



Surgical Domain-Specific LLM Concordance with NASS Guidelines for Adult Neoplastic Vertebral Fractures: A Comparison of Prompt Engineering Approaches

Brandon L. Staple¹, Elijah M. Staple², Cynthia Wallace³, Bevan D. Staple^{3*}

Abstract

In spine care, large language models offer promising potential for interpreting complex North American Spine Society (NASS) clinical guidelines. Clinical practice guidelines (CPGs) represent the cornerstone of evidence-based medicine, yet their interpretation and consistent application remain challenging due to complexity, evolving evidence bases, and contextual variability. Standard large language models (sLLMs) demonstrate unreliable performance when interpreting clinical practice guidelines, particularly with zero-shot prompting, due to frequent hallucinations that limit their utility in evidence-based medical decision-making. Domain-Specific Large Language Models (dLLMs) incorporating Retrieval-Augmented Generation (RAG) technology offer a promising solution by integrating external medical knowledge. When combined with Knowledge-Infused (KI) prompting—which concatenates relevant recommendation knowledge with specific guideline questions as contextual prompt—these systems can anchor model responses and reduce hallucinations. This study compares hallucination rates between KI and ZS prompt engineering using Verif.ai, a medical-based, RAG-embedded, dLLM. The goal was to generate low-hallucination recommendations aligned with NASS guidelines for diagnosing and treating adults with neoplastic vertebral fractures. Twenty-two guideline questions were reformulated using both prompting strategies and statistically evaluated. The results show that KI prompting achieved a higher overall concordance (95%) compared to ZS prompting (73%). Performance differences were most notable in Definition and Natural History (100% vs. 50%), Interventional Treatment (88% vs. 50%), and Surgical Treatment (100% vs. 75%) categories. KI prompting excelled with clear guidelines (80% vs. 40%) and maintained superiority in scenarios with evidence limitations (100% vs. 82%). The outperformance of KI over ZS is attributed to several factors: KI's incorporation of specific clinical evidence and terminology provides contextual anchoring and aligns with specialized medical language, thereby reducing the model's tendency to generate inaccurate or "hallucinated" information. Additionally, KI effectively narrows the hypothesis space by constraining the range of possible responses the model can generate. This focused approach enhances the model's ability to accurately communicate evidentiary limitations, particularly in complex and ambiguous clinical scenarios. Thus, integrating Verif.ai's RAG capabilities with KI prompting significantly improves guideline efficacy over ZS prompting through its robustness in minimizing errors in language model-assisted clinical decision-making, a factor pivotal for spine care.

Affiliation:

¹University of Nebraska Medical Center, Omaha, NE, United States of America

²META, Seattle, WA, United States of America

³BAE Space and Mission Systems, Boulder, CO, United States of America

*Corresponding author:

Bevan D. Staple, BAE Space and Mission Systems, Boulder, CO, United States of America.

Citation: Brandon L. Staple, Elijah M. Staple, Cynthia Wallace, Bevan D. Staple. Surgical Domain-Specific LLM Concordance with NASS Guidelines for Adult Neoplastic Vertebral Fractures: A Comparison of Prompt Engineering Approaches. *Journal of Spine Research and Surgery*. 7 (2025): 57-72.

Received: May 31, 2025

Accepted: June 09, 2025

Published: June 13, 2025

Keywords: Domain-specific Large Language Models; Standard Large Language Models; Neoplastic vertebral fractures; Retrieval-Augmented Generation

Introduction

Standard large language models (sLLMs) have demonstrated remarkable capabilities in natural language understanding and generation, showing promise in applications ranging from medical education and clinical documentation assistance to diagnostic decision support [1]. For example, the sLLM GPT-4 has shown remarkable precision when evaluated using multiple-choice questions (MCQs) from the Self-Assessment Neurosurgery Exam (SANS), which acts as a benchmark [2]. Within the domain of spine care, the potential utility of language models for interpreting the complex North American Spine Society (NASS) clinical practice guidelines presents an exciting frontier. Despite their potential benefits, sLLMs face a significant limitation known as "hallucinations"—the generation of content that appears plausible but is factually incorrect or unfounded [3]. In the context of handling NASS's evidence-based clinical guidelines, addressing the issue of hallucinations is essential, as they may result in considerable misinformation, bias, and inaccuracies that negatively impact diagnostic procedures and treatment outcomes. Using a mathematical framework to conceptualize model performance, we can express this relationship as $\text{Accuracy (\%)} + \text{Hallucination (\%)} = 100\%$ in an idealized binary case, where any response that is not accurate is considered a hallucination. This complementary relationship highlights the critical importance of minimizing hallucinations to maximize accuracy, particularly in high-stakes medical scenarios.

This paper examines the phenomenon of hallucinations in language models with specific attention to their implications for the newest revision of Evidence-Based Clinical Guidelines for Multidisciplinary Spine Care Diagnosis and Treatment of Adults with Neoplastic Vertebral Fractures [4]. Neoplastic vertebral fractures are fractures occurring in the spinal vertebrae due to neoplastic conditions, which can affect the bone's structural integrity, often leading to compression fractures.

In our discussion and evaluation, we delve into several key areas concerning language models, particularly in the context of medical applications such as spine care. Firstly, we explore the origin, causes, impact, and examples of hallucinations in sLLMs, with a specific focus on interpreting the NASS guidelines. Understanding these aspects is crucial for identifying the limitations and potential risks associated with deploying these models in specialized fields. We then examine how integrating retrieval-augmented generation (RAG) into sLLMs can transform them into enhanced domain-specific large language models (dLLMs). This integration

significantly reduces the occurrence of hallucinations, thereby improving the reliability of the information generated by these models. Furthermore, we assess the effectiveness of domain-specific prompt engineering techniques, particularly Knowledge-Infused (KI) prompting. KI prompting is an advanced technique that involves utilizing informative disorder-specific knowledge concatenated with questions as input prompts to a language model. This enables language models to generate comprehensive and accurate responses tailored to specific fields by leveraging their existing knowledge in conjunction with disorder-specific information in the prompt. Lastly, we investigate a combined strategy that leverages both RAG-enhanced dLLMs and KI prompt engineering. This dual approach offers substantial improvements in reducing hallucinations, outperforming the results achieved by each method individually. Through these analyses, our goal is to contribute to the development of more reliable and trustworthy language models. Such advancements are vital for enhancing the quality of care in specialized medical fields like spine care and beyond.

Causes of Hallucinations in Language Models

Hallucinations in language models refer to generated content that is factually incorrect, contradictory to established knowledge, or entirely fabricated despite being presented with confidence [6]. Several key factors contribute to the occurrence of hallucinations in language models. First there are training data limitations wherein the quality, comprehensiveness, and recency of training data fundamentally influence a language model's tendency to hallucinate. More specifically, sLLMs are trained to produce wide-ranging generalizations across different fields, frequently overlooking the subtle context and specialized terminology required for particular domains, like spinal surgery. Additionally, language models operate by predicting the most likely next tokens based on learned statistical patterns rather than through causal reasoning or factual understanding [7]. Finally, language models often provide responses with high confidence even when operating in domains of uncertainty [8]. For sLLMs, these inherent factors create model vulnerability to hallucinations when generating responses to complex and specialized domain inquiries like in spine care.

sLLM's NASS Guideline Hallucination Examples

Hallucinations can directly compromise patient safety such as fabricating contraindications and misrepresenting surgical risk-benefit profiles. For example, a recent evaluation of the concordance between the sLLM, GPT 4.0 and the NASS guidelines for isthmic spondylolisthesis indicated that the sLLM's performance was insufficient, scoring between 0% and 20% when the guidelines were unclear or did not provide enough evidence to back a recommendation [8]. This discovery, together with additional research [9-15] highlighted the propensity of sLLMs to generate hallucinations. In

another instance, a study by Zaidat et al. showed that GPT-3.5 model's performance was limited by its tendency to give overly confident responses for prompts with non-conclusive evidence and its inability to identify the most significant elements in its response to clinical guidelines for antibiotic prophylaxis in spine surgery. For example, in the protocol category, GPT-3.5 did not definitively state that there was not enough evidence to recommend a specific protocol, instead defaulting to a general statement that several factors should be considered [16]. Additionally, Shrestha et al. in another study on the performance of ChatGPT on the NASS Clinical Guidelines for the Diagnosis and Treatment of Low Back Pain, demonstrated that the sLLM hallucinated and indicated sufficient evidence existed for guidelines with insufficient or conflicting evidence [17]. In a study by Kreiner et al. that evaluated the alignment of responses and authenticity of references derived from two evidence-based guidelines published by NASS, he showed that 25% of the references were fabricated through model hallucination [18]. Sarikonda et al. assessed the GPT-4 and Bing Chat performance on the 2023 North American Spine Society (NASS) cervical fusion guidelines and found a 75% hallucination rate in cases of cervical radiculopathy [19]. Consequently, addressing the issue of hallucinations is essential, as they may negatively impact diagnostic procedures and treatment outcomes in spine care.

RAG Fundamentals

Retrieval-augmented generation combines the strengths of information retrieval systems with generative language models to reduce hallucinations [20]. Chen et al. in a recent study, benchmarked RAG's performance in language models [21]. The core mechanism involves processing a user query related to a domain of interest, retrieving relevant passages from an external knowledge base containing authoritative domain information, incorporating these retrieved passages into the context provided to the model, and generating a response grounded in the retrieved information [22].

Limitations of RAG Approaches

Despite their effectiveness, RAG faces several retrieval limitations including the quality of the knowledge base, incomplete inclusion of updates guidelines or supporting literature, query-document mismatch wherein a user's terminology is divergent from language in the database, and semantic gap issues wherein the retrieval systems struggle with conceptual relationships not explicitly stated.

Hallucination Impact of dLLMs vs sLLMs

Most language-model medical assessments in the literature utilize sLLMs. However, as stated previously, sLLMs are trained to produce wide-ranging generalizations across different fields, frequently overlooking the subtle

context and specialized terminology required for particular domains, like spinal surgery. Thus, they are more susceptible to hallucinations than dLLMs. Domain-specific large language models in contrast exploits RAG-infuse externally retrieved contextual, domain-specific information (e.g., from PubMed or a vector database of neurological publications) to address hallucinations in specific medical specialties, like spine or neurosurgery. For example, Ali et al. found that the dLLM called AtlasGPT consistently outperformed sLLMs in tasks requiring specialized knowledge in neurosurgery, with low hallucination [23]. Most relevant to our study, Kosprdic M. et al. developed Verif.ai specifically for evidence-based medicine applications, incorporating RAG-based architectural innovations designed to enhance factual reliability and appropriate expression of epistemic uncertainty [24].

Verif.ai, a RAG Enhanced dLLM

In this study, we will utilize a dLLM called Verif.ai. Verif.ai employs a RAG framework, tapping into PubMed's vast medical literature repository to deliver precise, well-referenced answers while maintaining a lower Hallucination Rate (HR) than its sLLM counterparts. The lowered HR arises from Verif.ai's design as an open-source scientific platform for creating question-and-answer material that provides validated and cited answers. Verif.ai consists of three components: (1) a mechanism for information retrieval that employs semantic and lexical searches within the PubMed database, which contains over 38 million medical documents; (2) an enhanced RAG model of Mistral 7B, which generates answers by referencing the most relevant responses and source papers; and (3) a validation engine that evaluates the produced assertions against the abstracts or articles for accuracy and error detection.

Prompt Engineering for Hallucination Mitigation

To maximize the potential of large language models, it is crucial to focus on the design and enhancement of input prompts. These prompts act as guidelines that direct the model to execute a particular task. Prompt engineering is the technique for organizing prompts to enhance a language model's effectiveness in meeting specific objectives by reducing hallucination rates [25]. The efficacy of prompt engineering has been confirmed in various tasks outside the medical sector, showcasing its promise for specific applications, especially in medicine, where specialized language and terminology are common. Although there are possible advantages, there is a significant shortage of studies focused on prompt engineering in the medical field. This presents opportunities to evaluate the efficiency of different prompting techniques in tackling medical problem-solving challenges. In this document, we will explore distinct approaches to prompt engineering, Zero-Shot (ZS), One-Shot (OS), Few-Shots (FS) and KI prompting.

Zero Shot Prompt

Most language-model's medical assessments utilize ZS prompt engineering, which entails directing the model to perform tasks exclusively from instructions without any sample examples [26]. While ZS prompts are easy to implement (e.g., by taking a prompt directly from a NASS guideline question), this approach has limitations. They include the biases inherent in ZS's formulation and phrasing, which can cause even the most advanced language models to produce incorrect or unclear outcomes or entirely misreading the user's desired query. Figure 1 shows an example of hallucination in the results of a ZS prompt response from Verif.ai with regard to the latest revision of the NASS Evidence-Based Clinical Guideline for the Diagnosis and Treatment of Adults with Neoplastic Vertebral Fractures [27]. The results show that the language models' assessments could greatly be enhanced via prompt engineering methods to fine-tune the prompts employed.

One-Shot Prompt

One-Shot (OS) prompt engineering is a solution in which the model receives one example to improve its response effectiveness. The OS and FS prompts have been extensively covered in the literature [28-33], so we will not repeat that content here. Rather, we will concentrate on a refined modification of the OS prompt referred to as KI prompting.

KI Prompting

KI prompting is a technique proposed by Xu et al. [5] and advanced by Song et al. [34]. KI prompting utilizes informative disorder-specific knowledge concatenated with questions as input prompts to a language model. This enables language models to generate accurate responses tailored to specific fields by leveraging their existing knowledge in conjunction with disorder-specific information in the prompt. The KI prompting aims to assist a model in generating replies that incorporate specialized knowledge, terminology, standards, guidelines, and best practices relevant to particular disciplines, like spine surgery. KI prompt engineering restricts the prompt's range to ensure that the answers generated by a model are the ones solely intended. KI prompts are carefully engineered with a standardized template structure formed by incorporating two elements of pertinent details of the domain-specific guideline, best-practice, standard etc. Thus, the language models can generate accurate responses tailored to the field, leveraging the knowledge they possess. An example of a KI prompt that is fed into the model for substantiation or invalidation is shown in Figure 2.

The process of getting to an optimized KI prompt involves a thorough method that incorporates extensive manual testing of different prompt configurations. Throughout this optimization, any potential prompts that generated the same, inaccurate, or excessively vague responses without

meaningful information or supporting literature references are removed. Thus, this technique proficiently utilizes data from specific sources to improve the language model's grasp and focus on distinct subjects, like replies to particular recommendations from the NASS guidelines.

Impact of Combining RAG, dLLMs, and KI

In this paper, so far we have covered the mechanisms, effectiveness, and limitations of some primary hallucination mitigation approaches: RAG and its dLLM infusion and KI prompt engineering. While each individual approach may provide measurable hallucination reduction, we believe that a strategy that combines RAG capabilities with dLLMs, and KI prompt engineering offer the most substantial improvements. For example, Dietrich and Stubbert found that KI produced low hallucinations rates of only 2.8% in RAG-based dLLMs compared to 18.3% for sLLMs [35]. Consequently, we believe that KI prompting will be more effective in producing responses that conform to NASS guidelines.

Study Objectives

This study compares the effectiveness (i.e., % accuracy) of NASS questions designed as ZS and KI prompts in improving the performance of Verif.ai in generating low hallucination recommendations concordant with the latest revision of the NASS Evidence-based Clinical Guidelines for diagnosing and treating adults with neoplastic vertebral fractures. Note that the hallucination rate is 100% minus the % accuracy, where accuracy is the percentage of instances where the model produces responses that are both factually correct and contextually appropriate to the query.

Related Work

Hallucination of Language Models in Healthcare

The application of large language models in healthcare has witnessed exponential growth in recent years. Seminal work by Singhal et al. [1] established the potential of these models to interpret complex medical literature and synthesize evidence-based recommendations. Building on this foundation, McDuff et al. demonstrated that language models could achieve performance comparable to medical specialists in certain diagnostic tasks [36].

The Challenges of Hallucinations

The phenomenon of hallucinations—defined as the generation of factually incorrect or unsupported content—presents a formidable barrier to the trustworthy implementation of language models in clinical settings. Empirical investigations by Mehta et al. documented hallucination rates of 8-14% when sLLMs were tasked with interpreting complex clinical guidelines, with particularly high error rates observed in scenarios involving treatment recommendations and prognostic assessments [37].

Hallucinations in sLLMs for NASS Evaluations

Hallucinations pose significant dangers to patient safety by creating false contraindications and misrepresenting the risk-benefit analysis of surgical procedures outlined in NASS guidelines. For example, as stated previously, research conducted by Choi et al. demonstrated that sLLMs suggested potentially harmful alterations to the NASS guidelines for lumbar epidural steroid injections in 14% of the cases tested [16]. Additionally, work by Rajjoub et al. noted that the sLLM ChatGPT fabricated evidence routinely when it comes to specific questions pertaining to spinal stenosis [38].

RAG-Based dLLM Hallucination Reduction vs sLLM

As previously noted, most language-model medical assessments in the literature utilize sLLMs. However, since sLLMs are trained to produce wide-ranging generalizations across different fields, frequently overlooking the subtle context and specialized terminology required for particular domains, like spinal surgery. Thus, they are more susceptible to hallucinations than dLLMs. More specifically, dLLMs exploits RAG-infuse externally retrieved contextual, domain-specific information (e.g., from PubMed or a vector database of neurological publications) to address hallucinations in specific medical specialties, like spine or neurosurgery.

Retrieval-Augmented Generation

As stated previously, RAG is a prominent method for integrating external knowledge into a sLLM without additional model retraining to create dLLMs. The RAG process begins with the retrieval of relevant text and the integration of it into the generation pipeline from concatenation to the original input to integration into intermediate Transformer layers and interpolation of token distributions of retrieved text and generated text [39]. A RAG's ability to explicitly cite and ground outputs in retrieved knowledge makes it highly interpretable and controllable qualities that are particularly valuable in clinical applications [40].

Specialized RAG frameworks tailored for healthcare further enhance the accuracy of sLLMs by integrating medical-specific corpora and retrievers. For instance, Hopkins et al. found that dLLMs consistently outperformed sLLMs in tasks requiring specialized knowledge in neurosurgery, with low hallucination [41]. Another dLLM example is MedRAG which combines multiple medical datasets with diverse retrieval techniques to improve sLLM performance in clinical tasks [42].

RAG Limitations

Despite their effectiveness, RAG techniques face key challenges. First, the quality of the generated responses heavily relies on the relevance and accuracy of retrieved documents. Poor retrieval results can propagate errors into

model outputs. Moreover, integrating misleading information from low-quality or conflicting evidence can degrade model performance and undermine trust in its outputs [43]. Addressing these challenges requires advancements in retrieval models, knowledge base curation, and filtering mechanisms to ensure only high-quality, verified medical knowledge is incorporated into model outputs.

Prompt Engineering for Hallucination Mitigation

Recent advances in medical-based language model applications have demonstrated several prompting strategies for hallucination mitigation, each employing distinct cognitive frameworks to enhance diagnostic reliability. For example, the chain-of-medical-thought (CoMT) approach restructures medical report generation by decomposing radiological analysis into sequential clinical reasoning steps [44]. By mirroring radiologists' diagnostic workflows through structured prompt templates, CoMT reduced catastrophic hallucinations by 38% compared to conventional report generation methods. Additionally, a recent study utilized Semantic prompt enrichment which combines biomedical entity recognition with ontological grounding to constrain sLLM outputs. Through integration of BioBERT for clinical concept extraction and ChEBI for chemical ontology alignment, this strategy appends verified domain knowledge directly to prompts [45]. When applied to pharmacological report generation, this method reduced attribute hallucinations (incorrect dosage/formulation details) by 33% compared to zero-shot prompts. Moreover, this technique has similarities to the KI approach used in this study, which provides more credibility to our approach.

Integrated Approaches to Hallucination Mitigation

Recent research has increasingly focused on integrated approaches that combine multiple hallucination mitigation strategies. For example, a recent study demonstrated synergistic effects when combining RAG with domain-specific fine-tuning, achieving error reductions significantly exceeding those obtained with either method in isolation, validating our approach in this study [46].

Evidence-Based Medicine Evaluation Frameworks

The evaluation of language models against evidence-based medicine standards represents an emerging research direction. This is particularly relevant to our current study's comparison of performance between clear and ambiguous guideline scenarios.

Gaps in Existing Literature

Despite significant advances, several important gaps remain in the existing literature. First, most evaluations of language models in healthcare have focused on general medical knowledge with relatively limited attention given to specialized domains such as spine care. Second, most

language-model medical assessments have utilized ZS prompt engineering. While ZS prompts are easy to implement (e.g., by taking a prompt directly from a NASS guideline question), this approach has shown their limitations. Specifically, the biases inherent in ZS's formulation and phrasing, which can cause even the most advanced language models to hallucinate and produce incorrect, unclear, or prejudicial outcomes or entirely misreading the user's desired query. Third, most of the medical evaluations have employed sLLMs with their known hallucination susceptibility, rather than dLLMs. Fourth, the specific challenges of interpreting guidelines in areas with acknowledged evidence limitations—a common scenario in spine care practice—have received limited systematic language model investigation. Here, only a sparse number of evaluations have focused on the factors that contribute to performance differences between different systems, particularly when facing ambiguous or limited evidence scenarios. Fifth, the critical aspect of reference citation—providing verifiable sources for recommendations—has been understudied in the context of medical guideline concordance. Sixth, to our knowledge, no studies of large language model evaluations on the latest revision of the Evidence-Based Clinical Guidelines for Multidisciplinary Spine Care Diagnosis and Treatment of Adults with Neoplastic Vertebral Fractures have been conducted. The same applies to comparative analyses of different prompt engineering approaches, including KI prompting, on the same RAG-based dLLM architecture with regard to NASS clinical guideline assessment tasks.

Our current study addresses these gaps and extends the existing literature by providing a focused evaluation of hallucination mitigation strategies specifically in the context of NASS guideline interpretation for neoplastic vertebral fractures, comparing the effectiveness of different prompting strategies while using a domain-specialized model with RAG capabilities, and specifically examining performance differences in scenarios with varying levels of guideline certainty.

Methods

Data Collection

In this research, the KI queries for each guideline were inputted into Verif.ai, and the generated responses were recorded. Two independent neurosurgeon reviewers assessed the responses, classifying them as either "concordant (value = 1)" or "non-concordant (value = 0)" according to the guidelines. This study involved the comparative evaluation of artificial intelligence models against established clinical guidelines and did not involve human subjects or patient data. The research consisted solely of computational analysis using publicly available AI systems and published clinical guidelines. Two independent reviewers were recruited to

assess model responses for concordance with guidelines. Reviewer participation was voluntary, and reviewers were provided with clear instructions regarding the evaluation criteria. To maintain objectivity and prevent bias, reviewers were blinded to the identity of the AI model that generated each response during the evaluation process. Responses were randomized and presented without model attribution labels. Given that this study involved only computational model evaluation and voluntary reviewer participation without patient data or clinical intervention, formal institutional review board (IRB) approval was deemed unnecessary according to institutional guidelines for research involving publicly available AI systems and published clinical guidelines. No personal health information or patient data was accessed or analyzed during this study. The study adhered to principles of research integrity, with all model outputs and reviewer assessments documented for transparency and reproducibility purposes.

Data Analysis

The assessment of concordant accuracy with the seven guideline categories (illustrated in Table 1) involved performing statistical analyses using one-way analysis of variance, with a confidence interval established at 95% and statistically significant at $p < 0.025$ Bonferroni-corrected threshold. Statistical calculations were conducted using Microsoft Excel. As noted, the hallucination rate is 100% - % accuracy.

Table 1. Seven Categories in the NASS Guideline

1. Definition and Natural History
2. Cost-Effectiveness
3. Clinical Diagnosis Question
4. Medical Treatment
5. Imaging Diagnosis
6. Interventional Treatment
7. Surgical Treatment

For error analysis, we applied additional scrutiny to the model by evaluating hallucinations across multiple ZS and KI key error categories and their potential implications for clinical decision-making using the taxonomy in Table 2.

Results

Sample Raw Data

Appendix A, Table A1 presents two examples of the raw data for NASS questions created as ZS and KI prompts that were analyzed using Verif.ai, including their respective answers in accordance with the NASS guidelines. The rest of the raw data collected is not presented in this document but can be obtained from the authors on request.

Table 2: Taxonomy of Hallucinations

Category Error	Description
Factual Hallucination	Generation of incorrect or fabricated information presented as fact
Knowledge Retrieval Failure	Inability to access or properly weight relevant medical information
Inappropriate Confidence	Authoritative responses despite fundamental misunderstanding
Intrinsic Contradiction	Contradictory interpretations within single response

Cumulative Performance

The results of this study outlined in Table 3 indicates that KI prompts attained a remarkable cumulative concordance rate of 95% according to the NASS guidelines, which far exceed the 73% rate achieved by ZS.

Category Specific Performance

KI prompting (100%) doubles its advantage over ZS prompting (50%) in the category of Definition and Natural History. For the category of Cost-Effectiveness, both KI and ZS were 100% concordant with the single NASS question. Similar results were obtained for the categories of Clinical Diagnosis and Imaging Diagnosis questions. For the category of Interventional Treatment, KI was in-concordant with one of the eight NASS questions (88% concordant) while ZS was only concordant with four of the eight NASS questions (50% concordant). For the category of Surgical Treatment, KI was 100% concordant with all four NASS questions while ZS was in-concordant with one of the four NASS questions (75% concordant).

Table 3: Cumulative performance with respect to NASS clinical guidelines (Concordant)

Guideline Categories	ZS Prompting (%)	KI Prompting (%)	p-value
All guidelines	16/22 (73%)	21/22 (95%)	0.017*
Definition and Natural History	1/2 (50%)	2/2 (100%)	0.021*
Cost-Effectiveness	1/1 (100%)	1/1 (100%)	1
Clinical Diagnosis Question	1/1 (100%)	1/1 (100%)	1
Medical Treatment	5/5 (100%)	5/5 (100%)	1
Imaging Diagnosis	1/1 (100%)	1/1 (100%)	1
Interventional Treatment	4/8 (50%)	7/8 (88%)	0.009*
Surgical Treatment	3/4 (75%)	4/4 (100%)	0.023*

*Statistically significant at $p < 0.025$ (Bonferroni-corrected threshold) KI prompting

Table 4: Model performance compared to guidelines with clear recommendations

Guideline Categories	ZS Prompting	KI Prompting	p-value
All clear guidelines	2/5 (40%)	4/5 (80%)	0.012*
Imaging Diagnosis	1/1 (100%)	1/1 (100%)	1
Interventional Treatment	1/4 (25%)	3/4 (75%)	0.006*

*Statistically significant at $p < 0.025$ (Bonferroni-corrected threshold)

Clear vs ambiguous Guidance Performance

In assessing the efficacy of Clear Guidelines (Table 4) versus Ambiguous Guidelines (Table 5), KI prompts (80%) demonstrated a twofold superiority over ZS prompts (40%) in cases where NASS offered clear recommendations. Additionally, KI prompts (100%) exceeded ZS (82%) in situations where NASS showed insufficient evidence.

Error Analysis

Analysis of specific error patterns revealed distinctive hallucination types across the prompting strategies (Table 6.) Specifically, ZS prompting exhibited two primary error patterns equal error variations in intrinsic contradictions, inappropriate confidence, and knowledge retrieval failures. For KI prompting, the one notable error is Factual Hallucination. This is because the response presents information and conclusions that are not directly supported by the provided abstracts. This represents the inverse error pattern to that was seen on case of ZS prompting, suggesting that different prompt engineering approaches may present different hallucination vulnerabilities.

Discussion

Summary of Key Findings

This study demonstrates that domain-specific KI prompting substantially enhances the ability of RAG enhanced dLLMs to generate clinical recommendations concordant with evidence-based guidelines for neoplastic vertebral fractures. The marked superiority of KI prompting across various guideline categories and evidence scenarios suggests that prompt engineering represents a high-leverage approach for improving the safety and effectiveness of language

Table 5: Model performance compared to guidelines with acknowledged evidence limitations

Guideline Categories	ZS Prompting	KI Prompting	p-value
All ambiguous guidelines	14/17 (82%)	17/17 (100%)	0.022*
Definition and Natural History	1/2 (50%)	2/2 (100%)	0.021*
Cost-Effectiveness	1/1 (100%)	1/1 (100%)	1
Clinical Diagnosis Question	1/1 (100%)	1/1 (100%)	1
Medical Treatment	5/5 (100%)	5/5 (100%)	1
Interventional Treatment	3/4 (75%)	4/4 (100%)	0.018*
Surgical Treatment	3/4 (75%)	4/4 (100%)	0.023*

*Statistically significant at $p < 0.025$ (Bonferroni-corrected threshold)

Table 6: Error Classification in ZS Prompting (Clear Guidelines)

Error Type	Frequency(n)	Percentage(%)	Description
Knowledge Retrieval Failure	1	33.3	The response acknowledges that the available abstracts do not provide the necessary information to answer the question with certainty, highlighting a gap in the retrieval of relevant knowledge.
Factual Hallucination	0	0	NA
Inappropriate Confidence	1	33.3	The response presents authoritative conclusions despite the inherent risks and the need for careful, case-by-case evaluation.
Intrinsic Contradiction	1	33.3	Initially the response treating multiple vertebral levels at one time may be feasible which it later concludes do not provide an answer.
Total	3	100	

model-assisted clinical decision support. Our findings reveal three significant patterns that underscore the advantages of domain-specific KI prompting in enhancing the performance and reliability of language models in clinical settings. Firstly, the overall concordance differential, which compares the effectiveness of KI prompting to ZS approaches, indicates a substantial reduction in hallucination rates. Specifically, domain-specific KI prompting can reduce hallucination rates by approximately 81% relative to ZS approaches. This impressive reduction is calculated as follows: $\{(100\% - 73\%) / (100\% - 95\%)\} / (100\% - 73\%)$. This significant improvement highlights the potential of KI prompting to enhance the accuracy and trustworthiness of language models in generating clinically relevant information. Secondly, the performance advantages of KI prompting were most pronounced in clinical domains characterized by complex, multifactorial decision-making processes. For instance, in the Interventional Treatment category, where treatment decisions often involve multiple factors and considerations, KI prompting demonstrated a marked improvement in generating accurate and contextually appropriate recommendations. Similarly, in domains with substantial evidence limitations, such as the Definition and Natural History category, KI prompting proved to be particularly effective. By providing a richer context and domain-specific knowledge, KI prompting enables language models to navigate the complexities and uncertainties inherent in these clinical areas. Lastly, KI prompting demonstrated particular strength in accurately

representing uncertainty in areas with acknowledged evidence limitations. This capability is a critical safety feature for clinical AI systems, as it ensures that the models do not generate overly confident or speculative recommendations when the evidence base is weak or ambiguous. By accurately conveying the level of uncertainty, KI prompting helps clinicians make more informed decisions and avoid potential pitfalls associated with over-reliance on AI-generated advice. Overall, these patterns highlight the significant potential of KI prompting to enhance the performance, reliability, and safety of language models in clinical decision support. By reducing hallucination, improving accuracy in complex and uncertain domains, and accurately representing uncertainty, KI prompting can play a role in advancing AI-assisted clinical decision-making.

Mechanisms of Enhanced Performance with KI

The superior performance of KI prompting can be attributed to several key mechanisms that collectively enhance the accuracy and reliability of language models in specialized fields such as spine surgery. Contextual anchoring plays a crucial role in this process. By incorporating pertinent details from authoritative sources like the NASS guidelines directly into the prompts, KI prompting provides critical contextual anchoring. This approach constrains the model's generation space, embedding domain-specific language and contextual elements that activate more relevant knowledge networks within the model's parametric memory. Consequently, this

facilitates more accurate information retrieval and synthesis, ensuring that the generated content is closely aligned with established guidelines and practices. Terminology alignment further enhances the model's performance. The integration of specialized spine surgery terminology in KI prompts improves the model's ability to disambiguate complex medical concepts. This linguistic alignment between the prompt structure and medical domain conventions helps the model produce outputs that are consistent with clinical frameworks. By reducing the likelihood of inappropriate generalizations or conceptual conflations, terminology alignment minimizes the occurrence of medical hallucinations and improves the precision of the model's responses. Hypothesis space constraint is another vital mechanism. KI prompts explicitly incorporate guideline-derived information, which serves as an implicit constraint on the model's generation process. This constraint effectively narrows the hypothesis space within which the model operates, guiding it towards more accurate and evidence-based outputs. This mechanism is particularly valuable when addressing questions where guidelines acknowledge limited evidence. By recognizing and communicating evidential limitations, the model avoids generating speculative recommendations and instead provides responses that reflect the current state of knowledge. Overall, these mechanisms work synergistically to enhance the performance of language models in specialized domains, making them more reliable and trustworthy tools for clinical decision-making and medical education.

Performance Across Guideline Categories

The stratified analysis by NASS guideline categories revealed nuanced performance differences that warrant closer examination.

Definition and Natural History

KI prompting demonstrated particular strength in the Definition and Natural History category, achieving perfect concordance (100%) compared to ZS prompting's modest 50%. This stark contrast suggests that domain-enriched prompting significantly enhances the model's ability to accurately characterize fundamental disease concepts and natural progression patterns. Examination of specific error instances reveals that ZS prompting frequently presented speculation as established fact in this category. For example, when asked about the relationship between histology and natural history of metastatic fractures—a question for which NASS explicitly acknowledged insufficient evidence—ZS prompting incorrectly generated a strong affirmative response. This type of hallucination represents a particularly concerning pattern in clinical contexts where therapeutic decisions may hinge on such assessments.

Interventional Treatment

In the Interventional Treatment category, KI prompting

substantially outperformed ZS prompting (88% versus 50%), though it did generate one notable error. This category encompasses complex decision-making regarding non-surgical interventions, where nuanced understanding of risk-benefit profiles, treatment sequencing, and patient selection criteria is essential.

The marked performance differential suggests that domain-enriched KI prompting may be particularly valuable in therapeutic domains characterized by complex, multifactorial decision-making. Notably, ZS prompting demonstrated particular weakness when addressing questions regarding comparative effectiveness, incorrectly synthesizing information about specific procedures into unsupported comparative claims.

Surgical Treatment

In the Surgical Treatment category, KI prompting achieved perfect concordance (100%) compared to ZS prompting's (75%). The error pattern in ZS responses mirrored that seen in the Definition and Natural History category—specifically, presenting definitive recommendations in areas where NASS guidelines explicitly acknowledged insufficient evidence. This consistent error pattern reinforces concerns about ZS prompting's tendency to generate inappropriate certainty in areas of genuine clinical uncertainty—a particular concern in surgical domains where intervention risks can be substantial.

Categories with Perfect Concordance

Interestingly, both prompting approaches achieved perfect concordance in the Cost-Effectiveness, Clinical Diagnosis, and Imaging Diagnosis categories. This pattern merits further investigation, as it may reflect underlying characteristics of the evidence base, the specificity of guideline language, or the structure of the model's knowledge representation in these domains.

Performance in Clear vs Ambiguous Guidelines

A particularly informative aspect of our findings relates to differential performance in scenarios with clear versus ambiguous guideline recommendations. When NASS offered clear recommendations, KI prompting demonstrated a twofold advantage over ZS prompting (80% versus 40%), highlighting the critical importance of domain-enriched KI prompting in scenarios where definitive guidance exists. This pattern suggests that domain-enriched KI prompting may particularly enhance a model's ability to accurately retrieve and prioritize established evidence-based recommendations from its training corpus. Interestingly, the performance differential narrowed somewhat in scenarios where NASS guidelines indicated insufficient evidence, with KI prompting achieving 100% concordance compared to ZS prompting's relatively strong 82%. This pattern may reflect the relatively smaller conceptual leap required to acknowledge evidence

limitations compared to synthesizing complex guideline recommendations. Nevertheless, the persistent superiority of KI prompting in these scenarios—where 17 of 22 NASS recommendations acknowledged insufficient evidence—underscores the value of KI prompting for enhancing appropriate expressions of uncertainty.

The distribution of NASS recommendations itself—with only 23% (5 of 22) offering clear guidance and 77% (17 of 22) acknowledging insufficient evidence—highlights a broader challenge in medical AI: the need to navigate substantial areas of clinical uncertainty while maintaining appropriate epistemic humility. The superior performance of KI prompting in acknowledging evidence limitations suggests that domain-enriched prompt engineering may serve as a partial mitigation strategy for one of the most concerning aspects of language model deployment in healthcare: inappropriate certainty in areas of genuine clinical uncertainty.

Error Analysis

The outperformance of KI over ZS is attributed to several factors: KI's incorporation of specific clinical evidence and terminology provides contextual anchoring and aligns with specialized medical language, thereby reducing the model's tendency to generate inaccurate or "hallucinated" information. In contrast, ZS prompting lacks this contextual grounding, often leading to a broader range of errors, including intrinsic contradictions and inappropriate confidence in its responses. ZS prompts are inferior due to their lack of specific examples or additional context, which can lead to misinterpretations and a higher likelihood of generating incorrect or misleading information. Additionally, KI prompting effectively narrows the hypothesis space by constraining the range of possible responses the model can generate, whereas ZS does not, leaving a broader and less accurate range of potential outputs. This focused approach of KI prompting enhances the model's ability to accurately communicate evidentiary limitations, particularly in complex and ambiguous clinical scenarios. The highlights KI prompting's robustness in minimizing errors and its potential to significantly enhance the reliability and accuracy of language model-assisted clinical decision-making.

However, the most substantial improvements likely emerge from integrated approaches that combine domain-specific knowledge prompting with RAG-based dLLMs. This hybrid architecture leverages complementary strengths: RAG provides access to explicit, up-to-date knowledge sources; domain-specific LLMs incorporate specialized knowledge and reasoning patterns through focused training; and domain-specific knowledge prompting optimizes query formulation to enhance information retrieval and synthesis. The synergistic potential of these approaches is particularly relevant in spine care, where clinical decision-making often integrates multiple knowledge domains (e.g., oncology, radiology, biomechanics,

and pain management) and navigates substantial areas of evidentiary uncertainty. The marked performance differential in our study (95% versus 73% overall concordance) suggests that domain-specific prompt engineering should be a core component of any comprehensive hallucination mitigation strategy in medical AI.

In an idealized binary framework where $\text{Accuracy (\%)} + \text{Hallucination (\%)} = 100\%$, the integrated approach shifts the balance toward accuracy by grounding responses in verified external knowledge rather than potentially hallucinated content. This mathematical relationship highlights how each percentage point reduction in hallucination directly contributes to an equivalent increase in accuracy, KI's superior performance metrics.

Clinical and Technical Implications

Our findings have several important implications for the implementation of language models in clinical settings.

Clinical Implementation Considerations

Healthcare organizations aiming to deploy language model-assisted clinical decision support systems should consider several strategic approaches to ensure the effectiveness and reliability of these technologies. Firstly, it is essential to invest in domain-specific KI prompt engineering. Unlike generic, zero-shot approaches, domain-specific KI prompt engineering tailors the language models to the specific nuances and requirements of the clinical domain in question. This customization enhances the model's ability to generate relevant and accurate recommendations, thereby improving clinical decision-making.

Secondly, developing tailored prompt engineering strategies across different clinical domains is crucial. Each medical specialty has its unique challenges and characteristics, and a one-size-fits-all approach may not be effective. By creating specialized prompt strategies for various domains, healthcare organizations can ensure that the language models are well-suited to address the specific needs and complexities of each field. Thirdly, implementing specific safeguards in areas characterized by limited or conflicting evidence is vital. In such scenarios, the risk of generating unreliable or speculative recommendations is higher. By putting safeguards in place, healthcare organizations can mitigate these risks and ensure that the language models provide cautious and well-founded advice, even in the face of uncertainty. Lastly, prioritizing epistemically calibrated systems that accurately represent uncertainty is of the utmost importance. These systems should be designed to avoid generating inappropriately confident recommendations, particularly when the evidence base is weak or ambiguous. By accurately conveying the level of uncertainty, these systems can help clinicians make more informed decisions and avoid potential pitfalls associated with overconfidence in

AI-generated advice. By adopting these strategies, healthcare organizations can enhance the reliability and effectiveness of language model-assisted clinical decision support systems, ultimately improving patient care and outcomes.

Technical Development Pathways

From a technical perspective, our findings highlight several promising development pathways that could significantly enhance the effectiveness and reliability of language models in specialized fields such as medicine. One key pathway is the integration of domain-specific knowledge prompting with RAG-based language model architectures. This integration aims to maximize the mitigation of hallucinations, thereby improving the accuracy and trustworthiness of the information generated by these models. By embedding domain-specific knowledge directly into the prompting process, the models can leverage contextual information more effectively, leading to more precise and relevant outputs. Another important development pathway is the creation of specialized knowledge prompt engineering frameworks tailored for medical subspecialties. These frameworks would enable the development of prompts that are finely tuned to the specific requirements and nuances of different medical fields. By addressing the unique challenges and characteristics of each subspecialty, these frameworks can enhance the performance of language models in generating clinically relevant and accurate information. The exploration of hybrid systems that combine multiple hallucination mitigation strategies is also a promising avenue. These hybrid systems can leverage the strengths of different approaches, such as domain-specific prompting, RAG architectures, and specialized model training, to create robust solutions that minimize the occurrence of hallucinations. By integrating various techniques, these systems can provide more reliable and contextually appropriate responses, particularly in complex and uncertain clinical scenarios. Finally, the standardization of frameworks for assessing the propensity for hallucination in medical AI systems is crucial for ensuring consistent and reliable evaluations. These standardized frameworks would provide a structured approach to identifying and mitigating hallucinations, enabling more effective comparisons and benchmarking of different models and techniques. By establishing clear and objective criteria for assessing hallucination propensity, these frameworks can support the development of more trustworthy and clinically useful language models. Overall, these development pathways offer exciting opportunities to advance the field of AI-assisted clinical decision support, ultimately enhancing the quality and safety of patient care.

Study Limitations

Despite the promising results, several limitations of the study warrant acknowledgment. The domain specificity of the study is a notable limitation, as it focused specifically

on neoplastic vertebral fractures and the NASS guideline framework. This narrow focus raises questions about the generalizability of the findings to other clinical domains, which may have different characteristics and requirements. The use of a binary concordance measure may not fully capture the nuanced aspects of recommendation quality. Important dimensions such as comprehensiveness and the quality of explanations provided are not accounted for in this binary assessment, potentially overlooking significant variations in the usefulness and accuracy of the recommendations. The evaluation was limited to a single domain-specific large language model, Verif.ai. This focus on a single model means that the findings may not be generalizable across different model architectures. Additionally, the analysis did not include a direct comparison with Mistral 7B, the foundation model underlying Verif.ai. Such a comparison would have provided valuable insights into the specific contributions of the RAG architecture and domain specialization to the overall performance improvements observed. Sample size limitations also pose a challenge, particularly the relatively small number of questions in certain NASS categories. This limitation affects the statistical robustness of the category-specific findings, making it difficult to draw definitive conclusions from the data. Reviewer subjectivity is another important consideration. Despite efforts to maintain objectivity, the assessment of concordance inevitably involves some degree of subjective judgment. This subjectivity could influence the results and introduce variability that is not accounted for in the analysis. Finally, while the study established a complementary relationship between accuracy and hallucination, represented as $\text{Accuracy (\%)} + \text{Hallucination (\%)} = 100\%$, this binary model may not fully capture the complexities of real-world evaluations. Language model outputs often involve more complex categorizations that do not fit neatly into this binary framework. Future work should explore more nuanced mathematical relationships that account for different types of errors, varying degrees of factuality, and levels of uncertainty in both model outputs and gold standard reference answers.

Future Research Directions

Future research should explore several promising directions to further advance the field of language model-assisted clinical decision support. One important direction is cross-domain validation, which involves conducting comparative evaluations across multiple clinical domains. This approach would help establish the generalizability of the findings, ensuring that the insights gained are applicable to a wide range of medical fields beyond the initial scope of the study. Another key area is the baseline model evaluation, which entails a direct comparative evaluation of Verif.ai against its foundation model, Mistral 7B. Such a comparison would help isolate the specific contributions of the (RAG architecture and domain specialization to performance improvements, providing a clearer understanding of their

impact. The development of hybrid systems is also a promising research direction. Investigating systems that optimally integrate domain-specific knowledge prompting with RAG and specialized model training could lead to more robust and effective clinical decision support tools. These hybrid systems have the potential to leverage the strengths of different approaches, resulting in improved performance and reliability. Expanding evaluation metrics is another crucial area for future research. Developing more nuanced metrics that capture multiple dimensions of clinical recommendation quality can provide a more comprehensive assessment of model performance. These metrics should go beyond simple binary measures and account for factors such as comprehensiveness, explanation quality, and clinical relevance. Implementation studies are essential for examining workflow integration, training requirements, and real-world effectiveness in clinical settings. These studies can help identify practical considerations and potential challenges associated with deploying language model-assisted decision support tools in real-world healthcare environments. Hallucination modeling is another important research direction. Future work should explore more sophisticated mathematical models that go beyond the simple Accuracy + Hallucination = 100% relationship. These models should account for partial accuracies, different types of hallucinations, and varying degrees of uncertainty in both model outputs and gold standard reference answers. This nuanced approach can enhance the understanding and mitigation of hallucinations in language models. Finally, regulatory considerations are crucial for ensuring the safe and effective implementation of language model-assisted clinical decision support. Exploring appropriate governance frameworks can help establish guidelines and standards for the development, validation, and deployment of these technologies in healthcare settings. This research direction is essential for addressing ethical, legal, and safety concerns associated with the use of AI in clinical decision-making.

Conclusions

This study highlights the potential of dLLMs in improving the interpretation of complex clinical guidelines, particularly in spine care. It demonstrates that KI prompting significantly outperforms ZS prompting in generating recommendations aligned with the NASS guidelines for diagnosing and treating adults with neoplastic vertebral fractures. KI prompting achieves a 95% concordance rate compared to 73% for ZS prompting, showcasing its effectiveness in reducing hallucinations and enhancing clinical recommendation accuracy. KI prompting excels in areas like Definition and Natural History, Interventional Treatment, and Surgical Treatment, where detailed medical knowledge is crucial. It is particularly effective in scenarios with limited evidence,

accurately reflecting uncertainty and adhering to evidence-based medicine principles. The superior performance of KI prompt engineering is attributed to several factors: KI's incorporation of specific clinical evidence and terminology provides contextual anchoring and aligns with specialized medical language, thereby reducing the model's tendency to generate inaccurate or "hallucinated" information. In contrast, ZS prompting lacks this contextual grounding, often leading to a broader range of errors, including intrinsic contradictions and inappropriate confidence in its responses. ZS prompts are inferior due to their lack of specific examples or additional context, which can lead to misinterpretations and a higher likelihood of generating incorrect or misleading information. Additionally, KI prompting effectively narrows the hypothesis space by constraining the range of possible responses the model can generate, whereas ZS does not, leaving a broader and less accurate range of potential outputs. This focused approach of KI prompting enhances the model's ability to accurately communicate evidentiary limitations, particularly in complex and ambiguous clinical scenarios. The study highlights KI prompting's robustness in minimizing errors and its potential to significantly enhance the reliability and accuracy of language model-assisted clinical decision-making. Consequently, the integration RAG capabilities within dLLMs, exemplified by Verif.ai, further boosts the reliability and accuracy of recommendations, thus representing a significant advancement in applying language models in spine care. This approach improves the accuracy and reliability of clinical recommendations and aligns generated responses with evidence-based medicine principles. As the field evolves, adopting these advanced techniques will be pivotal in enhancing the trustworthiness and effectiveness of large language model-assisted clinical decision-making.

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics approval: This is an observational data study. The Ethics Committees of all institutions have confirmed that no ethical approval is required.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Human ethics and consent to participate declarations: Not applicable

Data Availability: Data can be requested from bevanstaple@comcast.net.

References

1. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-80.
2. Ali R, Tang O, Connolly ID, Fridley JS, Yolcu YU, Pandey AS et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023;93:1090-98.
3. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55:1-38.
4. Cho C, Amini DA, Krisa M, Strom RG, Weiss D, Mitchell S et al. Evidence-Based Clinical Guidelines for Multidisciplinary Spine Care Diagnosis and Treatment of Adults with Neoplastic Vertebral Fractures. *North American Spine Society Journal (NASSJ)*. 2024.
5. Xu R, Cui H, Yu Y, Kan X, Cui W, Huang J et al. Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models. 2024:15496-523.
6. Huang L, Yu W, Sun L, Guo D, Liu Y, Niu M et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. 2023.
7. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E et al. Sparks of artificial general intelligence: Early experiments with GPT-4. 2023.
8. Lin C, Cho C, Strom RG, Mitchell S, Hadjipavlou A, Wang J et al. A comparative analysis between ChatGPT versus NASS clinical guidelines for adult isthmus spondylolisthesis. *NASSJ*. 2025;PII:S2666-5484(25)00019-8.
9. Ahmed W, Saturno M, Rajjoub R, Arroyave JS, Yolcu YU, Fridley JS et al. ChatGPT versus NASS clinical guidelines for degenerative spondylolisthesis: a comparative analysis. *Eur Spine J*. 2024;33:4182-203.
10. Nori H, King N, McKinney SM, Horvitz E, King D. Capabilities of GPT-4 on medical challenge problems. *arXiv*. 2023.
11. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;27:120.
12. Kayastha A, Lakshmanan K, Valentine M, Zaidat B, Yolcu YU, Fridley JS et al. Lumbar disc herniation with radiculopathy: a comparison of NASS guidelines and ChatGPT. *North Am Spine Soc J NASSJ*. 2024;19:100333.
13. Khoilyan A, Cho C, Strom RG, Mitchell S, Hadjipavlou A, Wang J et al. Evaluation of GPT-4 concordance with north American spine society guidelines for lumbar fusion surgery. *North American Spine Society Journal (NASSJ)*. 2025;21:100580.
14. Lang P, Bae J, Kim JH, Ahn JS, Lee SH, Cho WH et al. Analyzing large language models' responses to common lumbar spine fusion surgery questions: A comparison between ChatGPT and Bard. *Neurospine*. 2024;21:633-41.
15. Duey A, Nietsch K, Zaidat B, Yolcu YU, Pandey AS, Fridley JS et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. *Spine J*. 2023;23:1684-91.
16. Zaidat B, Shrestha N, Rosenberg AM, Yolcu YU, Pandey AS, Fridley JS et al. Performance of a large language model in the generation of clinical guidelines for antibiotic prophylaxis in spine surgery. *Neurospine*. 2024;21:128-46.
17. Shrestha N, Zaidat B, Yolcu YU, Pandey AS, Fridley JS. Performance of ChatGPT on NASS Clinical Guidelines for the Diagnosis and Treatment of Low Back Pain A Comparison Study. *Spine*. 2024;49:640-51.
18. Kreiner D, Cho C, Strom RG, Mitchell S, Hadjipavlou A, Wang J et al. The double-edged sword of generative AI: surpassing an expert or a deceptive "false friend"? *Spine J*. 2025;S1529-9430(25)00122-6.
19. Sarikonda A, Yolcu YU, Pandey AS, Fridley JS. Evaluating the Adherence of Large Language Models to Surgical Guidelines: A Comparative Analysis of Chatbot Recommendations and North American Spine Society (NASS) Coverage Criteria. *Cureus*. 2024;16.
20. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. 2020;33:9459-74.
21. Chen J, Lin H, Han X, Sun L. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024;16:17754-62.
22. Gao J, Ren J, Yang S, Li D, Qiu C, Zhang K et al. Retrieval-augmented generation for large language models: A survey. 2023.
23. Ali R, Hopkins BS, Patel B, Ali M, Beecher S, Krieg SM et al. AtlasGPT: a language model grounded in neurosurgery with domain-specific data and document retrieval. *J Neurosurg*. 2025;18:1-8.
24. Kosprdic M, Bashir N, Novakovic B, Labus A, Bogdanovic Z, Barjaktarovic L et al. Verif.ai: Towards an Open-Source Scientific Generative Question-Answering System with Referenced and Verifiable Answers. *The Sixteenth International Conference on Evolving Internet Athens, Greece*. 2024.

25. Barkley L, van der Merwe B. Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models. arXiv:2410.19385v1. 2024.
26. Li Y. A Practical Survey on Zero-shot Prompt Design for In-context Learning. arXiv:2309.13205. 2023.
27. Cho C, Amini DA, Krisa M, Strom RG, Weiss D, Mitchell S et al. Guideline summary review: an evidence-based clinical guideline for the diagnosis and treatment of adults with neoplastic vertebral fractures. *Spine J.* 2025;S1529-9430(25)00168-8.
28. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv preprint. 2023.
29. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P et al. Language Models are Few-Shot Learners. arXiv preprint. 2020.
30. Patel D, Raut G, Zimlichman E, Dagan A, Forkert ND, Joshi A et al. The Limits of Prompt Engineering in Medical Problem-Solving: A Comparative Analysis with ChatGPT on calculation based USMLE Medical Questions. 2023.
31. Zagher J, Naguib M, Bjelogrić M, Eng LS, Krzyzanowska MK, Goldenberg A et al. Prompt Engineering Paradigms for Medical Applications: Scoping Review. *J Med Internet Res.* 2024;26:e60501.
32. Wu J, Nishida T, Moghimi S, Weinreb RN, Goldbaum MH, Liebmann JM et al. Effects of prompt engineering on large language model performance in response to questions on common ophthalmic conditions. *Taiwan J Ophthalmol.* 2024;14:454-7.
33. Shah K, Xu AY, Sharma Y, Wang Z, Jaremko JL, Mushahwar VK et al. Large Language Model Prompting Techniques for Advancement in Clinical Medicine. *J Clin Med.* 2024;13.
34. Xu R, Cui H, Yu Y, Kan X, Cui W, Huang J et al. Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand. 2024:15496-523.
35. Dietrich N, Stubbert B. Evaluating Adherence to Canadian Radiology Guidelines for Incidental Hepatobiliary Findings Using RAG-Enabled LLMs. *Can Assoc Radiol J.* 2025;27:8465371251323124.
36. McDuff D, Schaekermann M, Tu T, Karthikesalingam A, Webster D, Corrado G et al. Towards Accurate Differential Diagnosis with Large Language Models. arXiv:2312.00164v1. 2023.
37. Mehta S, Gupta A, Ekbal A, Bhattacharyya P. Halu-NLP at SemEval-2024 Task 6: MetaCheckGPT - A Multi-task Hallucination Detection using LLM uncertainty and meta-models. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. 2024.
38. Rajjoub R, Arroyave JS, Zaidat B, Yolcu YU, Pandey AS, Fridley JS et al. ChatGPT and its role in the decision-making for the diagnosis and treatment of lumbar spinal stenosis: a comparative analysis and narrative review. *Global Spine J.* 2023;14:21925682231195783.
39. Asai A, Min S, Zhong Z, Chen D. Retrieval-based language models and applications. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*. 2023:41-6.
40. Rodriguez JA, Alsentzer E, Bates DW. Leveraging large language models to foster equity in healthcare. *J Am Med Inform Assoc.* 2024;31:2147-50.
41. Hopkins B, Ali R, Patel B, Beecher S, Krieg SM, Horbinski C et al. AtlasGPT: dawn of a new era in neurosurgery for intelligent care augmentation, operative planning, and performance. *J Neurosurg.* 2024;140:1211-4.
42. Xiong G, Jin Q, Lu Z, Zhang J, Peng J, Dou D et al. Benchmarking retrieval-augmented generation for medicine. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand. 2024:6233-51.
43. Wan A, Wallace E, Klein D. What evidence do language models find convincing? arXiv preprint arXiv:2402.11782. 2024.
44. Jiang Y, Chen J, Yang D, Wang S, Li H, Liu J et al. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. arXiv preprint arXiv:2406.11451. 2024.
45. Penkov S. Mitigating hallucinations in large language models via semantic enrichment of prompts: Insights from biobert and ontological integration. In: *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*. 2024:272-6.
46. Budakoglu G, Emekci H. Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning and Their Synergistic Fusion for Enhanced Performance. *IEEE Access.* 2025;PP:1-1.



This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC-BY\) license 4.0](https://creativecommons.org/licenses/by/4.0/)

APPENDIX A.

Table A1. Two examples of NASS questions prompt engineered as ZS and KI into Verif.ai and responses generated from the NASS Evidence-Based Clinical Guideline for the Diagnosis and Treatment of Adults with Neoplastic Vertebral Fractures.

Clinical Questions (ZS)	Guideline Recommendation	Verif.ai Response (ZS)	Clinical Questions (KI)	Verif.ai Response (KI)
Natural History Question 1: Does underlying histology affect the natural history of metastatic neoplastic vertebral fractures?	A systematic review of the literature yielded no studies to adequately address this question.	Yes, underlying histology does affect the natural history of metastatic neoplastic vertebral fractures. The abstracts provided indicate that vertebral biopsy before vertebroplasty can reveal unexpected positive histology, which can impact patient management. In a series of 324 biopsies performed just before vertebroplasty, 2% of cases had histology positive for an underlying pathology, including multiple myeloma, lymphoma/leukemia, spondylitis, metastasis, and hemangioma. The incidence of positive biopsies was significantly higher in patients younger than 73 years old, suggesting that the age of the patient may also be a factor in the natural history of metastatic neoplastic vertebral fractures (PUBMED:33843004). Additionally, the natural history of osteoporotic vertebral fractures is benign, with many patients experiencing sufficient pain relief during the first three months with conservative treatment. However, at the end of one year of conservative treatment, 40% may still have pain with VAS-scores ≥ 4 . The preventive measures for osteoporotic vertebral fractures include an active lifestyle, adequate nutrition, prevention of immobilization with physical therapy, and pharmacological therapy after fracture risk assessment (PUBMED:35301844). In contrast, the natural history of neoplastic locoregional disease, such as peritoneal carcinomatosis, can be crucial to guide the choice of the correct locoregional treatment. For instance, peritoneal carcinomatosis is considered a main step of disease progression for ovarian cancer and often for gastrointestinal tumors. However, when the tumors are confined to the surface of the peritoneum, they can be considered as localized diseases, and selected patients with neoplastic deposits confined within a definite site or region of the body can be considered candidates for curative treatment (PUBMED:16767897). Overall, the natural history of metastatic neoplastic vertebral fractures is influenced by the underlying histology, which can affect the severity of the fracture, the patient's prognosis, and the treatment options available.	For the question, "does underlying histology affect the natural history of metastatic neoplastic vertebral fractures?" is the answer, no studies adequately address this question?	Based on the provided abstracts, there is no study adequately addressing the question of whether underlying histology affects the natural history of metastatic neoplastic vertebral fractures. The abstracts focus on various aspects of vertebral fractures, such as the utility of vertebral biopsy before vertebroplasty (PUBMED:33843004), the management of oligometastatic disease in esophago-gastric cancer (PUBMED:36358619), the diagnosis of osteoporosis (PUBMED:7299006), the safety of vertebroplasty (PUBMED:16908547), the natural history and pathophysiology of flexible flatfoot (PUBMED:19963167), the comparison of coronary surgery with medical management (PUBMED:1013007), the safety and functional improvement of kyphoplasty (PUBMED:18561792), the treatment of pathologic vertebral compression fractures (PUBMED:32032083), and a practical approach to children with recurrent fractures (PUBMED:26138844). However, none of these abstracts provide information on the natural history of metastatic neoplastic vertebral fractures in relation to underlying histology. Therefore, the answer to the instruction is yes, no studies adequately address this question.

Clinical Questions (ZS)	Guideline Recommendation	Verif.ai Response (ZS)	Clinical Questions (KI)	Verif.ai Response (KI)
Imaging Diagnosis Question 1: Which imaging modalities and findings are most sensitive and specific for the accurate diagnosis of and treatment planning for neoplastic vertebral fractures?	Findings on routine MRI sequences are suggested as a method to differentiate osteoporotic from neoplastic vertebral fractures. Grade of Recommendation: B	<p>The abstracts provided do not directly address the specific sensitivity and specificity of various imaging modalities for the diagnosis of neoplastic vertebral fractures. However, they do discuss the use of different imaging modalities and techniques for characterizing vertebral fractures and differentiating between osteoporotic and neoplastic fractures.</p> <p>For instance, magnetic resonance imaging (MRI) with spin-echo echo-planar diffusion-weighted imaging (SE-EPI-DWI) has been shown to provide reliable information to support MRI diagnosis of neoplastic versus osteoporotic fractures, with apparent diffusion coefficient (ADC) values being a useful adjunctive parameter (PUBMED:22411038). Additionally, MRI with T1-weighted spin-echo and short inversion time inversion-recovery sequences has been used to evaluate the occurrence, location, and shape of the fluid sign in acute vertebral compression fractures, which can be an additional sign of osteoporosis and rarely occurs in metastatic fractures (PUBMED:12461253). Furthermore, a scoring system based on MRI and computed tomography (CT) findings have been devised to enhance the accuracy of imaging diagnosis of fractures caused by benign or malignant spinal lesions (PUBMED:16075165). MRI has also been used to establish a diagnostic scoring system for discriminating osteoporotic vertebral fractures from those caused by malignant tumors (PUBMED:16075165).</p> <p>In summary, while the abstracts do not provide specific sensitivity and specificity data, they suggest that MRI, particularly with the use of diffusion-weighted imaging and scoring systems, can be useful in differentiating between osteoporotic and neoplastic vertebral fractures.</p>	<p>For the question, "Which imaging modalities and findings are most sensitive and specific for the accurate diagnosis of and treatment planning for neoplastic vertebral fractures?" is the recommendation that Findings on routine MRI sequences are suggested as a method to differentiate osteoporotic from neoplastic vertebral fractures adequately address this question?</p>	<p>Yes, the recommendation that findings on routine MRI sequences are suggested as a method to differentiate osteoporotic from neoplastic vertebral fractures adequately addresses the question. The abstracts provided support the use of MRI, specifically sagittal T1-weighted MRI, in differentiating metastatic and osteoporotic vertebral fractures (PUBMED:31771804). Characteristic findings on MRI, such as pedicle or posterior element involvement, convex posterior border of the vertebral body, epidural infiltration, and diffuse homogeneous low signal intensity, are useful in the differential diagnosis of metastatic and osteoporotic vertebral fractures (PUBMED:31771804). Additionally, diffusion-weighted MRI (DWI) has been shown to provide reliable information to support MRI diagnosis of neoplastic versus osteoporotic fractures, with ADC values significantly higher in neoplastic fractures compared to osteoporotic ones (PUBMED:22411038). Furthermore, a scoring system based on MRI and CT findings has been devised to enhance the accuracy of imaging diagnosis of fractures caused by benign or malignant spinal lesions (PUBMED:16075165). These findings suggest that MRI, particularly with the use of specific sequences and scoring systems, is a sensitive and specific imaging modality for the accurate diagnosis and treatment planning for neoplastic vertebral fractures.</p>