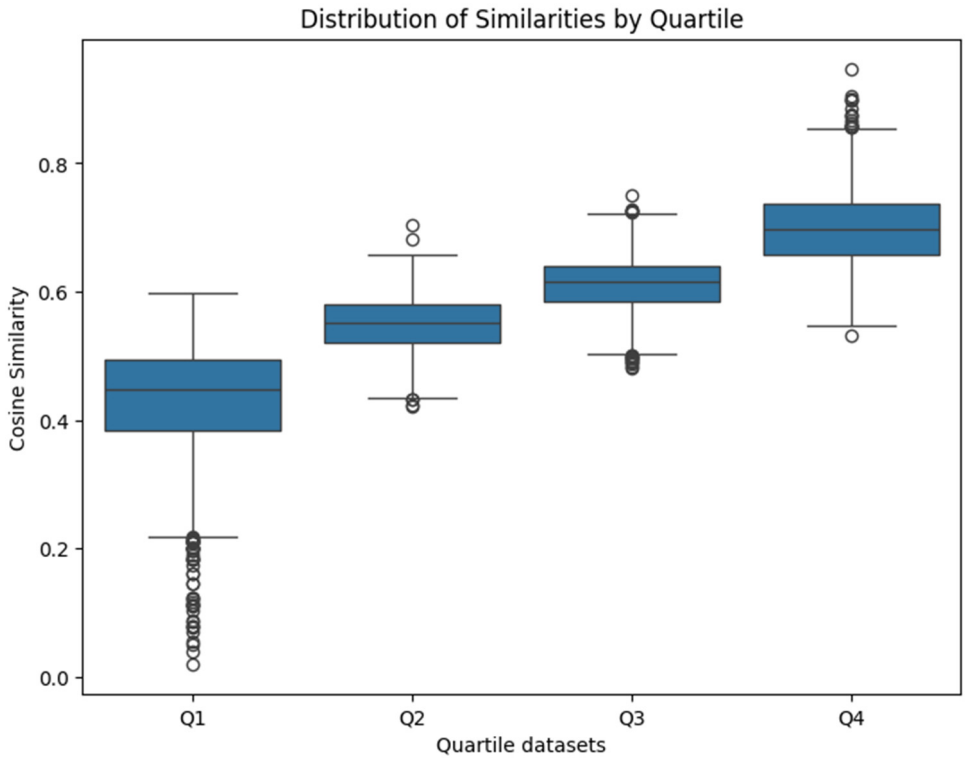


Figure 1. Distribution of cosine similarity scores by quartile. Quartiles (Q1–Q4) were generated by ranking articles based on their cosine similarity to the mean embedding vector of the nine target articles. Q1 contains articles with the lowest similarity, while Q4 contains those with the highest similarity. The box plots display the median, interquartile range, and outliers for each quartile.

Q1 displayed the widest range of scores, with numerous outliers toward the lower end of the similarity scale, likely articles easy to filter out. Conversely, Q4 demonstrated a tighter distribution with higher similarity values. This trend reflects the increasing difficulty of distinguishing between relevant and non-relevant articles as similarity scores rise.

3.1. Open Hermes

The performance of the Open Hermes model was evaluated using two prompting strategies: the single prompt and the double prompt (comprising initial filtering followed by detailed evaluation). When using the single prompt, Open Hermes demonstrated high accuracy across all quartiles but consistently struggled to identify relevant articles (True Positives), as shown in Table 2.

Table 2. Metrics for Open Hermes (Single Prompt).

Quartile	TN	FP	FN	TP
Q1	190	1	8	1
Q2	191	0	9	0
Q3	191	0	9	0
Q4	188	3	9	0

For Quartile 1, the model achieved an accuracy of 95.5%, because it correctly classified 190 non-target papers out of 191. The model had only 1 true positive (TP) and 8 false negatives (FN), resulting in a moderate precision of 50% but low recall and F1-score of 11.1% and 18.2%, respectively. For

 doi:10.20944/preprints202501.2029.v1

Quartiles 2, 3, and 4, the model achieved perfect or near-perfect accuracy (around 95.5%), but no relevant articles were identified, leading to undefined precision and F1-scores (Fig. 2A). While the model effectively rejected irrelevant articles (as evidenced by high true negative counts), it consistently failed to identify relevant studies across quartiles, which defies the purpose of the screening.

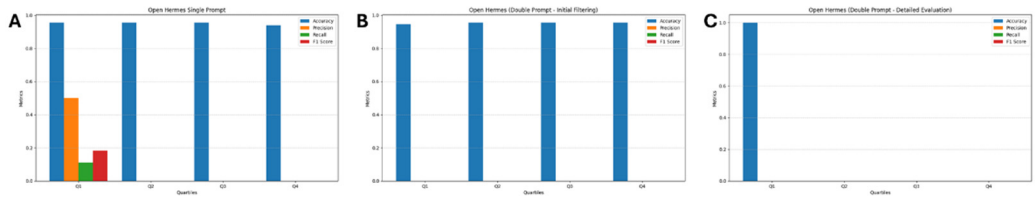


Figure 2. Performance metrics (Accuracy, Precision, Recall, and F1 Score) for Open Hermes across different evaluation setups and quartiles. (A) Results for the single-prompt configuration. (B) Results for the double-prompt configuration after initial filtering. (C) Results for the double-prompt configuration after detailed evaluation. Metrics are shown across quartiles (Q1–Q4), with Accuracy represented in blue, Precision in orange, Recall in green, and F1 Score in red.

The double prompt strategy, designed to improve performance by introducing an initial filtering step, showed mixed results. In the initial filtering phase, Open Hermes primarily focused on rejecting articles that lacked explicit mentions of EMD or BG (Table S1). For Quartile 1, the model achieved 94.5% accuracy, with 189 true negatives (TN) and 2 false positives (FP), but no true positives were identified. For Quartiles 2, 3, and 4, the model achieved 95.5% accuracy, rejecting all irrelevant articles, yet failed to detect any relevant studies, with precision and F1-scores remaining undefined (Fig. 2B).

In the subsequent detailed evaluation phase, Open Hermes was re-applied to assess articles that passed the initial filtering. However, the model's performance remained suboptimal, as obviously no true positives were identified across any quartile, because all target articles had been filtered out during the first round (Table S2, Fig. 2C).

3.2. Flan T5

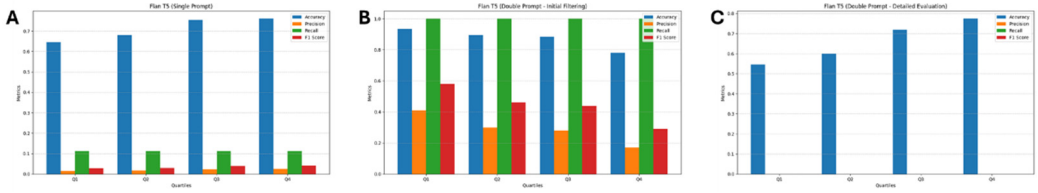


Figure 3. Performance metrics (Accuracy, Precision, Recall, and F1 Score) for Flan T5 across different evaluation setups and quartiles. (A) Results for the single-prompt configuration. (B) Results for the double-prompt configuration after initial filtering. (C) Results for the double-prompt configuration after detailed evaluation. Metrics are shown across quartiles (Q1–Q4), with Accuracy represented in blue, Precision in orange, Recall in green, and F1 Score in red.

The performance of Flan T5 was evaluated using two strategies: a single prompt approach and a double prompt approach, which included both initial filtering and detailed evaluation phases, as done with Open Hermes. Under the single prompt approach, Flan T5 demonstrated moderate accuracy across all quartiles but struggled with precision, recall, and F1-scores due to a high number of false positives and limited true positives (Table S3). In Quartile 1, the model achieved an accuracy of 64.5%, with 128 true negatives (TN), 63 false positives (FP), 8 false negatives (FN), and 1 true positive (TP). Precision was 1.6%, recall was 11.1%, and the F1-score was 2.7%, highlighting the model's limitations in identifying relevant articles. Similar trends were observed across Quartiles 2, 3, and 4, with accuracy ranging from 68% to 76.0%. Precision, recall, and F1-scores remained consistently low, as the model failed to effectively distinguish relevant articles from irrelevant ones

 doi:10.20944/preprints202501.2029.v1

(Fig. 3A). The high number of false positives across all quartiles indicates the model's tendency to err on the side of inclusivity, accepting a large number of irrelevant articles.

When using the double prompt strategy, Flan T5's performance improved significantly during the initial filtering phase, where the model focused on rejecting articles that did not meet broad inclusion criteria (e.g., lack of explicit mentions of EMD or BG). In this phase, the model achieved strong recall, correctly identifying all relevant articles (FN = 0) across all quartiles (Table 3). However, precision was limited due to a substantial number of false positives.

Table 3. Metrics for Flan T5 (Double Prompt - Initial Filtering).

Quartile	TN	FP	FN	TP
Q1	178	13	0	9
Q2	170	21	0	9
Q3	168	23	0	9
Q4	147	44	0	9

For example, in Quartile 1, the model achieved 93.2% accuracy, with 178 TN, 13 FP, 0 FN, and 9 TP. Precision was 40.9%, recall was 100%, and the F1-score was 58% (Fig. 3B). Accuracy across quartiles ranged from 78% to 89.5%, with recall remaining perfect but precision declining as false positives increased, particularly in Quartile 4 (17%), where false positives reached 44.

In the detailed evaluation phase, the model's performance declined significantly (Table 4). Despite achieving reasonable true negative counts, Flan T5 failed to identify any true positives across all quartiles (TP = 0). In Quartile 1, the model achieved 54.5% accuracy, with 12 TN, 1 FP, 9 FN, and 0 TP, while precision, recall, and F1-scores were zero. Similar trends were observed in Quartiles 2, 3, and 4, with accuracy ranging from 60.0% to 77.4% (Fig. 3C). These results highlight the limitations of the model's ability to refine article selection during the detailed evaluation phase, as it struggled to achieve any balance between recall and precision.

Table 4. Metrics for Flan T5 (Double Prompt - Detailed Evaluation).

Quartile	TN	FP	FN	TP
Q1	12	1	9	0
Q2	18	3	9	0
Q3	23	0	9	0
Q4	41	3	9	0

3.3. GPT-2

Using the single-prompt strategy, GPT-2 demonstrated consistent behavior across all quartiles (Q1–Q4). The model classified all articles as "Accepted," resulting in confusion matrices where true negatives (TN) and false negatives (FN) were absent, while all true positives (TP = 9) and false positives (FP = 191) were recorded (Table S4). This led to a precision of 4.5% across all quartiles, while recall remained perfect at 100% due to the complete acceptance of target articles (Fig. 4A).

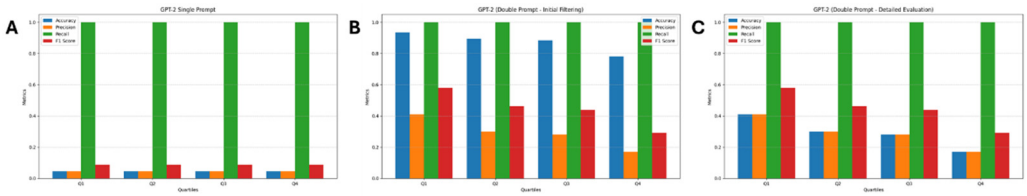


Figure 4. Performance metrics (Accuracy, Precision, Recall, and F1 Score) for GPT-2 across different evaluation setups and quartiles. (A) Results for the single-prompt configuration. (B) Results for the double-prompt configuration after initial filtering. (C) Results for the double-prompt configuration after detailed evaluation. Metrics are shown across quartiles (Q1–Q4), with Accuracy represented in blue, Precision in orange, Recall in green, and F1 Score in red.

 doi:10.20944/preprints202501.2029.v1

However, this behavior inflated the number of false positives, significantly lowering the F1 score to 8.7%. The overall accuracy for the single-prompt approach was 4.5%, underlining the inadequacy of this strategy for discerning relevant from irrelevant articles.

In contrast, the double-prompt approach introduced two stages: initial filtering and detailed evaluation. During the initial filtering, GPT-2 demonstrated improvement by rejecting some irrelevant articles (Table 5).

Table 5. GPT-2 (Double Prompt - Initial Filtering).

Quartile	TN	FP	FN	TP
Q1	178	13	0	9
Q2	170	21	0	9
Q3	168	23	0	9
Q4	147	44	0	9

Across the quartiles, TN increased significantly (e.g., Q1: TN = 178), reducing FP (e.g., Q1: FP = 13). This adjustment boosted accuracy to about 90% for Q1-Q3 quartiles, thought it was 78% in Q4. Precision declined substantially over the quartiles (Q1: 40.9%, Q2: 30%, Q3: 28.1%, Q4: 17%). Recall remained constant at 100% in the initial filtering phase, ensuring that no target articles were excluded prematurely, while F1 score declined from 58% in Q1 to 29% in Q4 for the increasing number of false positive articles (Fig. 4B). In the subsequent detailed evaluation phase (Table 6), the model showed minimal change, with no TN or FN identified. FP matched TP across all quartiles, emphasizing that all initially accepted articles were retained in the final stage.

Table 6. GPT-2 (Double Prompt - Detailed Evaluation).

Quartile	TN	FP	FN	TP
Q1	0	13	0	9
Q2	0	21	0	9
Q3	0	23	0	9
Q4	0	44	0	9

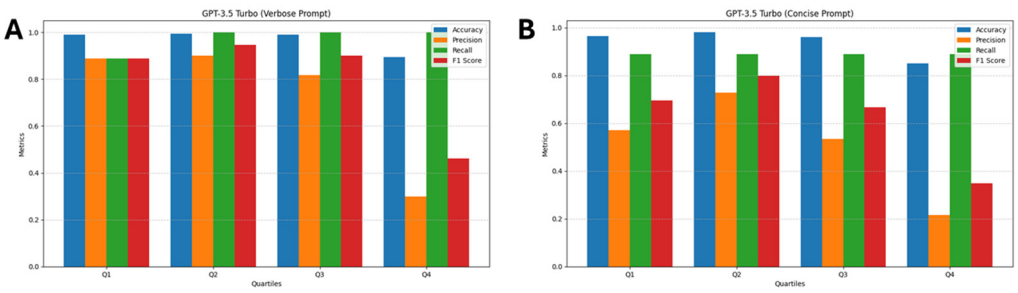
Precision for the detailed evaluation phase was consistent with the initial filtering results, but accuracy and F1 scores were not improved (Fig. 4C).

3.4. GPT-3.5 turbo

In the results for GPT-3.5 Turbo, we observed distinct performance trends when employing verbose and concise prompt strategies. The verbose prompt demonstrated consistent performance across Quartiles 1 to 4, with high true negative (TN) counts and minimal false positives (FP) (Table 7).

Table 7. GPT-3.5 Turbo (Verbose Prompt).

Quartile	TN	FP	FN	TP
Q1	190	1	1	8
Q2	190	1	0	9
Q3	189	2	0	9
Q4	170	21	0	9



 doi:10.20944/preprints202501.2029.v1

Figure 5. Performance metrics (Accuracy, Precision, Recall, and F1 Score) for GPT-3.5 Turbo evaluated using two different prompt styles: (A) Verbose Prompt and (B) Concise Prompt. Metrics are presented across quartiles (Q1–Q4), with Accuracy represented in blue, Precision in orange, Recall in green, and F1 Score in red.

For Quartile 1 (Q1), the model achieved an accuracy of 99.0%, a precision of 88.9%, a recall of 88.9%, and an F1-Score of 88.9%, reflecting balanced performance across metrics (Fig. 5A). In Quartiles 2 (Q2) and 3 (Q3), the accuracy remained high at 99.5% and 99.0%, respectively, with precision and recall both exceeding 88.9%. However, in Quartile 4 (Q4), formed by a pool of articles with higher similarity to the target articles, there was a notable decline in precision to 30%, despite recall remaining at 100%, attributed to an unsurprising increase in false positives, and F1 score was down to 46.1%.

The concise prompt strategy yielded comparable overall accuracy but with more variation in precision and false-positive rates (Table 8).

Table 8. GPT-3.5 Turbo (Concise Prompt).

Quartile	TN	FP	FN	TP
Q1	185	6	1	8
Q2	188	3	1	8
Q3	184	7	1	8
Q4	162	29	1	8

In Q1, the model achieved an accuracy of 96.5%, a precision of 57.1%, a recall of 88.9%, and an F1-Score of 69.44% (Fig. 5B). For Q2, precision improved to 72.7%, with accuracy reaching 98.0%, indicating better filtering of irrelevant articles. In Q3, precision dropped to 53.3%, with an F1-Score of 66.7%, highlighting challenges in balancing false positives and true positives. Q4 exhibited the most significant drop in performance, with precision reduced to 21.6% and an F1-Score of 34.8%, largely due to an increase in false positives.

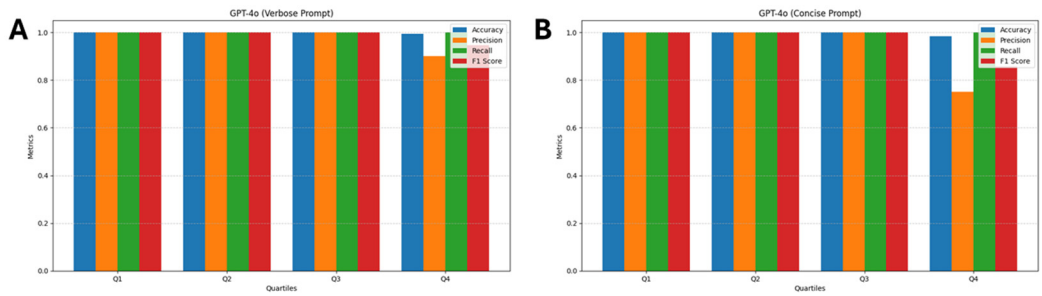
These findings underscore that the soft prompt strategy outperformed the concise prompt in terms of precision and F1-Score across most quartiles, though the concise prompt occasionally achieved better TN counts. Both strategies faced challenges in Quartile 4, particularly with managing false positives, which adversely affected precision.

3.4. GPT-4o

The GPT-4o model with the verbose prompt demonstrated consistently excellent performance across all quartiles, with near-perfect accuracy, precision, recall, and F1-Scores (Table 9).

Table 9. GPT-4o (Verbose Prompt).

Quartile	TN	FP	FN	TP
Q1	191	0	0	9
Q2	191	0	0	9
Q3	191	0	0	9
Q4	190	1	0	9



 doi:10.20944/preprints202501.2029.v1

Figure 6. Performance metrics (Accuracy, Precision, Recall, and F1 Score) for GPT-4o evaluated using two different prompt styles: (A) Verbose Prompt and (B) Concise Prompt. Metrics are presented across quartiles (Q1–Q4), with Accuracy represented in blue, Precision in orange, Recall in green, and F1 Score in red.

In Quartile 1 (Q1), the model achieved 100% accuracy, precision, and recall, resulting in an F1-Score of 100% (Fig. 6A). Similar results were observed in Quartiles 2 and 3 (Q2 and Q3), where the model exhibited flawless performance, correctly identifying all true positives (TP) and true negatives (TN) without any false positives (FP) or false negatives (FN). For Quartile 4 (Q4), there was a minimal drop in precision to 90%, attributed to 1 false positive. Despite this, recall remained at 100%, and the overall F1-Score was 94.7%.

Using the concise prompt, GPT-4o maintained consistently high performance, with accuracy and recall at 100% across Quartiles 1 to 3 (Table 10).

Table 10. GPT-4o (Concise Prompt).

Quartile	TN	FP	FN	TP
Q1	191	0	0	9
Q2	191	0	0	9
Q3	191	0	0	9
Q4	188	3	0	9

However, in Quartile 4, the model encountered three false positives, reducing precision to 75% with an F1-Score of 85.7% (Fig. 6B). Across all quartiles, the model demonstrated robust classification capabilities, consistently identifying true positives without missing any relevant articles (FN = 0).

4. Discussion

This study investigated a multi-step approach for automating the initial stages of article screening in a systematic review on periodontal regeneration as a tool to compare the performance of different Large Language Models (LLMs). The first step involved creating experimental datasets to measure LLMs' performance in classifying the articles, by collecting thematically related articles from Medline and ranking and clustering them by their semantic similarity to nine target RCTs identified in a recent systematic review [17] using sentence embeddings; this stage effectively split an otherwise very large corpus (>16000) into four quartiles of increasing similarity. The second step used LLMs to apply PICOT-based inclusion criteria, examining how single-prompt instructions, double-prompt procedures, and verbose versus concise prompts influenced performance.

The three lower-end LLMs (OpenHermes (7B), Flan T5, and GPT-2) showed poor performance when dealing with the standard prompt; OpenHermes failed to identify any target article with the only exception of Q1 dataset, i.e. the dataset with the lowest resemblance to the target article, where 1 target article was identified. Flan T5 performed only marginally better, identifying 1 target articles out of every dataset. GPT-2 accepted all articles in the dataset, completely missing the task.

These models were then tested through a double-prompt strategy, in which the first prompt broadly filtered out articles that did not mention essential interventions (Emdogain or bone graft), and the second prompt then applied the detailed inclusion criteria. With OpenHermes, the double-prompt procedure proved ineffective, as it did not improve any metrics. Flan T5 responded differently, initially excelling in the first prompt by retaining all truly relevant studies, only to underperform in the second step by failing to accept them again. This inverted outcome produced perfect recall in the broad filter but then caused complete or near-complete rejection of target articles in the detailed evaluation, highlighting the model's sensitivity to re-contextualization between prompt stages.

GPT-2 did reject a subset of articles in the initial filter, yet still remained heavily biased toward inclusion in the second step. Although GPT-2, when challenged with a simpler prompt, avoided missing any relevant articles, its over-acceptance burdened the process with excessive false positives, reflecting a trade-off that might be tolerable only if minimizing missed studies is the overriding priority, or as a general screening to short-list the articles to examine using a cheaper model.

 doi:10.20944/preprints202501.2029.v1

The more advanced GPT-3.5 Turbo and GPT-4o were primarily tested with a soft, verbose prompt that instructed acceptance if there was any possibility of meeting inclusion criteria, unless these criteria were explicitly violated. Both models used a more concise variation as well. Given their performance, no double prompt was deemed necessary. GPT-3.5 Turbo, when prompted verbosely, consistently captured all or nearly all target articles in Quartiles 1–3 and had only minor difficulties in Quartile 4, where a more substantial number of false positives appeared. This behavior indicated that while GPT-3.5 Turbo is generally reliable, increasingly complex or topic-similar abstracts (as it could happen in real-life datasets generated by accurate search strategies) can lead it to err on the side of inclusion. GPT-4o, by contrast, demonstrated near-perfect precision and recall in Quartiles 1–3, with only a handful of false positives in the highest-similarity quartile. It showed relatively little variation between verbose and concise prompts, indicating better capacity to handle fine-grained inclusion and exclusion criteria consistently. The near-ideal performance of GPT-4o in most quartiles did not completely avoid misclassifications, but it significantly reduced the tension between missing relevant articles and accepting too many irrelevant ones.

These differences among models and prompt variants underscore the importance of carefully selecting and configuring LLMs to conduct automated screening of the literature for systematic reviews. Smaller or older-generation models such as OpenHermes (7B), Flan T5, and GPT-2 are prone either to exclude truly relevant articles (particularly with multi-step prompts) or to accept far too many irrelevant ones, and they are very sensitive to changes in prompt design. While the double-prompt approach was introduced to filter out obviously irrelevant papers, it occasionally undermined performance when the model failed to maintain consistent logic between the two stages. However, our data show that both Flan T5 and GPT-2 can effectively handle a simple prompt (like the one we used for the initial filtering in the double prompt approach) and are therefore still probably usable for initial literature screenings. By contrast, more advanced LLMs like GPT-3.5 Turbo and GPT-4o appear to manage single-prompt scenarios effectively, especially when prompts are fine-tuned to emphasize recall but not at the cost of excessive false positives.

However, our datasets have a limited size, so relying solely on these models without human verification still carries risk in real screenings, particularly in more nuanced systematic reviews where abstracts can be vague or incomplete.

Overall, our findings suggest that several LLMs carry significative potential to handle the automated screening of the biomedical literature, and even smaller, leaner and generally "surpassed" model can still be effective in eliminating off-topic articles. Although GPT-4o generally emerged as the most consistently reliable model in detecting relevant RCTs and minimizing misclassification, periodic human checks remain vital to mitigate residual errors.

By tailoring the prompting strategy to a given model's strengths, or by supplementing weaker models with additional filtering rules, researchers can achieve a screening pipeline that substantially reduces labor without sacrificing the comprehensiveness expected in evidence-based reviews.

5. Conclusions

In conclusion, we showed that advanced LLMs proved effective for automated screening in a systematic review on periodontal regeneration. Thanks to sentence embeddings we generated 4 increasingly difficult databases and advanced models like GPT-3.5 Turbo and GPT-4o met the dual challenge of capturing nearly all relevant RCTs while limiting false positives and thus effectively reducing human burden.

Less powerful models, while unable to handle a complex and structured prompt to conduct the literature screening, improved somewhat when subjected to a double-prompt strategy but nonetheless struggled with maintaining balanced performance across prompt stages. These outcomes highlight the critical role of prompt design, and model selection in optimizing both recall and precision. In our setting, GPT-4o delivered near-ideal outcomes in all quartiles.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Table S1. Metrics for Open Hermes (Double Prompt - Initial Filtering); Table S2: Metrics

for Open Hermes (Double Prompt - Detailed Evaluation); Table S3: Metrics for Flan T5 (Single Prompt); Table S4: GPT-2 Single Prompt Metrics.

Author Contributions: Conceptualization, C.G., M.M. and E.C..; methodology, C.G.; software, C.G.; formal analysis, C.G. and M.T.C.; data curation, S.G. and M.T.C.; writing—original draft preparation, C.G. and M.M.; writing—review and editing, S.G. and E.C.; All the authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mulrow, C.D. Systematic Reviews: Rationale for Systematic Reviews. *BMJ* **1994**, *309*, 597–599, doi:10.1136/bmj.309.6954.597.
2. Betrán, A.P.; Say, L.; Gülmezoglu, A.M.; Allen, T.; Hampson, L. Effectiveness of Different Databases in Identifying Studies for Systematic Reviews: Experience from the WHO Systematic Review of Maternal Morbidity and Mortality. *BMC Med Res Methodol* **2005**, *5*, 6, doi:10.1186/1471-2288-5-6.
3. Grivell, L. Mining the Bibliome: Searching for a Needle in a Haystack? *EMBO Rep* **2002**, *3*, 200–203, doi:10.1093/embo-reports/kvf059.
4. Landhuis, E. Scientific Literature: Information Overload. *Nature* **2016**, *535*, 457–458.
5. Scells, H.; Zuccon, G.; Koopman, B.; Deacon, A.; Azzopardi, L.; Geva, S. Integrating the Framing of Clinical Questions via PICO into the Retrieval of Medical Literature for Systematic Reviews. In Proceedings of the Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; ACM: New York, NY, USA, November 6 2017; pp. 2291–2294.
6. Anderson, N.K.; Jayaratne, Y.S.N. Methodological Challenges When Performing a Systematic Review. *The European Journal of Orthodontics* **2015**, *37*, 248–250, doi:10.1093/ejo/cjv022.
7. Dennstädt, F.; Zink, J.; Putora, P.M.; Hastings, J.; Cihoric, N. Title and Abstract Screening for Literature Reviews Using Large Language Models: An Exploratory Study in the Biomedical Domain. *Syst Rev* **2024**, *13*, 158, doi:10.1186/s13643-024-02575-4.
8. Khraisha, Q.; Put, S.; Kappenberg, J.; Warraitch, A.; Hadfield, K. Can Large Language Models Replace Humans in Systematic Reviews? Evaluating <scp>GPT</Scp> -4's Efficacy in Screening and Extracting Data from Peer-reviewed and Grey Literature in Multiple Languages. *Res Synth Methods* **2024**, *15*, 616–626, doi:10.1002/jrsm.1715.
9. Dai, Z.-Y.; Shen, C.; Ji, Y.-L.; Li, Z.-Y.; Wang, Y.; Wang, F.-Q. Accuracy of Large Language Models for Literature Screening in Systematic Reviews and Meta-Analyses 2024.
10. Delgado-Chaves, F.M.; Jennings, M.J.; Atalaia, A.; Wolff, J.; Horvath, R.; Mamdouh, Z.M.; Baumbach, J.; Baumbach, L. Transforming Literature Screening: The Emerging Role of Large Language Models in Systematic Reviews. *Proceedings of the National Academy of Sciences* **2025**, *122*, doi:10.1073/pnas.2411962122.
11. Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; Lenert, L.A. The Emergence of Large Language Models (LLM) as a Tool in Literature Reviews: An LLM Automated Systematic Review. *arXiv preprint arXiv:2409.04600* **2024**.
12. Elliott, J.H.; Synnot, A.; Turner, T.; Simmonds, M.; Akl, E.A.; McDonald, S.; Salanti, G.; Meerpohl, J.; MacLehose, H.; Hilton, J.; et al. Living Systematic Review: 1. Introduction—the Why, What, When, and How. *J Clin Epidemiol* **2017**, *91*, 23–30, doi:10.1016/j.jclinepi.2017.08.010.
13. Ren, J.; Fok, M.R.; Zhang, Y.; Han, B.; Lin, Y. The Role of Non-Steroidal Anti-Inflammatory Drugs as Adjuncts to Periodontal Treatment and in Periodontal Regeneration. *J Transl Med* **2023**, *21*, 149, doi:10.1186/s12967-023-03990-2.

14. Mijiritsky, E.; Assaf, H.D.; Peleg, O.; Shacham, M.; Cerroni, L.; Mangani, L. Use of PRP, PRF and CGF in Periodontal Regeneration and Facial Rejuvenation—A Narrative Review. *Biology (Basel)* **2021**, *10*, 317, doi:10.3390/biology10040317.
15. Miron, R.J.; Moraschini, V.; Estrin, N.E.; Shibli, J.A.; Cosgarea, R.; Jepsen, K.; Jervøe-Storm, P.; Sculean, A.; Jepsen, S. Periodontal Regeneration Using Platelet-rich Fibrin. Furcation Defects: A Systematic Review with Meta-analysis. *Periodontol 2000* **2024**, doi:10.1111/prd.12583.
16. Woo, H.N.; Cho, Y.J.; Tarafder, S.; Lee, C.H. The Recent Advances in Scaffolds for Integrated Periodontal Regeneration. *Bioact Mater* **2021**, *6*, 3328–3342, doi:10.1016/j.bioactmat.2021.03.012.
17. Fidan, I.; Labreuche, J.; Huck, O.; Agossa, K. Combination of Enamel Matrix Derivatives with Bone Graft vs Bone Graft Alone in the Treatment of Periodontal Intrabony and Furcation Defects: A Systematic Review and Meta-Analysis. *Oral Health Prev Dent* **2024**, *22*, 655–664.
18. Reimers, N.; Gurevych, I. Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. *arXiv preprint arXiv:1908.10084* **2019**.
19. Stankevičius, L.; Lukoševičius, M. Extracting Sentence Embeddings from Pretrained Transformer Models. *Applied Sciences* **2024**, *14*, 8887, doi:10.3390/app14198887.
20. Li, B.; Han, L. Distance Weighted Cosine Similarity Measure for Text Classification. In; 2013; pp. 611–618.
21. Cichosz, P. Assessing the Quality of Classification Models: Performance Measures and Evaluation Procedures. *Open Engineering* **2011**, *1*, 132–158, doi:10.2478/s13531-011-0022-9.
22. Cottam, J.A.; Heller, N.C.; Ebsch, C.L.; Deshmukh, R.; Mackey, P.; Chin, G. Evaluation of Alignment: Precision, Recall, Weighting and Limitations. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data); IEEE, December 10 2020; pp. 2513–2519.
23. Abbade, L.P.F.; Wang, M.; Sriganesh, K.; Mbuagbaw, L.; Thabane, L. Framing of Research Question Using the PICOT Format in Randomised Controlled Trials of Venous Ulcer Disease: A Protocol for a Systematic Survey of the Literature. *BMJ Open* **2016**, *6*, e013175, doi:10.1136/bmjopen-2016-013175.
24. Scheyer, E.T.; Velasquez-Plata, D.; Brunsvold, M.A.; Lasho, D.J.; Mellonig, J.T. A Clinical Comparison of a Bovine-Derived Xenograft Used Alone and in Combination with Enamel Matrix Derivative for the Treatment of Periodontal Osseous Defects in Humans. *J Periodontol* **2002**, *73*, 423–432, doi:10.1902/jop.2002.73.4.423.
25. Sculean, A.; Chiantella, G.C.; Windisch, P.; Gera, I.; Reich, E. Clinical Evaluation of an Enamel Matrix Protein Derivative (Emdogain) Combined with a Bovine-Derived Xenograft (Bio-Oss) for the Treatment of Intrabony Periodontal Defects in Humans. *Int J Periodontics Restorative Dent* **2002**, *22*, 259–267.
26. Sculean, A.; Barbé, G.; Chiantella, G.C.; Arweiler, N.B.; Berakdar, M.; Brecx, M. Clinical Evaluation of an Enamel Matrix Protein Derivative Combined with a Bioactive Glass for the Treatment of Intrabony Periodontal Defects in Humans. *J Periodontol* **2002**, *73*, 401–408, doi:10.1902/jop.2002.73.4.401.
27. Hoidal, M.J.; Grimard, B.A.; Mills, M.P.; Schoolfield, J.D.; Mellonig, J.T.; Mealey, B.L. Clinical Evaluation of Demineralized Freeze-Dried Bone Allograft with and without Enamel Matrix Derivative for the Treatment of Periodontal Osseous Defects in Humans. *J Periodontol* **2008**, *79*, 2273–2280, doi:10.1902/jop.2008.080259.
28. Aspriello, S.D.; Ferrante, L.; Rubini, C.; Piemontese, M. Comparative Study of DFDBA in Combination with Enamel Matrix Derivative versus DFDBA Alone for Treatment of Periodontal Intrabony Defects at 12 Months Post-Surgery. *Clin Oral Investig* **2011**, *15*, 225–232, doi:10.1007/s00784-009-0369-y.
29. Jaiswal, R.; Deo, V. Evaluation of the Effectiveness of Enamel Matrix Derivative, Bone Grafts, and Membrane in the Treatment of Mandibular Class II Furcation Defects. *Int J Periodontics Restorative Dent* **2013**, *33*, e58-64, doi:10.11607/prd.1428.
30. Peres, M.F.S.; Ribeiro, E.D.P.; Casarin, R.C. V; Ruiz, K.G.S.; Junior, F.H.N.; Sallum, E.A.; Casati, M.Z. Hydroxyapatite/β-Tricalcium Phosphate and Enamel Matrix Derivative for Treatment of Proximal Class II Furcation Defects: A Randomized Clinical Trial. *J Clin Periodontol* **2013**, *40*, 252–259, doi:10.1111/jcpe.12054.
31. Queiroz, L.A.; Santamaria, M.P.; Casati, M.Z.; Ruiz, K.S.; Nociti, F.; Sallum, A.W.; Sallum, E.A. Enamel Matrix Protein Derivative and/or Synthetic Bone Substitute for the Treatment of Mandibular Class II Buccal Furcation Defects. A 12-Month Randomized Clinical Trial. *Clin Oral Investig* **2016**, *20*, 1597–1606, doi:10.1007/s00784-015-1642-x.

32. Lee, J.-H.; Kim, D.-H.; Jeong, S.-N. Adjunctive Use of Enamel Matrix Derivatives to Porcine-Derived Xenograft for the Treatment of One-Wall Intrabony Defects: Two-Year Longitudinal Results of a Randomized Controlled Clinical Trial. *J Periodontol* **2020**, *91*, 880–889, doi:10.1002/JPER.19-0432.
33. White, J. PubMed 2.0. *Med Ref Serv Q* **2020**, *39*, 382–387, doi:10.1080/02763869.2020.1826228.
34. Chapman, B.; Chang, J. Biopython: Python Tools for Computational Biology. *ACM Sigbio Newsletter* **2000**, *20*, 15–19.
35. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.
36. Mckinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the Proceedings of the 9th Python in Science Conference; van der Walt, S., Millman, J., Eds.; 2010; pp. 51–56.
37. Thakkar, H.; Manimaran, A. Comprehensive Examination of Instruction-Based Language Models: A Comparative Analysis of Mistral-7B and Llama-2-7B. In Proceedings of the 2023 International Conference on Emerging Research in Computational Science (ICERCS); IEEE, 2023; pp. 1–6.
38. Oza, J.; Yadav, H. Enhancing Question Prediction with Flan T5-A Context-Aware Language Model Approach. *Authorea Preprints* **2023**.
39. Patwardhan, I.; Gandhi, S.; Khare, O.; Joshi, A.; Sawant, S. A Comparative Analysis of Distributed Training Strategies for GPT-2. *arXiv preprint arXiv:2405.15628* **2024**.
40. Williams, C.Y.K.; Miao, B.Y.; Butte, A.J. Evaluating the Use of GPT-3.5-Turbo to Provide Clinical Recommendations in the Emergency Department. *medRxiv* **2023**, 2010–2023.
41. Islam, R.; Moushi, O.M. Gpt-4o: The Cutting-Edge Advancement in Multimodal Llm. *Authorea Preprints* **2024**.
42. Cao, C.; Sang, J.; Arora, R.; Kloosterman, R.; Cecere, M.; Gorla, J.; Saleh, R.; Chen, D.; Drennan, I.; Teja, B. Prompting Is All You Need: LLMs for Systematic Review Screening. *medRxiv* **2024**, 2024–2026.
43. Bisong, E. Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Bisong, E., Ed.; Apress: Berkeley, CA, 2019; pp. 59–64 ISBN 978-1-4842-4470-8.
44. Sentence-Transformers/All-Mpnet-Base-V2 Available online: https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (accessed on 10 February 2024).
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. 2011. *Available online: scikit-learn. org (accessed on 20 December 2021)* **2019**.