Review

How to Write Effective Prompts for Screening Biomedical Literature Using Large Language Models

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

Posted Date: 6 February 2025

doi: 10.20944/preprints202502.0396.v1

Keywords: systematic reviews; prompt engineering; GPT; zero-shot learning; biomedical screening

Review

How to Write Effective Prompts for Screening Biomedical Literature Using Large Language Models

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

[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: Large language models (LLMs) have emerged as powerful tools for (semi-)automating the initial screening of abstracts in systematic reviews, offering the potential to significantly reduce the manual burden on research teams. This paper provides a broad overview of prompt engineering principles and highlights how traditional PICO (Population, Intervention, Comparison, Outcome) criteria can be converted into actionable instructions for LLMs. We analyze the trade-offs between "soft" prompts, which maximize recall by accepting articles unless they explicitly fail an inclusion requirement, and "strict" prompts, which demand explicit evidence for every criterion. Using a periodontics case study, we illustrate how prompt design affects recall, precision, and overall screening efficiency, and discuss metrics (accuracy, precision, recall, F1 score) to evaluate performance. We also examine common pitfalls, such as overly lengthy prompts or ambiguous instructions, and underscore the continuing need for expert oversight to mitigate hallucinations and biases inherent in LLM outputs. Finally, we explore emerging trends, including multi-stage screening pipelines and fine-tuning, while noting ethical considerations related to data privacy and transparency. By applying systematic prompt engineering and rigorous evaluation, researchers can optimize LLM-based screening processes, allowing for faster and more comprehensive evidence synthesis across biomedical disciplines.

Keywords: systematic reviews; prompt engineering; GPT; zero-shot learning; biomedical screening

1. Introduction

Systematic reviews and meta-analyses are vital tools in evidence-based medicine because they synthesize data from multiple studies to provide robust insights into the effectiveness, safety, or comparative value of different healthcare interventions [1–3]. Yet one of the most demanding aspects of conducting a systematic review is screening the often thousands of abstracts retrieved from broad database searches—a process that is time-intensive, cognitively fatiguing, and subject to human error [4]. This workload can pose a substantial barrier to timely and accurate evidence synthesis, especially as the pace of publishing in biomedical research accelerates exponentially, and the need for constant updates becomes pressing [5].

Systematic reviews typically begin by searching databases such as PubMed, Embase, Scopus, or the Cochrane Library with highly sensitive queries [6]. These queries are often broad so as not to miss potentially relevant studies [7]. The result can be thousands of unique hits and normally more than one independent human reviewers then assess titles and abstracts, discarding studies that clearly do not meet the inclusion criteria [8]. Even in well-coordinated teams, this process can be lengthy and prone to errors from fatigue or inconsistent interpretations [9].

A structured framework commonly used in clinical research for defining questions and inclusion criteria is PICO (Population, Intervention, Comparison, Outcome) [10]. Researchers have diversely expanded this tool to PICOT (adding Time or Type of Study [11]), PICOS (adding Study Design), or reformulated it to SPIDER (Sample, Phenomenon of interest, Design, Evaluation, Research type) [12,13]. Generally speaking, PICO is a versatile tool, which can be used to better formulate a relevant clinical question, or to structure a systematic literature search [14]. When screening for relevant studies for a systematic review, these PICO criteria can effectively become filters to summarize the requirements that must be met for inclusion. If an article's abstract meets certain criteria—e.g., "adult population," "RCT design," "minimum six-month follow-up"—it moves forward, whereas articles that fail these requirements are excluded [15].

Recent innovations in artificial intelligence (AI) and natural language processing (NLP) have led to the development of several tools to help scholars conduct easier and faster systematic reviews [16–18]. Recently, the rise of large language models (LLMs) capable of generating human-like text and performing complex textual reasoning tasks has opened the way to unexpected possibilities [19]. Among these, the GPT (Generative Pre-trained Transformer) family—particularly GPT-3.5 and GPT-4—[20], and, more recently the Deepseek family [21] have attracted significant attention for their remarkable ability to interpret and produce language at a near-human level. These models promise to automate or semi-automate various steps in the systematic review workflow, including the notoriously laborious initial screening of titles and abstracts [20]. To put it briefly, researchers have started to use LLMs to comb through the literature and identify relevant articles for systematic reviews, instead of leaving this task to human teams [22].

If an LLM can reliably read an abstract and decide whether it meets the PICO criteria, researchers might only need to examine a reduced set of articles that the model deemed "ACCEPT." This approach cuts the time spent on obviously irrelevant articles and dramatically reduce the screening burden.

However, successfully deploying LLMs for such screening tasks requires more than simply providing these algorithms an abstract and asking, "Should this be included?" All LLM models need a precise set of instructions, known as prompt, to operate effectively. A correct prompt engineering, i.e. crafting clear, context-rich instructions, is therefore key to produce the desired output [23] and recent publication have highlighted the relevance of these skills for the medical profession [24,25]. Prompt engineering becomes especially important in situations like zero-shot or few-shot classification, where the model is not fine-tuned on a large set of domain-specific examples but must instead rely on the prompt itself to guide its reasoning [26].

This article will briefly explore what large language models are, the differences between zero-shot and few-shot approaches, and how these concepts apply to the screening of biomedical literature. After establishing these foundations, we shall provide a detailed tutorial on writing prompts—ranging from softer to stricter styles—that guide LLMs to accept or reject abstracts based on PICO-based inclusion criteria. We also discuss best practices, pitfalls, iterative testing, and potential future directions of AI-augmented systematic review workflows. Our hope is that researchers in dentistry and medicine fields ranging from periodontology to oncology, and beyond, can use these techniques to accelerate their reviews, maintain high standards of methodological rigor, and ultimately enhance evidence-based decision-making in healthcare.

2. Understanding Large Language Models Beyond the Buzzwords

2.1. The Transformer Revolution

The evolution of Large Language Models (LLMs) possibly began with Statistical Language Models (SLMs) in the 1990s, which relied on probabilistic methods to predict word sequences. These were succeeded by Neural Language Models (NLMs) in the early 2010s, utilizing deep learning to better capture word relationships. The introduction of Pre-trained Language Models (PLMs), including BERT and GPT, revolutionized natural language processing by applying unsupervised training on massive corpora before fine-tuning for specific applications. The modern era of LLMs probably began with OpenAI's GPT series, marking a shift toward massive-scale transformer-based models capable of generating human-like text [19].

In 2018, OpenAI introduced GPT-1, demonstrating the potential of generative pre-training to improve downstream tasks [27]. One year later, GPT-2 expanded the parameter count to 1.5 billion, yielding notably improved fluency and coherence, though it remained prone to factual inaccuracies [28]. In 2020, OpenAI released GPT-3, a 175-billion-parameter model that popularized few-shot learning, allowing users to guide the model with minimal examples in the prompt [29]. GPT-3's ability to engage in zero-shot reasoning and process deep contextual understanding expanded its application range, though factual precision remained an issue [30]. The Transformer architecture, introduced in 2017, played a crucial role in enabling these advancements by efficiently handling long-range dependencies, outperforming recurrent-based approaches [31–33]. Transformers use a mechanism called "self-attention" to weigh the importance of different words in a context, allowing the model to handle long-range dependencies far more efficiently than earlier recurrent networks (e.g., LSTM or GRU) [34,35].

By 2022, OpenAI refined its model with GPT-3.5, improving coherence and reducing susceptibility to generating off-topic or erroneous statements [36]. This version became widely accessible through ChatGPT, significantly enhancing conversational fluency and adherence to complex instructions [37]. In 2023, OpenAI released GPT-4, also available in an optimized form (GPT-4o), further pushing the boundaries of context management and logical consistency, though its exact architecture and parameter count remain partially undisclosed. GPT-4 has shown particular promise in specialized fields like biomedical research, where tasks such as summarizing scientific articles or screening abstracts for systematic reviews benefit from its advanced natural language processing capabilities [38–40].

Around the same time, other organizations entered the LLM race, offering alternatives to OpenAI's models. Meta introduced the Llama family, including Llama 2, in semi-open-source formats, giving researchers more flexibility in local deployment [41]. Google unveiled Gemini, reported to incorporate multimodal reasoning and compete directly with GPT-4 [42]. Anthropic developed Claude, focusing on safe and transparent language generation [43], while Alibaba introduced Qwen, emphasizing multilingual capabilities and domain specialization [44]. Additional models, such as DeepSeek, reflect the trend of expanding LLM architectures, with each new system claiming unique advantages—whether in safety, scale, multilingual support, or reduced hallucinations [45]. Though many of these emerging models have yet to be extensively tested for biomedical applications, they provide alternative pathways, particularly for domain-specific automated screening tasks.

Importantly, some of these LLMs can be downloaded and run locally, provided sufficient computational resources are available [46], or deployed using platforms like Google Colab with access to high-end GPUs [47]. Others, including GPT-3.5 and higher, are primarily accessed via Application Programming Interfaces (APIs), allowing researchers to use powerful hosted models without requiring dedicated hardware [48]. In some cases, such as the more advanced GPT-o1, or GPT-o3 model families, limited API access is currently available only to selected users. Local deployment appeals to users needing full control over the model's environment or those concerned about data privacy, while API-based access enables scalability and ease of integration. This distinction is particularly relevant for systematic review projects, where considerations of privacy, cost, and computational power influence the choice of model deployment.

What sets large language models apart is their scale [49]. By training on trillions of tokens (words or subword units) [50], LLM develop a remarkably broad understanding of language, style, factual knowledge, and even some reasoning capabilities [45,51–53]. LLMs can perform tasks such as summarizing text, translating languages, answering questions, and classifying content—all without being explicitly programmed or fine-tuned for each specific task, but as a result of their massive-scale training [54].

2.2. Strengths, Caveats, and Uncharted Territories

Using LLMs to screen the biomedical literature is a promising approach, with the potential of dramatically reducing the burden currently carried by the investigators [20,22,55]. However, it should be noted that in practice, many biomedical abstracts may omit critical details (e.g., exact follow-up durations or explicit randomization methods) [56]. In such cases, the translation of PICO criteria into prompts may face challenges, necessitating fallback strategies (e.g., flagging ambiguous abstracts for human review).

Recent LLMs excel at parsing and generating text, allowing them to process a prompt and respond with coherent, contextually relevant, and sometimes creative answers, thanks to a broad knowledge base acquired during pre-training [57]. Despite these strengths, they also exhibit key limitations that can affect their utility for specialized tasks. One notable issue involves hallucinations, where the model confidently supplies factually incorrect information [58]. For example, it might reference a randomized controlled trial in "Journal X, 2018" that does not actually exist, complete with fabricated authors and findings. In the context of systematic reviews, such hallucinations can be particularly harmful if researchers mistakenly trust these invented details, thereby compromising the reliability of study selection or data extraction.

Another challenge is the sensitivity to prompting, in which minor rephrasing of instructions can produce markedly different outputs [59–61]. As an illustration, changing a guideline from "Accept if participants are X" to "Reject if participants are not X" could lead to unforeseen consequences in a given model, even though the underlying logic is similar.

A further limit arises from domain gaps, where the model's broad training does not always capture specialized terminology or conventions in fields like biomedicine—an LLM might misinterpret "CAL" (clinical attachment level in periodontics) as referring to "calories," for instance [62].

Finally, LLMs often mimic patterns in their training data rather than performing true logical inference [63], causing them to replicate style and phrasing without genuinely evaluating the content. In spite of these shortcomings, LLMs remain highly adaptable tools, capable of classifying scientific abstracts or performing other tasks with minimal additional training, provided that users structure prompts carefully and remain vigilant to potential errors.

2.3. Where to Draw the Line: Titles, Abstracts, or Full Text?

Determining which portion of an article to use for screening is a critical decision in systematic reviews, and it has substantial implications for both accuracy and efficiency [64]. Traditionally, reviewers begin by examining titles alone, which can be rapidly scanned to discard clearly irrelevant studies—like those focusing on a different patient population or unrelated interventions [9]. We have successfully used titles in numerous Topic Modeling investigations, because titles very effectively summarize the theme of a paper [65–67]. However, titles often provide few details and can lead to a high number of false positives, pushing many irrelevant items into the next stage.

The next, more informative option is the abstract [68], which is widely accessible through major free databases (e.g., Medline, Google Scholar) without any paywall restrictions, unlike the full texts which may require a fee [69]. Abstracts typically summarize the study's population, interventions, primary outcomes, and, in many cases, methodological details such as randomization or follow-up duration [70]. In many respects, abstracts represent a modern evolution of summaries; once relegated to a section at the end of an article, they have now become the predominant—and sometimes exclusive—means of conveying essential information in bibliographic searches [71]. By using this brief but comprehensive snapshot, readers, reviewers (and now, LLMs) can make an informed decision about eligibility. Screening abstracts represents a middle ground that balances depth of information with practicality: the text is concise enough for AI algorithms to parse quickly, yet usually detailed enough to capture key inclusion or exclusion criteria. In addition, abstracts are standardized in many biomedical reports, making them relatively uniform across studies [72].

A potential expansion of screening is to consider full-text articles, which can eliminate ambiguity by allowing the reviewer or LLM to access all methodological and outcome details [73]. This approach, however, poses significant logistical and computational challenges. First, acquiring full-text—especially for large search results—may require additional subscriptions or interlibrary loans, and many databases do not host the entire paper [74–77]. Second, in the case of LLMs, feeding long documents to AI systems can be computationally expensive, potentially exceeding token limits in certain large language model APIs. Full-text processing also increases inference time per article, which, for thousands of items, can become quite costly in terms of both machine time and financial cost, if using paid APIs.

For systematic reviews seeking to incorporate large language models, adopting an abstract-based screening step thus emerges as an efficient starting point, aligning well with existing workflows, cost constraints, and the easily retrievable nature of abstracts across major bibliographic databases.

3. Evaluation Metrics for Classification Tasks

3.1. The Basics of Classification Metrics

When assessing the effectiveness of LLMs in classifying abstracts for a systematic review, it is helpful to first understand four fundamental terms [78]. A true positive (TP) describes a situation where the model labels an article as relevant (ACCEPT) and the article indeed meets the inclusion criteria of the review. A false positive (FP) arises when the model also labels an article as relevant, yet that article proves to be irrelevant upon verification. By contrast, a true negative (TN) occurs if the model rejects an article (REJECT) and this rejection is correct, while a false negative (FN) is a scenario in which a relevant article is mistakenly rejected; in case of a systematic review, this is probably the most feared scenario, i.e. missing useful evidence. Literature searches for systematic reviews tend to be design to minimize FN articles, at the expenses of a higher number of FP, which can be tradionally filtered out during the subsequent manual screening.

Building upon these concepts, a range of metrics can be calculated to evaluate model performance [79]. Accuracy measures the proportion of correct predictions (both ACCEPT and REJECT) across all screened articles and is computed as:

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

Although accuracy gives a quick overview of performance, it can be misleading in highly imbalanced datasets [80]. Suppose, for instance, that the LLM is malfunctioning and simply labels every abstract as relevant. Such LLM will achieve high accuracy because it correctly captures all true positives. Yet this outcome is practically useless, as the model fails to distinguish between relevant and irrelevant studies. More revealing in this context is precision, the fraction of predicted relevant articles that are truly relevant, expressed as:

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

Recall, also referred to as sensitivity, focuses on how well the model captures the pool of genuinely relevant articles [81], calculated as

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

In systematic reviews, where missing valid studies can weaken the overall analysis, recall is often emphasized, because this measurement gauges the amount of FNs. To provide a balanced view of both precision and recall, the F1 score offers a single metric by taking the harmonic mean of these two measures [82]:

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

Balancing these metrics is particularly critical in systematic reviews, where the primary concern is to minimize false negatives so that truly relevant research is not lost. Nonetheless, a model that indiscriminately accepts too many irrelevant articles may overwhelm reviewers at the next stage of full-text assessment, thus mitigating the gains from automation. Consequently, it is important for researchers to weigh recall, precision, and other metrics carefully when designing prompts and deploying LLM-based screening strategies in order to achieve an optimal blend of efficiency and completeness.

3.2. Bringing Metrics to Life: A Practical Example

Consider an LLM applied to screening 1,000 articles. Suppose the model identifies 200 articles as relevant, of which 150 are truly relevant (TP = 150) and 50 are not (FP = 50). Meanwhile, this model misses 10 relevant studies that it incorrectly rejects (FN = 10), and correctly rejects 790 irrelevant ones (TN = 790). The relevant metrics can be therefore calculated as follows (Figure 1):

$$\text{Accuracy: } (150 + 790) / 1{,}000 = 94\%$$

$$\text{Precision: } 150 / (150 + 50) = 75\%$$

$$\text{Recall: } 150 / (150 + 10) = 93.75\%$$

$$\text{F1 Score: } 2 \times (0.75 \times 0.9375) / (0.75 + 0.9375) \approx 0.83\ (83\%)$$

In systematic reviews, the recall of nearly 94% may be acceptable, but 50 false positives remain a substantial number for full-text review. Adjusting the prompt to be stricter might lower recall but boost precision.

Advanced users could parse chain-of-thought or confidence scores, though many GPT endpoints do not expose raw probabilities. Moreover, chain-of-thought outputs can be unreliable or verbose. Typically, using discrete “ACCEPT”/“REJECT” outputs—guided by prompt instructions—suffices for systematic review screening.

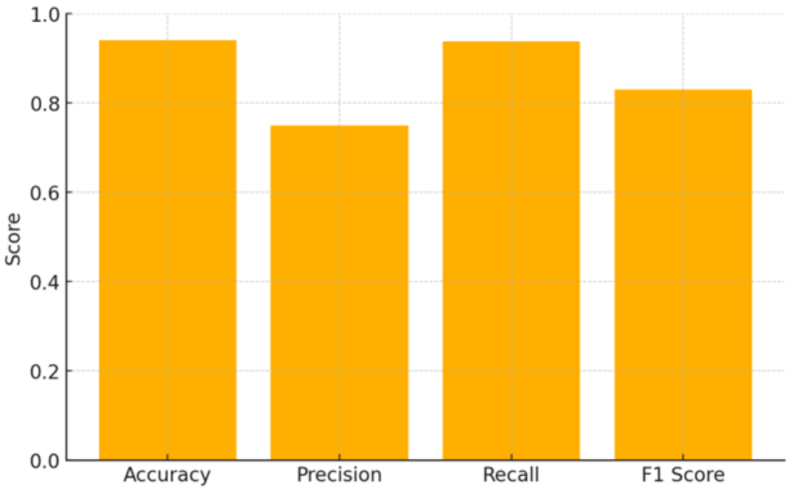

Figure 1. This histogram exemplifies the performance metrics of an LLM screening 1,000 articles. Accuracy is high (94%), but the trade-off between recall (93.75%) and precision (75%) suggests a tendency toward false positives. The F1 score (83%) highlights the balance between these factors, indicating potential for optimization through stricter prompting.

4. Speak and It Shall Be Done: The Art of Prompt Engineering

4.1. Building a Good Prompt

A prompt is essentially the set of instructions or contextual information that a user provides to an LLM in order to guide its output [83]. At its simplest, a prompt might be just a quick question typed into a search bar—yet it can also be highly detailed, specifying style, tone, structure, and even the precise format of the desired answer [84]. Generally speaking, well-structured prompts typically feature four key elements [85]. First, they contain an instruction, which defines the task the model should perform or the behavior it should adopt (e.g., "Summarize the following abstract"). Second, they include context, which offers background knowledge or domain-specific details that help the model produce more accurate and relevant responses (for example, clarifying that it is an AI assistant aiding with a systematic review about periodontal regeneration). Third, a prompt provides input data, such as the specific text or question the model must process, so it knows exactly what content to analyze (like an abstract describing an RCT). Finally, prompts often specify an output indicator, meaning they state the desired format or type of response—whether a short paragraph, a bullet list, or a more elaborate explanation.

Users can combine these elements to align an LLM's responses with their specific goals, such as translating text or classifying biomedical abstracts. In the specific context of literature screening, a prompt would incorporate context, eligibility criteria, and decision instructions, often drawing on a PICO (or PICO-like) framework. For instance, one might start by noting, "*You are assisting in a systematic review on periodontal regeneration comparing Bone Morphogenetic Proteins (BMP) and bone graft (BG)*," thereby framing the task within a particular domain. The user then outlines the key requirements—adult population (≥18 years), intrabony defects, RCT design, minimum six-month follow-up—and finally instructs the model exactly how to respond, e.g. "*If the abstract satisfies these inclusion criteria, respond with 'ACCEPT'; otherwise, respond with 'REJECT.' No additional text.*" This structure would ensure that the LLM knows precisely what features to look for, where to find them, and how to format its classification.

The structure of a prompt is crucial because LLMs do not inherently "know" how to proceed with every query; instead, they rely on these external directives to frame and interpret the data. A well-designed prompt fosters clear and consistent responses, whereas unclear or incomplete prompts can lead to confusion or unrelated output. Ultimately, prompts function as a crucial bridge between the model's broad, pretrained knowledge and the user's objectives—turning an otherwise general-purpose AI system into a targeted and more reliable tool for tasks like summarizing research findings or identifying relevant articles in a systematic review.

4.2. Zero, One, or Few Shots?

When deploying LLMs for screening tasks in systematic reviews, there are multiple strategies that vary according to the amount and nature of guidance provided in the prompt. The simplest approach is zero-shot learning, where the model is tasked with classifying new abstracts (e.g., as "ACCEPT" or "REJECT") based on the PICO criteria, without exposure to any labeled examples [86]. In this scenario, the LLM relies entirely on its pre-trained knowledge and the specific instructions within the prompt. Researchers can specify inclusion and exclusion rules—such as requiring adult patients (≥18 years old), randomized controlled trials (RCTs), or a minimum follow-up—then feed each abstract through this single prompt. The zero-shot approach is appealing for large-scale automation due to its simplicity and minimal token usage, although it may struggle in domains with ambiguous or complex criteria [87–89].

An intermediate option is the one-shot approach, which provides exactly one labeled example in the prompt. For instance, the user might include a single abstract that meets the criteria and is labeled "ACCEPT," and rely on the model to extrapolate the classification logic for all other abstracts. While this single example can clarify basic requirements, it may still be inadequate for capturing diverse edge cases or subtleties in a specialized field like periodontology.

A more robust method is few-shot learning, where two or more labeled examples appear in the prompt, illustrating both "ACCEPT" and "REJECT" decisions [90,91]. This strategy offers the model a small but representative range of scenarios, enabling it to better handle ambiguous abstracts. In systematic reviews, a few-shot approach is often feasible if a handful of known relevant and non-relevant studies exist—for instance, when updating a previous review. However, the downside is greater token consumption and increased complexity, as the prompt must accommodate not only the PICO-based instructions but also sample abstracts that illustrate borderline cases. Researchers must therefore decide if the performance gains justify the more extensive prompt design.

Recently, some practitioners have experimented with few-shot chain-of-thought prompts, where the examples include not only labeled abstracts but also brief rationales indicating why an article was accepted or rejected [92]. By exposing the LLM to the reasoning process, the model may generate more accurate or consistent classifications. Yet this approach can further inflate token usage, and revealing too much "chain-of-thought" may inadvertently cause the model to overfit the examples rather than applying general rules.

In practice, the choice among zero-shot, one-shot, few-shot, or few-shot chain-of-thought depends on the complexity of the inclusion criteria, the typical ambiguity in the abstracts, and the resources available. Zero-shot setups work well if the rules are clear and the model's underlying knowledge base is sufficient (and no known abstracts are available), whereas few-shot prompts are usually recommended when domain-specific nuances frequently arise and users already have viable examples of the abstracts they are looking for. Ultimately, researchers should test each strategy on a small validation set of abstracts to see which yields the most desirable balance of recall and precision for their particular review.

4.3. Soft Versus Strict: Setting the Bar for Inclusion

An important dimension of prompt design is determining how permissive or strict the model should be when assessing abstracts. A permissive (or "soft") prompt instructs the LLM to accept an article unless it clearly violates a predefined requirement. For example, a prompt might state that the only grounds for rejection are explicit contradictions such as an exclusively pediatric population if the inclusion criteria include age>18 years old. This approach typically yields high recall, ensuring that potentially relevant studies are not inadvertently excluded; however, it may also result in an increased number of false positives, thereby requiring more manual review later in the process.

In contrast, a strict prompt directs the LLM to reject an article unless it explicitly meets every inclusion criterion. This strategy can lower the number of false positives by demanding clear evidence—such as explicit mention of adult patients, a randomized controlled trial (RCT) design, and a specified follow-up duration—but it may inadvertently exclude relevant studies whose abstracts do not detail every required element. Given that many systematic reviews prioritize sensitivity (i.e., minimizing false negatives), relying exclusively on soft prompts can lead to an overwhelming number of articles to review in subsequent full-text screening [12]. A balanced or two-stage approach, where the first pass uses a soft prompt and the second pass applies a stricter filter, can allow for thorough screening without excessive manual follow-up.

4.4. Illustrative Examples

To show how these strategies might appear in practice, one could start with a soft prompt such as:

"You are assisting in a systematic review on periodontal regeneration comparing Bone Morphogenetic Proteins (BMP) plus any bone graft (BG) versus BG alone.

We include RCTs with at least six months of follow-up in adults (≥18 years) with intrabony or furcation defects.

If the abstract suggests or does not contradict these criteria, respond with 'ACCEPT.'

Only respond with 'REJECT' if it clearly violates any requirement—exclusively pediatric population, follow-up shorter than six months, or no mention of randomization or EMD.

Output only the word 'ACCEPT' or 'REJECT.'"

This design ensures high recall because the model will accept studies unless they definitively fail an inclusion rule. By contrast, a non-soft, stricter prompt might read:

"You are an expert reviewer screening for RCTs of BMP+BG vs. BG alone in adult periodontitis patients (≥18 years) with intrabony or furcation defects.

The study must explicitly mention randomization, adult age, combination of EMD and bone graft, a control with BG alone, and a follow-up of at least six months.

If any criterion is absent or unclear, respond 'REJECT.'

Otherwise, respond 'ACCEPT.'

No additional text."

Here, failing to see explicit mention of a required element leads directly to rejection, thus lowering false positives at the risk of overlooking abstracts that might still be relevant if the missing detail is confirmed later in the full text.

4.5. Avoiding Pitfalls with Prompt Refinement

Although advanced LLMs such as GPT-4 or GPT-4o offer promising avenues for semi-automating literature screening, several pitfalls can undermine their performance if not addressed. One frequent issue involves the use of overly long prompts, in which excessive background information dilutes the model's focus on core classification rules [93]. These models can certainly process large amounts of text, but they may become confused or deviate from their primary task if instructions are buried in lengthy or tangential material.

Another obstacle arises when prompts contain contradictory or ambiguous criteria—for instance, instructing the model to reject any abstract lacking follow-up details yet also advising it not to penalize uncertainty [94]. Such conflicting instructions often lead to inconsistent or unpredictable responses. Moreover, LLMs may sometimes hallucinate information by confidently inferring characteristics that are not actually stated, such as labeling a study "randomized" simply because it is described as "prospective." Encouraging the model to accept only abstracts that explicitly mention "RCT," "randomized," or other unambiguous terms helps reduce these erroneous leaps. Prompt engineering is seldom perfected in one pass, so an iterative process of testing, refining, and retesting the prompts with a small validation set is often very important to identify where false positives and false negatives occur and to recalibrate the instructions accordingly [95].

Despite significant advances in LLM capabilities, human oversight remains indispensable, especially for edge cases that involve unclear abstracts or unusual study designs. While the model can filter out large swaths of irrelevant articles, expert judgment is crucial in verifying the accuracy of the final screening decisions and ensuring that no critical evidence is mistakenly excluded. Although the example presented here focuses on periodontal regeneration, similar strategies are applicable to other biomedical domains such as oncology, cardiology, and infectious diseases. Furthermore, as these models continue to evolve, there is potential for deeper integration into standard review workflows, with near-instant triage of abstracts and real-time prompt adjustments based on user feedback. However, ethical and logistical concerns persist, including transparency of the screening process, data privacy, and the possibility of systematic biases in LLM training data [96]. Addressing these challenges through careful documentation of prompts, awareness of training limitations, and appropriate regulatory oversight is essential for maximizing the benefits of AI-assisted systematic reviews while maintaining rigorous scientific standards [97].

5. Future Horizons and Potential Advancements

Although the latest LLM models perform impressively in zero-shot or few-shot modes [98–100], further refinements may come from fine-tuning or advanced instruction tuning. In a fine-tuning scenario, one would gather a large corpus of labeled abstracts—designated "ACCEPT" or "REJECT" according to specific criteria—and retrain the model to internalize these rules. While this approach enhances classification accuracy, it requires substantial computational resources and high-quality training data [101,102].

Another promising approach is to implement multi-stage screening to balance high recall with manageable false positives [103]. First, a simple prompt maximizes recall by eliminating only clearly irrelevant studies, such as those on animal models or unrelated conditions. In a second pass, a stricter prompt would then apply more exacting criteria, thereby reducing the number of false positives that remain. Finally, human reviewers would verify any borderline articles that either meet all criteria but remain questionable or exhibit contradictory outcomes between the first and second passes, using their expert domain knowledge on a much narrower corpus. By layering these stages, researchers could preserve both the efficiency gains of automated screening and the rigor of expert oversight.

Language barriers may also influence LLM performance, as retrieved articles often include abstracts in multiple languages. Publication language has long been an obstacle in evidence synthesis, partly due to the lower accessibility of non-English studies [104]. While modern LLMs can process multilingual texts, their effectiveness varies across languages due to uneven training data [105–109]. Furthermore, prompt engineering must still consider differences in terminology and structure across languages, particularly when dealing with domain-specific contexts [110].

Future research will also focus on integrating LLM-based screening into existing review pipelines. This includes addressing practical challenges such as token limits, latency, computational cost, and data security, as well as developing robust interfaces that combine automated screening with human validation.

Finally, the growing influence of LLMs raises ethical and transparency issues that researchers cannot ignore [111]. These models often function as "black boxes," making it difficult to trace their decision processes, although chain of thought reasoning can improve data interpretability [112]. While they excel in text generation, verifying the factual accuracy of their outputs remains a challenge [113]. This raises ethical concerns when AI-assisted screening is used to extract clincial evidence, thus possibly influencing clinical guidelines or policies [97]. Maintaining detailed records of prompts, model versions, and modifications will be essential for accountability [96,114]. Additionally, biases in training data could lead to the underrepresentation of certain populations or interventions, subtly skewing results [115–117]. By systematically documenting AI-based screening methods and remaining vigilant about potential biases, researchers will be able to maintain the integrity of systematic reviews as these technologies become increasingly integrated into evidence synthesis.

6. Conclusions

Advanced Large language Models have a great potential to reduce the workload of screening abstracts in systematic reviews. However, achieving accurate, consistent, and transparent results hinges on writing effective prompts that translate your PICO criteria into unambiguous, testable instructions. This article has illustrated how “soft” and “non-soft” prompt strategies differ in recall and precision, discussed zero-shot versus few-shot classification approaches, and provided concrete examples of how these methodologies might be applied in a periodontal regeneration case study.

Crafting prompts is more an iterative, practical art than a one-size-fits-all procedure. The general principles—clarity, explicitness about acceptance/rejection triggers, short and consistent output formats, and thorough testing on a validation set—apply broadly across biomedical domains. Some researchers will prefer a heavily inclusive, soft prompt to ensure near-100% recall, while others may rely on a strict approach to limit false positives. Few-shot examples can guide the model more accurately for specialized or ambiguous criteria, but at the cost of longer prompts and slightly more complexity.

Although AI-assisted screening will not entirely replace human judgment, it allows systematic review teams to concentrate their time on high-value tasks such as critical appraisal of full-text articles, data extraction, and synthesis. As LLMs continue to advance and specialized domain-specific models gain traction, it is likely that such tools will become standard in systematic reviewing. Nonetheless, vigilance regarding biases, hallucinations, or contradictory outputs remains important, and final oversight by trained researchers is essential for preserving the integrity of evidence-based medicine.

By learning how to effectively craft prompts and harness the abilities of LLMs, whether in zero-shot or few-shot modes—researchers can bring greater efficiency, consistency, and scalability to systematic reviews. This, in turn, has the potential to accelerate scientific progress and facilitate more timely, comprehensive insights into the ever-expanding biomedical literature.

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: Please add: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Appendix A

Practical Guidelines for Prompt Engineering

This appendix provides a comprehensive set of guidelines for practitioners aiming to implement LLM-based screening in systematic reviews. The following step-by-step instructions are designed to ensure a systematic, transparent, and reproducible approach to crafting effective prompts, addressing both the technical and practical challenges inherent in translating clinical criteria into actionable instructions for LLMs.

A.1. Define Clear Criteria

Map PICO Elements Precisely:

- Population: Clearly state the target population (e.g., “adult patients, defined as individuals ≥18 years”).
- Intervention: Specify the intervention details (e.g., “use of EMD combined with bone graft”).
- Comparison: Define what the intervention is being compared against (e.g., “bone graft alone”).
- Outcome: Identify the primary outcome or endpoints (e.g., “clinical measures of periodontal regeneration”).
- Use Unambiguous Language: avoid vague terms. For example, write “adult patients (≥18 years)” instead of “adults.”; include specific keywords or phrases that the model can recognize as indicators of a particular criterion.
- Incorporate Contextual Examples (if necessary): where appropriate, provide a brief example or keyword list that exemplifies the criterion, ensuring the model captures the intended meaning.

A.2. Choose an Appropriate Prompting Approach

Zero-Shot vs. One-Shot vs. Few-Shot:

Zero-Shot: Use when there are no available abstracts. This approach is suitable for straightforward criteria but may be less effective for ambiguous abstracts.

One-Shot: Provide a single labeled abstract to clarify the expected output, useful when the criteria are mostly clear but benefit from a guiding instance.

Few-Shot: Include multiple examples (both “ACCEPT” and “REJECT” cases) to better capture the variability in abstract presentations, particularly in complex domains.

A.3. Employ Iterative Testing and Refinement

Pilot Testing: Use a small, representative validation set of abstracts to assess prompt performance.

doi:10.20944/preprints202502.0396.v1

Evaluate outcomes using key metrics such as precision, recall, and F1 score.

Feedback Loop: Adjust prompt wording, examples, or criteria specificity based on observed performance issues (e.g., high false positive or false negative rates).

Document each iteration, noting the changes made and the impact on model performance.

Performance Metrics: Regularly compute performance metrics to ensure that modifications are moving the prompt toward an optimal balance between recall and precision.

A.4. Integrate Human Oversight

Secondary Review Process: Establish a protocol for human review of ambiguous or borderline abstracts flagged by the LLM; define criteria or thresholds (e.g., a confidence score below a set level or inconsistent outputs) that trigger manual verification.

Expert Involvement: Use human feedback to further refine the prompt and improve model accuracy over time.

Appendix B

Comprehensive Example: Zero-Shot Soft Prompt

Below is a more extensive example that researchers can adapt to their own PICO(T) criteria. It is a zero-shot, soft prompt for screening:

System:

You are an AI assistant helping with a systematic review on periodontal regeneration.

User Prompt:

You will decide if each article should be ACCEPTED or REJECTED based on the following criteria:

Population (P): Adult patients (≥18 years) with intrabony or furcation periodontal defects. If the abstract does not mention age, or does not clearly describe non-adult populations, do not penalize.

Intervention (I): Must involve enamel matrix derivative (EMD) combined with a bone graft material (BG). If either EMD or BG is implied or partially mentioned, do not penalize.

Comparison (C): Ideally a group that uses BG alone, or some control lacking EMD. If not stated but not contradicted, do not penalize.

Outcomes (O): Must measure clinical attachment level (CAL) gain or probing depth (PD) reduction, or at least mention standard periodontal parameters. If the abstract does not state outcomes explicitly but mentions "periodontal regeneration," do not penalize.

Study design: Must be an RCT or strongly imply random allocation. If uncertain, do not penalize.

Follow-up: Minimum 6 months. If not stated or unclear, do not penalize unless it says <6 months.

Decision Rule: If no criterion is explicitly violated, respond only with "ACCEPT." If any criterion is clearly contradicted (e.g., non-randomized, pediatric population, <6 months follow-up), respond with "REJECT." Provide no additional explanation.

Title: {title}

Abstract: {abstract}

This prompt is relatively permissive ensuring high recall. It also captures the essence of PICO while acknowledging that abstracts can be incomplete. Researchers implementing a pipeline can start here, accept all articles GPT classifies as "ACCEPT," then briefly review them manually or apply a second, stricter screening if desired.

References

1. Mulrow, C.D. Systematic Reviews: Rationale for Systematic Reviews. *BMJ* **1994**, *309*, 597–599, doi:10.1136/bmj.309.6954.597.
2. Dickersin, K.; Scherer, R.; Lefebvre, C. Systematic Reviews: Identifying Relevant Studies for Systematic Reviews. *Bmj* **1994**, *309*, 1286–1291.
3. Parums, D. V Review Articles, Systematic Reviews, Meta-Analysis, and the Updated Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Guidelines. *Med Sci Monit* **2021**, *27*, e934475-1.
4. Waffenschmidt, S.; Knelangen, M.; Sieben, W.; Bühn, S.; Pieper, D. Single Screening versus Conventional Double Screening for Study Selection in Systematic Reviews: A Methodological Systematic Review. *BMC Med Res Methodol* **2019**, *19*, 132, doi:10.1186/s12874-019-0782-0.
5. 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.
6. Bramer, W.M.; Rethlefsen, M.L.; Kleijnen, J.; Franco, O.H. Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study. *Syst Rev* **2017**, *6*, 245, doi:10.1186/s13643-017-0644-y.
7. Scells, H.; Zuccon, G. Generating Better Queries for Systematic Reviews. In Proceedings of the The 41st international ACM SIGIR conference on research & development in information retrieval; 2018; pp. 475–484.
8. Cooper, C.; Booth, A.; Varley-Campbell, J.; Britten, N.; Garside, R. Defining the Process to Literature Searching in Systematic Reviews: A Literature Review of Guidance and Supporting Studies. *BMC Med Res Methodol* **2018**, *18*, 85, doi:10.1186/s12874-018-0545-3.
9. Gupta, S.; Rajiah, P.; Middlebrooks, E.H.; Baruah, D.; Carter, B.W.; Burton, K.R.; Chatterjee, A.R.; Miller, M.M. Systematic Review of the Literature: Best Practices. *Acad Radiol* **2018**, *25*, 1481–1490, doi:10.1016/j.acra.2018.04.025.
10. Eriksen, M.B.; Frandsen, T.F. The Impact of Patient, Intervention, Comparison, Outcome (PICO) as a Search Strategy Tool on Literature Search Quality: A Systematic Review. *Journal of the Medical Library Association* **2018**, *106*, doi:10.5195/jmla.2018.345.
11. 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.
12. Methley, A.M.; Campbell, S.; Chew-Graham, C.; McNally, R.; Cheraghi-Sohi, S. PICO, PICOS and SPIDER: A Comparison Study of Specificity and Sensitivity in Three Search Tools for Qualitative Systematic Reviews. *BMC Health Serv Res* **2014**, *14*, 579, doi:10.1186/s12913-014-0579-0.
13. Cooke, A.; Smith, D.; Booth, A. Beyond PICO. *Qual Health Res* **2012**, *22*, 1435–1443, doi:10.1177/1049732312452938.
14. Frandsen, T.F.; Bruun Nielsen, M.F.; Lindhardt, C.L.; Eriksen, M.B. Using the Full PICO Model as a Search Tool for Systematic Reviews Resulted in Lower Recall for Some PICO Elements. *J Clin Epidemiol* **2020**, *127*, 69–75, doi:10.1016/j.jclinepi.2020.07.005.
15. Brown, D. A Review of the PubMed PICO Tool: Using Evidence-Based Practice in Health Education. *Health Promot Pract* **2020**, *21*, 496–498, doi:10.1177/1524839919893361.
16. Pham, B.; Jovanovic, J.; Bagheri, E.; Antony, J.; Ashoor, H.; Nguyen, T.T.; Rios, P.; Robson, R.; Thomas, S.M.; Watt, J.; et al. Text Mining to Support Abstract Screening for Knowledge Syntheses: A Semi-Automated Workflow. *Syst Rev* **2021**, *10*, 156, doi:10.1186/s13643-021-01700-x.
17. Chai, K.E.K.; Lines, R.L.J.; Gucciardi, D.F.; Ng, L. Research Screener: A Machine Learning Tool to Semi-Automate Abstract Screening for Systematic Reviews. *Syst Rev* **2021**, *10*, 93, doi:10.1186/s13643-021-01635-3.
18. Gates, A.; Johnson, C.; Hartling, L. Technology-Assisted Title and Abstract Screening for Systematic Reviews: A Retrospective Evaluation of the Abstrackr Machine Learning Tool. *Syst Rev* **2018**, *7*, 45, doi:10.1186/s13643-018-0707-8.

doi:10.20944/preprints202502.0396.v1

19. Wang, Z.; Chu, Z.; Doan, T.V.; Ni, S.; Yang, M.; Zhang, W. History, Development, and Principles of Large Language Models: An Introductory Survey. *AI and Ethics* **2024**, doi:10.1007/s43681-024-00583-7.
20. Li, M.; Sun, J.; Tan, X. Evaluating the Effectiveness of Large Language Models in Abstract Screening: A Comparative Analysis. *Syst Rev* **2024**, *13*, 219, doi:10.1186/s13643-024-02609-x.
21. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *arXiv preprint arXiv:2307.06435* **2023**.
22. Khraisha, Q.; Put, S.; Kappenberg, J.; Warraitch, A.; Hadfield, K. Can Large Language Models Replace Humans in the Systematic Review Process? Evaluating GPT-4's Efficacy in Screening and Extracting Data from Peer-Reviewed and Grey Literature in Multiple Languages. *arXiv preprint arXiv:2310.17526* **2023**.
23. Giray, L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Ann Biomed Eng* **2023**, *51*, 2629–2633, doi:10.1007/s10439-023-03272-4.
24. Meskó, B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J Med Internet Res* **2023**, *25*, e50638, doi:10.2196/50638.
25. Zaghir, J.; Naguib, M.; Bjelogrlic, M.; Névéol, A.; Tannier, X.; Lovis, C. Prompt Engineering Paradigms for Medical Applications: Scoping Review. *J Med Internet Res* **2024**, *26*, e60501, doi:10.2196/60501.
26. Wahidur, R.S.M.; Tashdeed, I.; Kaur, M.; Lee, H.-N. Enhancing Zero-Shot Crypto Sentiment with Fine-Tuned Language Model and Prompt Engineering. *IEEE Access* **2024**.
27. Ashwathy, J.S.; SR, N.; Pyati, T. The Progression of ChatGPT: An Evolutionary Study from GPT-1 to GPT-4. *Journal of Innovations in Data Science and Big Data Management* **2024**, 38–44.
28. Bharathi Mohan, G.; Prasanna Kumar, R.; Parathasarathy, S.; Aravind, S.; Hanish, K.B.; Pavithria, G. Text Summarization for Big Data Analytics: A Comprehensive Review of GPT 2 and BERT Approaches. In; 2023; pp. 247–264.
29. Kalyan, K.S. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *Natural Language Processing Journal* **2024**, *6*, 100048, doi:10.1016/j.nlp.2023.100048.
30. Katrak, M. The Role of Language Prediction Models in Contractual Interpretation: The Challenges and Future Prospects of GPT-3. *Legal Analytics* **2022**, 47–62.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv Neural Inf Process Syst* **2017**, *30*.
32. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; 2020; pp. 38–45.
33. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A Survey of Transformers. *AI Open* **2022**.
34. Liu, J.; Chu, X.; Wang, Y.; Wang, M. Deep Text Retrieval Models Based on DNN, CNN, RNN and Transformer: A Review. In Proceedings of the 2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS); IEEE, 2022; pp. 391–400.
35. Shiri, F.M.; Perumal, T.; Mustapha, N.; Mohamed, R. A Comprehensive Overview and Comparative Analysis on Deep Learning Models. *CNN, RNN, LSTM, GRU* **2023**.
36. Gue, C.C.Y.; Rahim, N.D.A.; Rojas-Carabali, W.; Agrawal, R.; RK, P.; Abisheganaden, J.; Yip, W.F. Evaluating the OpenAI's GPT-3.5 Turbo's Performance in Extracting Information from Scientific Articles on Diabetic Retinopathy. *Syst Rev* **2024**, *13*, 135, doi:10.1186/s13643-024-02523-2.
37. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of ChatGPT-Related Research and Perspective towards the Future of Large Language Models. *Meta-Radiology* **2023**, *1*, 100017, doi:10.1016/j.metrad.2023.100017.
38. Zhang, C.; Zhang, C.; Zheng, S.; Qiao, Y.; Li, C.; Zhang, M.; Dam, S.K.; Thwal, C.M.; Tun, Y.L.; Huy, L.L. A Complete Survey on Generative Ai (Aigc): Is Chatgpt from Gpt-4 to Gpt-5 All You Need? *arXiv preprint arXiv:2303.11717* **2023**.
39. Tao, K.; Osman, Z.A.; Tzou, P.L.; Rhee, S.-Y.; Ahluwalia, V.; Shafer, R.W. GPT-4 Performance on Querying Scientific Publications: Reproducibility, Accuracy, and Impact of an Instruction Sheet. *BMC Med Res Methodol* **2024**, *24*, 139.

40. Baktash, J.A.; Dawodi, M. Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing. *arXiv preprint arXiv:2305.03195* **2023**.
41. Sindhu, B.; Prathamesh, R.P.; Sameera, M.B.; KumaraSwamy, S. The Evolution of Large Language Model: Models, Applications and Challenges. In Proceedings of the 2024 International Conference on Current Trends in Advanced Computing (ICCTAC); IEEE, 2024; pp. 1–8.
42. Irshad, M. Revolutionizing Healthcare Delivery: Evaluating the Impact of Google's Gemini AI as a Virtual Doctor in Medical Services. J Artif Intell. *Mach Learn & Data Sci 2024 2*, 1618–1625.
43. Annepaka, Y.; Pakray, P. Large Language Models: A Survey of Their Development, Capabilities, and Applications. *Knowl Inf Syst* **2024**, doi:10.1007/s10115-024-02310-4.
44. Xu, J.; Li, Z.; Chen, W.; Wang, Q.; Gao, X.; Cai, Q.; Ling, Z. On-Device Language Models: A Comprehensive Review. *arXiv preprint arXiv:2409.00088* **2024**.
45. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *arXiv preprint arXiv:2307.06435* **2023**.
46. Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; Huang, K. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *IEEE Communications Surveys & Tutorials* **2025**.
47. Patel, D.; Raut, G.; Cheetirala, S.N.; Nadkarni, G.N.; Freeman, R.; Glicksberg, B.S.; Klang, E.; Timsina, P. Cloud Platforms for Developing Generative AI Solutions: A Scoping Review of Tools and Services. *arXiv preprint arXiv:2412.06044* **2024**.
48. Ofoeda, J.; Boateng, R.; Effah, J. Application Programming Interface (API) Research. *International Journal of Enterprise Information Systems* **2019**, *15*, 76–95, doi:10.4018/IJEIS.2019070105.
49. Hadi, M.U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; Mirjalili, S. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *Authorea Preprints* **2023**.
50. Villalobos, P.; Ho, A.; Sevilla, J.; Besiroglu, T.; Heim, L.; Hobbhahn, M. Will We Run out of Data? Limits of LLM Scaling Based on Human-Generated Data. *arXiv preprint arXiv:2211.04325* **2024**, 13–29.
51. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* **2023**.
52. Wang, J.; Yang, Z.; Yao, Z.; Yu, H. Jmlr: Joint Medical Llm and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability. *arXiv preprint arXiv:2402.17887* **2024**.
53. Zhang, Y.; Mao, S.; Ge, T.; Wang, X.; de Wynter, A.; Xia, Y.; Wu, W.; Song, T.; Lan, M.; Wei, F. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. *arXiv preprint arXiv:2404.01230* **2024**.
54. Liu, Y.; He, H.; Han, T.; Zhang, X.; Liu, M.; Tian, J.; Zhang, Y.; Wang, J.; Gao, X.; Zhong, T. Understanding Llms: A Comprehensive Overview from Training to Inference. *arXiv preprint arXiv:2401.02038* **2024**.
55. Campos, D.G.; Fütterer, T.; Gfrörer, T.; Lavelle-Hill, R.; Murayama, K.; König, L.; Hecht, M.; Zitzmann, S.; Scherer, R. Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research. *Educ Psychol Rev* **2024**, *36*, 19, doi:10.1007/s10648-024-09862-5.
56. Drury, A.; Pape, E.; Dowling, M.; Miguel, S.; Fernández-Ortega, P.; Papadopoulou, C.; Kotronoulas, G. How to Write a Comprehensive and Informative Research Abstract. *Semin Oncol Nurs* **2023**, *39*, 151395, doi:10.1016/j.soncn.2023.151395.
57. Liang, X.; Wang, H.; Wang, Y.; Song, S.; Yang, J.; Niu, S.; Hu, J.; Liu, D.; Yao, S.; Xiong, F. Controllable Text Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2408.12599* **2024**.
58. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans Inf Syst* **2025**, *43*, 1–55, doi:10.1145/3703155.
59. Loya, M.; Sinha, D.A.; Futrell, R. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variation and Hyperparameters. *arXiv preprint arXiv:2312.17476* **2023**.
60. Errica, F.; Siracusano, G.; Sanvito, D.; Bifulco, R. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. *arXiv preprint arXiv:2406.12334* **2024**.

61. Sclar, M.; Choi, Y.; Tsvetkov, Y.; Suhr, A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting. *arXiv preprint arXiv:2310.11324* **2023**.
62. Shen, J.; Tenenholtz, N.; Hall, J.B.; Alvarez-Melis, D.; Fusi, N. Tag-LLM: Repurposing General-Purpose LLMs for Specialized Domains. *arXiv preprint arXiv:2402.05140* **2024**.
63. Felin, T.; Holweg, M. Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science* **2024**, *9*, 346–371, doi:10.1287/stsc.2024.0189.
64. Crowther, M.A.; Cook, D.J. Trials and Tribulations of Systematic Reviews and Meta-Analyses. *ASH Education Program Book* **2007**, *2007*, 493–497.
65. Guizzardi, S.; Colangelo, M.T.; Mirandola, P.; Galli, C. Modeling New Trends in Bone Regeneration, Using the BERTopic Approach. *Regenerative Med* **2023**, *18*, 719–734.
66. Colangelo, M.T.; Meleti, M.; Guizzardi, S.; Galli, C. A Macroscopic Exploration of the Ideoscape on Exosomes for Bone Regeneration. *Osteology* **2024**, *4*, 159–178.
67. Galli, C.; Cusano, C.; Meleti, M.; Donos, N.; Calciolari, E. Topic Modeling for Faster Literature Screening Using Transformer-Based Embeddings. In Proceedings of the Metrics; MDPI, 2024; Vol. 1, p. 2.
68. Mateen, F.J.; Oh, J.; Tergas, A.I.; Bhayani, N.H.; Kamdar, B.B. Titles versus Titles and Abstracts for Initial Screening of Articles for Systematic Reviews. *Clin Epidemiol* **2013**, 89–95.
69. Saloojee, H.; Pettifor, J.M. Maximizing Access and Minimizing Barriers to Research in Low- and Middle-Income Countries: Open Access and Health Equity. *Calcif Tissue Int* **2023**, *114*, 83–85, doi:10.1007/s00223-023-01151-7.
70. Herbst, E.; Kopf, S. Writing an Abstract. *Arthroskopie* **2024**, *37*, 258–261, doi:10.1007/s00142-024-00688-5.
71. Galli, C.; Colangelo, M.T.; Guizzardi, S. Linguistic Changes in the Transition from Summaries to Abstracts: The Case of the Journal of Experimental Medicine. *Learned Publishing* **2022**, *35*, 271–284, doi:10.1002/leap.1427.
72. Moher, D.; Schulz, K.F.; Altman, D.G. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials. *The lancet* **2001**, *357*, 1191–1194.
73. Hermont, A.P.; Cruz, P.V.; Occhi-Alexandre, I.G.P.; Bendo, C.B.; Auad, S.M.; Pordeus, I.A.; Martins, C.C. The Importance of Full Text Screening When Judging Eligibility Criteria in a Systematic Review. *Arquivos em Odontologia* **2022**, *58*, 160–165, doi:10.35699/2178-1990.2022.37521.
74. Jacso, P. Open Access to Scholarly Full-text Documents. *Online Information Review* **2006**, *30*, 587–594.
75. Singh, A.; Singh, M.; Singh, A.K.; Singh, D.; Singh, P.; Sharma, A. "Free Full Text Articles": Where to Search for Them? *Int J Trichology* **2011**, *3*, 75–79.
76. Lewis, C.L. The Open Access Citation Advantage: Does It Exist and What Does It Mean for Libraries? *Information Technology and Libraries* **2018**, *37*, 50–65.
77. Ye, A.; Maiti, A.; Schmidt, M.; Pedersen, S.J. A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis. *Future Internet* **2024**, *16*, 167.
78. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C.A.F.; Nielsen, H. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics* **2000**, *16*, 412–424, doi:10.1093/bioinformatics/16.5.412.
79. Bramer, W.M.; Giustini, D.; Kramer, B.M.R. Comparing the Coverage, Recall, and Precision of Searches for 120 Systematic Reviews in Embase, MEDLINE, and Google Scholar: A Prospective Study. *Syst Rev* **2016**, *5*, 39, doi:10.1186/s13643-016-0215-7.
80. Streiner, D.L.; Norman, G.R. "Precision" and "Accuracy": Two Terms That Are Neither. *J Clin Epidemiol* **2006**, *59*, 327–330, doi:10.1016/j.jclinepi.2005.09.005.
81. Straube, S.; Heinz, J.; Landsvogt, P.; Friede, T. Recall, Precision, and Coverage of Literature Searches in Systematic Reviews in Occupational Medicine: An Overview of Cochrane Reviews Recall, Precision Und Coverage von Literatursuchen in Systematischen Reviews Aus Dem Bereich Arbeitsmedizin: Ein Überblick Über Cochrane Reviews. *GMS Medizinische Informatik, Biometrie und Epidemiologie* **2021**, *17*.
82. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: An Overview. *arXiv preprint arXiv:2008.05756* **2020**.

83. Beurer-Kellner, L.; Fischer, M.; Vechev, M. Prompting Is Programming: A Query Language for Large Language Models. *Proceedings of the ACM on Programming Languages* **2023**, *7*, 1946–1969.
84. Chen, B.; Zhang, Z.; Langrené, N.; Zhu, S. Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review. *arXiv preprint arXiv:2310.14735* **2023**.
85. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv* **2023**, *55*, 1–35.
86. Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C.P.; Wang, X.-Z.; Wu, Q.M.J. A Review of Generalized Zero-Shot Learning Methods. *IEEE Trans Pattern Anal Mach Intell* **2022**, *45*, 4051–4070.
87. Li, Y. A Practical Survey on Zero-Shot Prompt Design for in-Context Learning. *arXiv preprint arXiv:2309.13205* **2023**.
88. Dang, H.; Mecke, L.; Lehmann, F.; Goller, S.; Buschek, D. How to Prompt? Opportunities and Challenges of Zero-and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. *arXiv preprint arXiv:2209.01390* **2022**.
89. Qi, B.; Zhang, K.; Li, H.; Tian, K.; Zeng, S.; Chen, Z.-R.; Zhou, B. Large Language Models Are Zero Shot Hypothesis Proposers. *arXiv preprint arXiv:2311.05965* **2023**.
90. Wang, X.; Yin, X.; Zhang, Y.; Zhang, Y. Related Work on Few-Shot Method: A Review. **2024**.
91. Dang, H.; Mecke, L.; Lehmann, F.; Goller, S.; Buschek, D. How to Prompt? Opportunities and Challenges of Zero-and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. *arXiv preprint arXiv:2209.01390* **2022**.
92. 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.
93. Kusano, G.; Akimoto, K.; Takeoka, K. Are Longer Prompts Always Better? Prompt Selection in Large Language Models for Recommendation Systems. *arXiv preprint arXiv:2412.14454* **2024**.
94. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927* **2024**.
95. Heston, T.F.; Khun, C. Prompt Engineering in Medical Education. *International Medical Education* **2023**, *2*, 198–205.
96. Ferdaus, M.M.; Abdelguerfi, M.; Ioup, E.; Niles, K.N.; Pathak, K.; Sloan, S. Towards Trustworthy Ai: A Review of Ethical and Robust Large Language Models. *arXiv preprint arXiv:2407.13934* **2024**.
97. Quttainah, M.; Mishra, V.; Madakam, S.; Lurie, Y.; Mark, S. Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study. *JMIR AI* **2024**, *3*, e51834, doi:10.2196/51834.
98. Matsui, K.; Utsumi, T.; Aoki, Y.; Maruki, T.; Takeshima, M.; Takaesu, Y. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *J Med Internet Res* **2024**, *26*, e52758.
99. Guo, E.; Gupta, M.; Deng, J.; Park, Y.-J.; Paget, M.; Naugler, C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res* **2024**, *26*, e48996.
100. Tran, V.-T.; Gartlehner, G.; Yaacoub, S.; Boutron, I.; Schwingshackl, L.; Stadelmaier, J.; Sommer, I.; Aboulayeh, F.; Afach, S.; Meerpohl, J. Sensitivity, Specificity and Avoidable Workload of Using a Large Language Models for Title and Abstract Screening in Systematic Reviews and Meta-Analyses. *medRxiv* **2023**, 2012–2023.
101. Anisuzzaman, D.M.; Malins, J.G.; Friedman, P.A.; Attia, Z.I. Fine-Tuning Llms for Specialized Use Cases. *Mayo Clinic Proceedings: Digital Health* **2024**.
102. Parthasarathy, V.B.; Zafar, A.; Khan, A.; Shahid, A. The Ultimate Guide to Fine-Tuning Llms from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. *arXiv preprint arXiv:2408.13296* **2024**.
103. Cohen, Y.; Aperstein, Y. A Review of Generative Pretrained Multi-Step Prompting Schemes–and a New Multi-Step Prompting Framework. **2024**.

104. Neimann Rasmussen, L.; Montgomery, P. The Prevalence of and Factors Associated with Inclusion of Non-English Language Studies in Campbell Systematic Reviews: A Survey and Meta-Epidemiological Study. *Syst Rev* **2018**, *7*, 1–12.
105. Zhu, S.; Xu, S.; Sun, H.; Pan, L.; Cui, M.; Du, J.; Jin, R.; Branco, A.; Xiong, D. Multilingual Large Language Models: A Systematic Survey. *arXiv preprint arXiv:2411.11072* **2024**.
106. Xu, Y.; Hu, L.; Zhao, J.; Qiu, Z.; Ye, Y.; Gu, H. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *arXiv preprint arXiv:2404.00929* **2024**.
107. Yuan, F.; Yuan, S.; Wu, Z.; Li, L. How Multilingual Is Multilingual LLM? *arXiv preprint arXiv:2311.09071* **2023**.
108. Thellmann, K.; Stadler, B.; Fromm, M.; Buschhoff, J.S.; Jude, A.; Barth, F.; Leveling, J.; Flores-Herr, N.; Köhler, J.; Jäkel, R. Towards Multilingual LLM Evaluation for European Languages. *arXiv preprint arXiv:2410.08928* **2024**.
109. Huang, K.; Mo, F.; Li, H.; Li, Y.; Zhang, Y.; Yi, W.; Mao, Y.; Liu, J.; Xu, Y.; Xu, J. A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers. *arXiv preprint arXiv:2405.10936* **2024**.
110. Huang, K.; Mo, F.; Li, H.; Li, Y.; Zhang, Y.; Yi, W.; Mao, Y.; Liu, J.; Xu, Y.; Xu, J. A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers. *arXiv preprint arXiv:2405.10936* **2024**.
111. Laakso, A.; Kemell, K.-K.; Nurminen, J.K. Ethical Issues in Large Language Models: A Systematic Literature Review. **2024**.
112. Zhang, Z.; Yao, Y.; Zhang, A.; Tang, X.; Ma, X.; He, Z.; Wang, Y.; Gerstein, M.; Wang, R.; Liu, G. Igniting Language Intelligence: The Hitchhiker's Guide From Chain-of-Thought Reasoning to Language Agents. *arXiv preprint arXiv:2311.11797* **2023**.
113. Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G.L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. Factuality Challenges in the Era of Large Language Models and Opportunities for Fact-Checking. *Nat Mach Intell* **2024**, *6*, 852–863, doi:10.1038/s42256-024-00881-z.
114. Wang, H.; Fu, W.; Tang, Y.; Chen, Z.; Huang, Y.; Piao, J.; Gao, C.; Xu, F.; Jiang, T.; Li, Y. A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy. *arXiv preprint arXiv:2501.09431* **2025**.
115. Lin, Z.; Guan, S.; Zhang, W.; Zhang, H.; Li, Y.; Zhang, H. Towards Trustworthy LLMs: A Review on Debiasing and Dehallucinating in Large Language Models. *Artif Intell Rev* **2024**, *57*, 243.
116. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S.S. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv preprint arXiv:2411.10915* **2024**.
117. Ranjan, R.; Gupta, S.; Singh, S.N. A Comprehensive Survey of Bias in Llms: Current Landscape and Future Directions. *arXiv preprint arXiv:2409.16430* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.