Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SafeLLM: Extraction as a Hallucination-Resistant Alternative to Rewriting in Safety-Critical Settings

Julia Ive

j.ive@ucl.ac.uk

University College London

Felix Jozsa

University College London

Evridiki Georgaki

University College London

Nabeel Sheikh

Somerset NHS Foundation Trust

Emma Cattell

Somerset NHS Foundation Trust

Nick Jackson

King's College Hospital

Paulina Bondaronek

University College London

Ciaran Scott Hill

University College London

Richard Dobson

University College London

Research Article

Keywords: Large language models, retrieval-augmented generation, natural language processing, extractive question answering, clinical guidelines, hallucination mitigation, safety-critical systems

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11288339/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SafeLLM: Extraction as a Hallucination-Resistant Alternative to Rewriting in Safety-Critical Settings

Julia Ive[1,6,*], Felix Jozsa[1,2,6,†], Evridiki Georgaki[1], Nabeel Sheikh[3], Emma Cattell[3], Nick Jackson[4], Paulina Bondaronek[1], Ciaran Scott Hill[2], and Richard Dobson[1,5]

[1]Institute of Health Informatics, University College London, London, UK
[2]National Hospital for Neurology and Neurosurgery, Queen Square, London, UK
[3]Somerset NHS Foundation Trust, UK
[4]King's College Hospital, Denmark Hill, London, UK
[5]King's College London, London, UK
[6]These authors contributed equally: Julia Ive, Felix Jozsa
[*]j.ive@ucl.ac.uk
[1]f.jozsa@ucl.ac.uk

ABSTRACT

Background: Large language models (LLMs) are increasingly used to access organisational documentation, including standard operating procedures (SOPs), HR policies and institutional guidelines. However, retrieval-augmented generation (RAG) approaches that rely on free-form rewriting can introduce hallucinations and unstable trade-offs between completeness and conciseness, particularly in safety- and compliance-critical settings.
Objectives: To evaluate extraction as a hallucination-resistant alternative to rewriting-based RAG, and to compare strategies that balance precision, recall and safety across document types and model scales.
Methods: We compare multiple prompting strategies, including line-number-based source selection, extraction of relevant guideline sentences with explicit annotation of safety-related content, and a multi-stage pipeline that refines draft answers using supporting evidence from the source guideline. Constrained extraction strategies enable more reliable selection of clinically relevant source content, while direct verbatim copying provides a strong precision-oriented baseline. Experiments are conducted on documents of varying length (local NHS guidelines in acute care and oncology, as well as UK-wide NICE guidelines) and structures, using both frontier-scale and smaller locally deployable models. Performance is assessed using automatic metrics and via human expert evaluation of relevance and completeness.
Results: Line-number selection provides the most consistent recall improvements for shorter, protocol-like guidelines across both large and small models, achieving medical-term recall of up to 95%. For substantially longer guidelines, draft-first extraction achieves the highest recall with selected frontier models, reaching up to 97%. Safety-oriented approaches improve precision but introduce systematic omissions, and multi-stage filtering further amplifies this trade-off.
Conclusions: Line-number selection provides a robust alternative to rewriting-based RAG for accessing organisational documentation. Drafting-based approaches may offer additional benefits in very long-document settings. These findings highlight the importance of adapting retrieval strategies to document structure, model scale and the risks of omission.

Keywords: Large language models; retrieval-augmented generation; natural language processing; extractive question answering; clinical guidelines; hallucination mitigation; safety-critical systems.

Introduction

Administrative and operational documents, including clinical guidelines, standard operating procedures (SOPs), and organisational policies, provide consensus-based instructions for decision-making and task execution[1]. Across domains, adherence to such documents is a core requirement of good professional practice and is associated with improved safety, compliance, and operational consistency[2–8].

In practice, high-level guidance is complemented by local or departmental documents that adapt recommendations to specific workflows, constraints, and organisational structures. Users consult these resources to answer focused, task-oriented questions, for example, what actions to take, which approvals are required, or how to escalate a situation. However, their use is often hindered by complexity and time pressure[9–12], challenges that are amplified by increasing workload and growing document volume[13–15]. In high-stakes or time-critical settings, these limitations can lead to delays, errors, or non-adherence.

Large Language Models (LLMs) offer a promising approach to document-based question answering (Q&A)[16], but their

deployment is constrained by hallucinations, i.e. generated content not supported by source evidence[17]. Retrieval-Augmented Generation (RAG) mitigates this by grounding responses in retrieved documents[18], improving factual consistency[19]. However, it does not fully eliminate errors: models may omit critical steps, introduce irrelevant details, or fail to faithfully reproduce source text, posing risks in safety- and compliance-sensitive contexts[3,20–22].

A central challenge in document-based Q&A is balancing interpretation with faithful reproduction. Some queries require integrating information distributed across multiple sections or adapting conditional instructions to a specific scenario. Others demand precise, verbatim reproduction to preserve meaning, particularly where exact wording carries operational or regulatory significance.

Current approaches focus on improving generation quality, rather than controlling when generation should occur[3,22]. Here, we take the opposite approach: we minimise generation. We constrain models to operate directly on source text through extraction, enabling them to identify and present relevant information without rewriting. This reframes the task from answer generation to evidence selection, reducing opportunities for hallucination while preserving fidelity to source material.

Our main contribution is the development and evaluation of an extraction framework for the use case of clinical question answering that minimises hallucinations while balancing the risk of adding information versus omitting clinically critical details. In contrast to rewriting-based approaches, we constrain LLMs to select and copy guideline text, targeting high fidelity to source material, an essential requirement in settings where even small inaccuracies in wording (e.g., drug dosing, contraindications, or escalation criteria) may impact patient safety. We evaluate this approach across multiple guideline use cases that vary in length and structure, including concise acute care protocols from University College London Hospital (UCLH), oncology-focused institutional guidelines from Somerset NHS Trust, and large-scale national recommendations (NICE). These settings capture a spectrum of real-world decision-making scenarios. Evaluation is conducted using human expert-developed question–answer pairs grounded in these documents, reflecting practical clinical information needs under time pressure. We further assess performance across a range of LLM configurations, including both open-weight and commercial models, to characterise robustness across deployment settings. Rather than attempting to make LLMs better generators, we argue that constraining them offers a more reliable path for deployment in safety-critical clinical environments.

Methods

Data

We evaluate our methods across three datasets spanning distinct document styles and scales: concise acute care guidelines (UCLH), moderate-length institutional guidelines (Somerset NHS Trust), and large-scale national recommendations (NICE).

The UCLH dataset consists of 96 clinically grounded questions grouped under 24 acute care topics, designed by FJ and NJ to reflect real-world scenarios encountered by frontline clinicians. Topics include, for example, *acute coronary syndrome*, where questions probe escalation decisions (e.g., "If first-line therapy does not relieve symptoms, when is a second dose indicated?"), and *acute kidney injury*, focusing on investigation and management steps (e.g., "Which blood tests are essential on admission?"). Reference answers were constructed via direct selection of relevant guideline lines to ensure grounding in source material.

The Somerset NHS Trust dataset follows a similar question design process but differs in annotation strategy. Here, reference answers were generated exclusively through clinician free writing (NS and FJ). These guidelines are predominantly oncology-oriented and capture acute complications of cancer and its treatment, such as *neutropenic sepsis*, *metabolic disturbances*, and *treatment-related toxicities*. Despite being moderate in length, they are strongly acute in nature, often focusing on time-critical recognition, escalation, and stabilisation decisions in vulnerable patient populations.

In contrast, the NICE dataset was generated synthetically following the protocol of Ding et al.[23], using `gpt-4o-mini` to produce clinical scenario–question–guideline paragraph triplets from full-length recommendation documents. To reflect realistic model inputs, we merge the scenario description and associated question into a single input prompt. All generated outputs were subsequently manually reviewed by human experts to ensure clinical validity. This dataset captures scenario-related inputs, for example within *chronic kidney disease* or *heart failure* management. A representative example is:

"A 68-year-old patient with a history of hypertension and type 2 diabetes presents with progressive fatigue and ankle swelling over several weeks. Blood tests show reduced eGFR and elevated creatinine. According to current guidelines, what investigations and initial management steps should be undertaken, and when should specialist referral be considered?"

This contrasts with the shorter, focused queries in UCLH and Somerset, such as: *"Which blood tests are essential in acute kidney injury?"* (see Appendix Tables A1 and A2).

All underlying guideline documents were extracted from PDF files using the `pdftotext` utility to preserve original wording, followed by manual cleaning to remove tables of content, correct formatting artefacts and ensure fidelity to source content.

Across datasets, question formulation varies substantially, from concise queries in UCLH (mean 9.6 words) to longer, scenario-based inputs in NICE (mean 44.85 words), with Somerset NHS Trust occupying an intermediate position. Importantly, the guideline corpora differ by an order of magnitude in length: short UCLH documents (mean 746 words), moderate-length

Table 1. Summary statistics of the evaluation datasets spanning three document-length regimes: concise acute care guidelines (UCLH), moderate-length institutional guidelines (Somerset NHS Trust), and large-scale national recommendations (NICE). Values are reported as mean (min–max). * NICE questions were sampled randomly keeping original proportions per topic.

	UCLH	Somerset NHS Trust	NICE
Questions (n)	96	50	100*
Guidelines (n)	25	6	10
Question length (words)	9.6 (3–17)	11.2 (2–21)	44.85 (26–72)
Answer length (words)	39.4 (1–139)	36.5 (5–145)	30.0 (9–87)
Guideline length (lines)	73 (34–152)	438 (317–636)	881 (258–1466)
Guideline length (words)	746 (375–1360)	3153 (2372–4510)	5224 (1437–8007)
Total corpus (words)	18,578	18,918	52,236
Total corpus (lines)	1,824	2,628	8,809

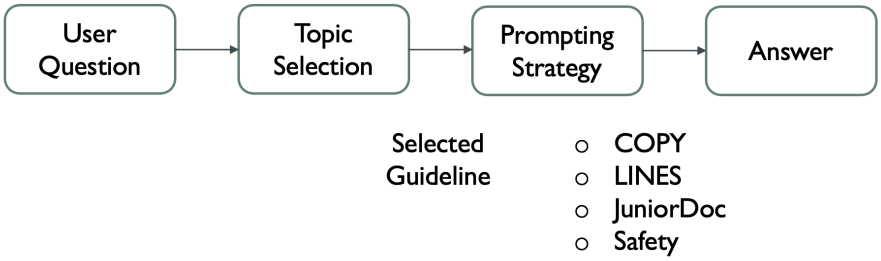


Figure 1. Overview of the proposed guideline question-answering pipeline. Given a user question, the system first performs **Topic Selection** to identify the most relevant clinical guideline from the guideline index. Once the relevant guideline is selected, the system applies one of several **prompting strategies** to extract the evidence required to answer the question. The evaluated strategies include: (i) **COPY**, where the model extracts verbatim sentences; (ii) **LINES**, where the model selects relevant guideline line numbers; (iii) **JuniorDoc**, a multi-agent workflow where a junior-doctor-style draft answer is generated, verified against the guideline, and filtered by a safety editor; and (iv) **Safety**, which extracts the guideline sentences relevant to the question and annotates safety-critical content to support safe evidence-based responses.

Somerset NHS Trust guidelines (mean 3,153 words), and long-form NICE recommendations (mean 5,224 words; maximum 8,007 words) (Table 1).

Prompting Strategies for Guideline-Based Clinical Question Answering

This section outlines the methods evaluated in this study, starting with a baseline retrieval approach based on semantic similarity between the user question and guideline text. We then present a set of prompting pipeline methods that combine an initial topic identification step with different prompting regimes for extracting relevant evidence from clinical guidelines (see Fig. 1, see summary of the approaches in Table 2).

Baseline Retrieval (BASE) As a baseline retrieval strategy, we construct a repository of information chunks derived from guideline text, each representing a self-contained unit of clinically relevant content that may be used to answer a question. From each guideline, we extract guideline sections and, for each section, the section title, section text, and a representative actionable question.

These sections are extracted automatically from guideline paragraphs using an LLM prompt. This section extraction prompt (Appendix Listing A1) decomposes each paragraph into smaller, clinically meaningful overlapping subparts. For each section, it then generates a short descriptive title and an actionable clinical question. By associating each text segment with a representative question, we aim to reduce sensitivity to phrasing variation between guideline text and associated queries. Following this decomposition, we retain the complete original paragraph as an additional section and generate a corresponding descriptive title and actionable question in a separate step.

At indexing time, each subpart is embedded using one of two representations (Figure 2): (a) *text*, based on guideline topic, section title, and section text concatenated into single string; or (b) *text + question*, which additionally incorporates the representative question (guideline topic, section title, section text and representative question concatenated into single string). At query time, the user question is embedded using the same model and compared to all embeddings (variant (a) or (b)) via

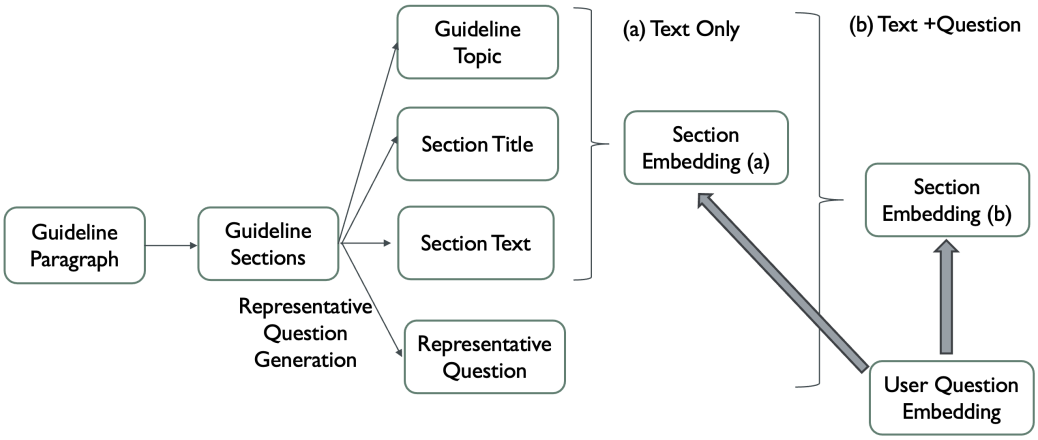


Figure 2. Construction of the representative question–section index used by the baseline retrieval strategies (BASE). Each guideline document is first segmented into paragraphs, from which clinically coherent overlapping sections are extracted. For every section, a representative question and section title are generated using an LLM. Each subpart is represented by four elements: the guideline topic, section title, representative question, and verbatim section text. Two indexing variants are considered: (a) *text*, where embeddings are computed from guideline topic, section title and text concatenated together into a single string; and (b) *text + question*, where embeddings are computed as in (a) plus the representative question. Representative questions act as a semantic interface between guideline text and user queries. At query time, the user question is embedded using the same embedding model and compared against section embeddings via cosine similarity.

cosine similarity. Retrieval is then performed by selecting the subparts whose representations are most semantically aligned with the query.

For section-based retrieval, duplicate (`guideline_id`, `section_id`) entries are removed before all candidate sections within the selected guideline are scored using cosine similarity between the question and section embeddings. Sections are ranked by decreasing cosine similarity, without clipping, shifting or min–max normalisation. The raw scores s_i are converted into non-negative weights using softmax.

Negative similarities are retained and receive lower, but still positive, probability mass, while zero vectors are assigned a cosine similarity of zero. Starting with the highest-ranked section, sections are selected sequentially until their cumulative probability reaches 0.95 or five sections have been selected, whichever occurs first.

Table 2. Summary of prompting strategies. All methods begin with Topic Selection, which identifies the most relevant guideline document before extraction.

Abbrev	Strategy	Key Idea	Hallucination Control
BASE	Baseline Retrieval	Retrieve relevant guideline sections using embedding similarity between user queries and those sections	Retrieval-only baseline
Topic	Topic Selection	Select the single most relevant guideline document before answering the question.	Limits search space to one guideline
COPY	Verbatim Sentence Extraction	Model copies only exact sentences from the guideline required to answer the question	Enforces copying
LINES	Line Selection	Model outputs line numbers corresponding to relevant guideline text. The final answer is reconstructed from those lines	Guarantees verbatim text. Minimises hallucination
Junior	Junior Doctor Workflow	Multi-agent reasoning: junior doctor draft answers are rewritten using supporting evidence from the source guideline	Grounding
Safety	Annotated Safety Extraction	Model selects the relevant set of guideline sentences that answer the question and annotates each with a role label (e.g., answer, detail, safety)	Copying with explicit role labelling

Step 1: Topic Selection (Topic) Topic selection is the first step across all strategies. Given a user query, the model selects exactly one guideline topic from a predefined index of topics. This design follows the workflow observed in clinical practice. Questions typically relate to a single diagnosis, meaning clinicians first identify the relevant guideline before consulting detailed

sections (see Appendix Listing A2).

Since our study focuses on acute medical problems encountered on an undifferentiated hospital ward, clinical questions can generally be mapped to a single primary issue, even for patients with substantial multimorbidity. Furthermore, many acute conditions, such as hyperkalaemia or hypoglycaemia, are managed according to standard protocols that are largely independent of a patient's underlying comorbidities. Where management does require adaptation for specific patient groups, such as those with chronic kidney disease, these considerations are typically captured within dedicated sections of the same guideline. This structure supports a topic-first retrieval strategy, in which a single guideline is identified before relevant subsections are selected.

Step 2: Evidence Retrieval Strategies

Verbatim Sentence Strategy (COPY) We first introduce COPY as a prompting baseline that probes the model's ability to directly copy relevant information from the guideline. The model is instructed to return only verbatim sentences that answer the user question, with explicit constraints forbidding paraphrasing and requiring exact reproduction of the source text.

This setup probes models' capacity for faithful extraction, providing a reference point for more constrained strategies (see Appendix Listing A3).

Line Selection Strategy (LINES) Building on this baseline, the LINES strategy introduces a more sophisticated mechanism. The guideline text is augmented with line numbers and provided to the model together with the user question. The model is then tasked with identifying the relevant set of line numbers required to answer the question.

The final answer is reconstructed deterministically by retrieving the corresponding guideline lines in post-processing. This approach enforces tighter control over the extracted content and guarantees the absence of unsupported text, as the output consists exclusively of selected guideline line numbers (see Appendix Listing A4).

Junior Doctor Workflow (Junior) We also design a multi-agent prompting strategy that mimics a realistic clinical reasoning workflow.

Agent 1 acts as a junior doctor and generates an initial answer based only on training memory, without consulting external documents. This simulates the real-world setting in which clinicians often rely on prior knowledge under time pressure. The agent is instructed to provide concise answers containing actionable details such as thresholds, medication doses, or monitoring steps (see Appendix Listing A5). Agent 2 then compares the draft answer against the official guideline text and selects the guideline sentences that best support, clarify, or correct it, preserving their original order (see Appendix Listing A6, see Figure 3).

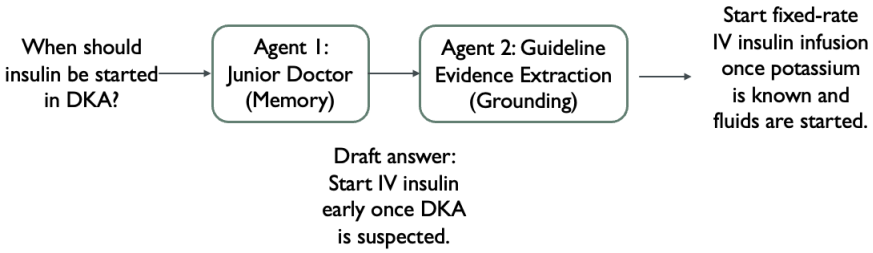


Figure 3. Illustration of the **JuniorDoc** prompting workflow. The user question is first processed by **Agent 1**, which acts as a junior clinician reasoning from memory and produces an initial draft answer. This draft may be incomplete or imprecise because it did not yet consult the guideline text. **Agent 2** then grounds the answer by extracting verbatim sentences from the relevant clinical guideline section that support, refine, or correct the initial draft. This multi-stage design combines clinical reasoning and evidence grounding.

Minimal Safety Extraction (Safety) We also evaluate a prompting strategy that instructs the model to extract the minimal safe set of sentences required to answer the user's question. The model is prompted to select the smallest subset of full sentences that (i) directly answer the question and (ii) include any safety-critical information. Each selected sentence is annotated with a single reason label indicating its role in the response: `answers_question`, `adds_required_detail`, or `safety_critical`. This prompt also enforces strict constraints: sentences must be copied verbatim from the guideline, paraphrasing is not allowed, and the original order must be preserved (see Appendix Listing A7).

The Safety strategy is motivated by the observation that guideline answers often contain both essential recommendations and surrounding explanatory text. By explicitly defining roles, the model is guided to retain only the most relevant content.

In addition to the sentence extraction methods, we implemented line-based variants of the Safety prompting strategy. The Safety-Lines variant applies the original Safety prompt but asks to output line numbers instead of sentences. For API settings, this approach also substantially reduces the number of generated tokens and enables lightweight downstream reconstruction of the final response directly from the original guideline (see Appendix Listing A8).

LLM Types

To evaluate the robustness of our methods across different model architectures and deployment settings, we conducted experiments using several open-weight LLMs available through the Ollama framework[1]. Specifically, we evaluated `gpt-oss-120b`[24] and `qwen2.5:32b`[25]. These models provide openly available weights, enabling reproducibility and supporting deployment in secure environments where external API access may be restricted (e.g., clinical or trusted research environments). Using open-weight models also allows evaluation under controlled and identical inference conditions. All models were run with default hyperparameters except for the temperature parameter, which was fixed at 0 to minimise non-deterministic variation.

For the baseline retrieval strategy (BASE), we additionally experimented with three embedding models available through Ollama: `nomic-embed-text`, `embeddinggemma`, and `qwen3-embedding`. These models differ primarily in their training objectives and underlying model families. `nomic-embed-text` is designed specifically for semantic search and retrieval tasks, `embeddinggemma` provides lightweight embeddings derived from the Gemma model family[26], and `qwen3-embedding` is based on the Qwen architecture and trained to produce dense representations aligned with instruction-tuned models.

For the open-access NICE guideline data, we additionally evaluate large-scale API-based models to assess how well our extraction strategies generalise to frontier models beyond locally deployable setups. The models evaluated include `deepseek-ai/deepseek-v3` (approximately 671B parameters) and `qwen/qwen3-235b-a22b-instruct-2507` (235B parameters), both accessed through the Replicate platform, and `gpt-5-mini`, accessed through Azure OpenAI in Microsoft Azure AI Foundry. For the longer Somerset guidelines, we additionally evaluate the 20B-parameter `gpt-oss:20b` model, run locally using Ollama. We exclude `gpt-4o-mini` from the evaluation because it was used to create the synthetic NICE dataset.

Evaluation

We evaluate the proposed methods using both automatic metrics and human assessment.

Automatic Evaluation We first report average answer length (AnsLen) to understand model verbosity, as longer outputs tend to increase recall-oriented metrics while potentially reducing precision.

We then evaluate overlap with reference answers using ROUGE-L[27], which can be interpreted simply as a form of *sentence-level overlap*: it measures how much of the reference content is recovered *in the correct order*, rewarding correct reconstruction of long chunks of text rather than fragmented matches.

To capture more fragmented lexical overlap, we report word-level precision, recall, and F1 (P, R, F1). Both reference and generated answers are normalised by removing non-alphanumeric characters, lowercasing, and splitting into tokens on space. Precision is defined as the proportion of generated tokens supported by the reference, recall as the proportion of reference tokens recovered, and F1 as their harmonic mean. This formulation is robust to word order and minor formatting differences.

Finally, we assess entity-level fidelity using medical named entity recognition (NER, medical term) precision, recall, and F1. We use MedCAT[28] for biomedical named entity recognition and concept linking. Specifically, we employ the latest available MedCAT model packs based on SNOMED CT and UMLS, accessed through UMLS-authenticated releases (v2.1.0 with the CogStack model pack v2_Snomed2025_MIMIC_IV_bbe806e192df009f).[2] MedCAT identifies clinical concepts in guideline texts and maps them to standard biomedical terminologies. Evaluation is performed over sets of unique extracted term spans, with precision and recall defined as the overlap between predicted and reference entity sets, and F1 as their harmonic mean. This provides a targeted measure of whether clinically salient concepts (e.g., drugs, conditions, symptoms) are correctly preserved, independent of surrounding text.

Together, these metrics provide complementary views of extraction quality: AnsLen captures verbosity, ROUGE-L reflects sentence-level overlap, word-level scores capture fragmented lexical overlap, and medical term metrics focus specifically on clinically relevant entities.

We report some of these complementary measures of extraction quality in diagrams, plotting medical-term recall against answer length to show the trade-off between clinical information coverage and verbosity. The ideal extraction strategy appears towards the upper-left of each plot, combining high medical-term recall with a short, concise answer. Table 3 summarises the automatic evaluation metrics and their roles in this study.

[1] https://ollama.com
[2] https://medcat.readthedocs.io/en/latest/main.html

Table 3. Automatic evaluation metrics highlighting complementary aspects of verbosity, coverage and clinical fidelity.

Metric	Type	Purpose in this study
AnsLen	Output length	Measures average answer length; captures verbosity
ROUGE-L	Sequence-level overlap	Measures ordered overlap with the reference (longest common subsequence); captures sentence-level recall and rewards coherent reconstruction
Word P/R/F1	Word-level precision, recall and F1	Word overlap robust to formatting differences. Precision reflects unsupported additions, recall reflects coverage, and F1 balances both
MedicalTerm P/R/F1	Medical term overlap	Evaluates preservation of clinically salient concepts

Table 4. Comparison of index characteristics across datasets. UCLH (acute care) shows concise, action-oriented structure; Somerset (oncology) is more narrative; NICE (national guidance) is highly heterogeneous.

Statistic	UCLH	Somerset	NICE
Number of guidelines	25	6	10
Number of sections	987	469	4,726
Sections per guideline, mean	39.5	78.2	472.6
Sections per guideline, range	6–74	48–125	12–849
Section length, words mean	31.4	65.8	18.6
Question length, words mean	8.7	8.2	8.3

Human Evaluation In addition to automatic metrics, we conducted a human evaluation to assess the clinical quality of the generated answers. While automatic metrics quantify textual overlap, they are limited in their ability to capture clinical correctness, contextual appropriateness, and the safety implications of missing or added information. In guideline-based question answering, small deviations in wording, omission of key recommendations, or inclusion of irrelevant details may not substantially affect metric scores but can have significant clinical consequences.

Three clinicians (FJ, EG and NS) evaluated the system outputs against the original guideline documents without access to the reference answers used for automatic evaluation. Outputs for the 96 questions were distributed among the evaluators, with each output receiving one primary rating. To assess inter-annotator agreement, all three clinicians independently evaluated a shared subset of 12 questions covering pericarditis, acute kidney injury and atrial fibrillation. Majority voting was used to assign the final labels for this shared subset, the primary evaluator's rating was retained for the remaining questions. Only one evaluator had participated in developing the UCLH question set.

The evaluators scored answers along two dimensions. The first criterion was *correctness*, indicating whether the information provided was relevant to the clinical question (categories: *very relevant*, *relevant*, *irrelevant*). The second criterion was *completeness*, indicating whether the answer adequately covered the relevant guideline content (categories: *minor omission*, *major omission*, *just right*, *minor addition*, *major addition*). Major omissions are particularly important in clinical contexts because missing key recommendations may lead to incomplete or unsafe guidance.

Results

Baseline Index Construction Applying the baseline indexing procedure (see Methods) across the three data sources yields corpora reflecting differences in each guideline style.

For the UCLH dataset, the indexing process produced 987 sections. Sections are relatively short (mean 31 words), guidelines are decomposed into an average of 41 sections (see Table 4). In the Somerset NHS Trust dataset, the index contains 469 sections across 6 guidelines (mean 78 sections per guideline). Sections are substantially longer than in UCLH (mean 66 words), reflecting a more narrative style with multiple clinical considerations rather than concise decision rules. The NICE dataset is substantially larger and more heterogeneous, comprising 4,726 sections across 10 guidelines (mean 473 sections per guideline). Sections are shorter on average (mean 19 words). Across all datasets, representative questions remain short and consistent (≈ 8 words). Examples are shown in Appendix Table A3. These questions capture key clinical core reasoning patterns, including diagnostic criteria, escalation decisions, and safety-critical alerts.

Baseline retrieval limitations. Our BASE retrieval strategy provides a retrieval baseline, consistently achieving high recall across all configurations (word-level recall = 0.65–0.86; Table 5). This indicates that relevant guideline content is retrieved within the selected sections. However, precision remains substantially lower (word precision = 0.12–0.24), reflecting that retrieved sections frequently include non-essential content.

Table 5. Automatic evaluation of baseline (BASE) retrieval across embedding models and representations for UCLH data. We compare representations based on section text, section topic and title (text only) and representations based on section text, section topic, title and representative LLM-generated question (text + question). Guideline sections are retrieved using cosine similarity between embedded user query and representations, and returned in full as answers. Metrics include ROUGE-L (R-L, sentence-level recall), word-level precision (P), recall (R), and F1, as well as medical term precision, recall, and F1. Answer length (AnsLen) is reported as mean word count (lower is better). BASE achieves consistently high recall across embeddings, but low precision and long outputs, indication non-essential content. Best results per representation are shown in **bold**, second best are underlined.

Embedding	Ans Len	R-L	P	R	F1	Term P	Term R	Term F1
text								
gemma	190.97	0.3026	0.2117	0.8372	0.3151	0.2614	0.8191	0.3717
nomic	245.81	0.1720	0.1244	0.6518	0.1872	0.1479	0.6136	0.2166
qwen3	<u>174.03</u>	<u>0.3218</u>	<u>0.2337</u>	<u>0.8369</u>	<u>0.3340</u>	<u>0.3032</u>	<u>0.8327</u>	<u>0.4089</u>
text + question								
gemma	257.65	0.2248	0.1514	0.7853	0.2381	0.1857	0.7635	0.2771
nomic	255.74	0.2048	0.1366	0.8035	0.2190	0.1779	0.7860	0.2684
qwen3	**181.77**	**0.3332**	**0.2418**	**0.8608**	**0.3430**	**0.3083**	**0.8514**	**0.4105**

Table 6. Embedding similarity statistics for baseline retrieval. Values represent cosine similarity between the query and the top-ranked guideline section. We compare representations based on section text, section topic and title (text only) and representations based on section text, section topic, title and representative LLM-generated question (text + question). Results show overall high similarity across models, with `nomic` achieving the highest mean similarity and lowest variance, and `gemma` showing lower similarity. Best values per representation are shown in **bold**.

Representation	Model	Mean	Std	Max
Text only				
	nomic	**0.76**	**0.05**	0.86
	qwen3	0.75	0.06	**0.88**
	gemma	0.65	0.06	0.83
Text + question				
	nomic	**0.75**	**0.04**	**0.84**
	qwen3	0.72	0.06	**0.84**
	gemma	0.68	0.07	0.82

BASE produces long outputs (mean 174–257 words), as entire guideline sections are returned rather than selectively extracted evidence. This reduces usability of those answers in time-critical clinical settings.

Across embedding strategies, `qwen3` achieves the strongest performance overall. In both representations, it yields the highest ROUGE-L and F1 scores, while also producing the shortest outputs among BASE variants (174 words for text-only).

Augmenting section representations with LLM-generated questions (text + question) produces mixed effects across embeddings. In particular, it is beneficial for `qwen3` and `nomic`, but decreases performance for `gemma`. Also, text + question representations tend to slightly increase output length.

All models achieve relatively high cosine similarity (mean 0.65–0.76, Table 6), confirming that retrieval identifies semantically relevant sections. `gemma` exhibits lower mean similarity and higher variance, consistent with its weaker downstream retrieval performance.

Prompting-based extraction strategies. Across all datasets and model families, prompting-based extraction substantially improves over baseline retrieval, not by retrieving more information, but by selecting it more precisely. Among these approaches for UCLH, LINES consistently emerges as the most reliable strategy for preserving clinically relevant content (Table 7, Figure 4).

Unlike COPY, which serves as a baseline for verbatim extraction, LINES explicitly enforces line-level selection of evidence. This constraint proves critical. Across settings, LINES achieves the highest recall across both word-level and entity-level metrics (e.g., word recall 0.94 and NER recall 0.95 for GPT-OSS; 0.90 and 0.91 for Qwen), with statistically significant improvements over COPY. In safety-critical settings, where missing information may have direct consequences, this recall advantage is decisive. LINES therefore provides the most effective balance of completeness and usability.

Table 7. Performance comparison on the UCLH dataset across prompting-based extraction strategies and model families. Metrics include ROUGE-L (R-L, sentence-level recall), word-level precision (P), recall (R), and F1, as well as medical term precision, recall, and F1. Answer length (AnsLen) is reported as mean word count (lower is better). Statistical significance is assessed via paired bootstrap resampling (10,000 draws with replacement), with 95% confidence intervals and two-sided p-values computed from the bootstrap distribution; significance (*) indicates $p < 0.05$ relative to the COPY baseline. LINES consistently achieves the highest recall among prompting methods, approaching the coverage of baseline retrieval while producing substantially shorter outputs. Bold highlights the top-performing methods per metric, second-best performance is underlined.

Method	AnsLen	R-L	P	R	F1	Term P	Term R	Term F1
GPT-OSS-120B								
copy	**44.96**	**0.7537**	**0.7884**	0.8276	**0.7667**	**0.7979**	0.8485	**0.7807**
lines	109.49	0.6231*	0.5380	**0.9406***	0.6253*	0.5631	**0.9485***	0.6540*
junior	118.00	0.4983*	0.4205	0.8745	0.5085*	0.4414	0.8753	0.5298*
safety	45.05	0.7082*	0.7371	0.8014	0.7169*	0.7424	0.8174	0.7295*
Qwen2.5-32B-Instruct								
copy	39.75	**0.7717**	**0.8199**	0.8054	**0.7762**	**0.8230**	0.8275	**0.7871**
lines	93.91	0.6800*	0.6349	**0.9040***	0.6820*	0.6584	**0.9128***	0.7036*
junior	146.33	0.4942*	0.4415	0.8317	0.5052*	0.4692	0.8354	0.5335*
safety	**39.44**	0.6794*	0.7163	0.7550*	0.6891*	0.7341	0.7676	0.7098*

COPY provides a strong precision-oriented baseline. SAFETY behaves very similarly to COPY, achieving precision close to COPY across settings, while producing outputs of similar length. JUNIOR improves recall relative to other methods, but consistently falls short of LINES while producing longer and less precise responses.

Extraction behaviour analysis. Appendix Table A4 provides qualitative insight into how different strategies work. For example, for Pericarditis Q10, the reference answer specifies three key elements: chest radiograph, transthoracic echocardiogram, and cardiology input. Both LINES and COPY recover all elements, while Safety keeps only the minimum subset.

For Q11 Pericarditis, LINES most closely matches the reference answer with all key elements. COPY also achieves high coverage but introduces minor omission of the cardiology follow-up. In contrast, the Safety strategy produces a more concise output, prioritising core treatment steps but omits important supporting details, notably PPI co-prescription, GI side-effect management, and cardiology follow-up. These trends are consistent across both GPT-OSS and Qwen models.

Qualitative analysis of JuniorDoc (Appendix Table A5) shows that Agent-2 effectively consolidates key guideline elements. However, additions of non-essential contextual information and omission of relevant details are frequent. For example, in AF, Agent-2 captures the core decision rule for cardioversion (onset <48 h with no precipitant), but introduces additional context (broader rhythm control pathways).

Performance of extraction strategies for longer documents and frontier models. The relationship between medical-term recall and answer length reveals a natural trade-off between completeness and conciseness across datasets and models (Figures 5a and 5b. Detailed results are reported in Appendix Tables A6 and A7). No single extraction strategy dominates across all settings. COPY generally produces concise, precise answers, whereas LINES and JUNIORDOC prioritise completeness, although their efficiency varies by document length and model.

The Somerset guidelines present a more challenging extraction setting, with medical-term recall remaining below 0.57 across all configurations (Figure 5a). LINES achieves the highest medical-term recall for GPT-OSS-20B (0.5670), GPT-OSS-20B with Qwen3 section retrieval (0.5529), and Qwen2.5-32B (0.5167), with average answer lengths of 83–94 words. For section retrieval, Qwen3 embeddings rank sections, which are retained until a cumulative similarity probability of 0.95 (computed with normalised similarity scores) is reached. This reduces the input context approximately threefold and is expected to reduce inference cost and latency, with only a modest reduction in medical-term recall from 0.5670 to 0.5529.

For the substantially longer NICE guidelines (Figure 5b), JUNIORDOC performs best with Qwen3-235B and DeepSeek-V3, achieving medical-term recall of 0.9321 and 0.9702 with average answer lengths of approximately 86 and 125 words, respectively. It also produces the highest word-level recall for both models. The performance of JUNIORDOC on the full NICE guidelines may reflect its two-stage design. Agent 1 first produces a clinically oriented draft identifying the information required, and agent 2 then retrieves the corresponding guideline evidence. This intermediate representation may help frontier models locate relevant evidence in very long contexts. However, its lower precision and F1 scores (Appendix Table A7) show that it can also retrieve content beyond the reference answer. JUNIORDOC is therefore most advantageous when completeness

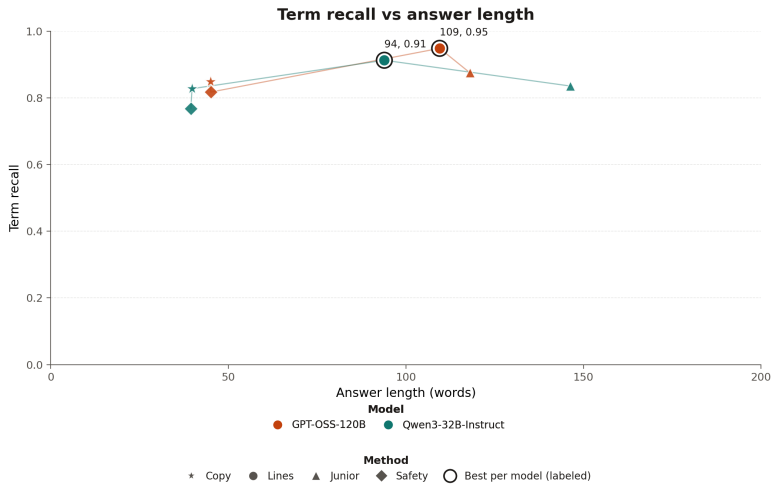


Figure 4. Medical-term recall versus answer length for the UCLH guidelines. Each point represents a prompting strategy applied to a model, with colour denoting the model and marker shape denoting the extraction strategy. Answer length is the mean number of words per response, and medical-term recall measures the proportion of reference medical terms recovered. Circled and labelled points indicate the strategy achieving the highest medical-term recall for each model.

is prioritised.

For GPT-5-mini, LINES provides the stronger recall–length balance, achieving medical-term recall of 0.9773 and word-level recall of 0.9506 in approximately 164 words. By comparison, JUNIORDOC achieves medical-term recall of 0.9585 in approximately 220 words. The benefit of JUNIORDOC is therefore model-dependent.

SAFETY-LINES broadly preserves the performance of SAFETY, slightly increasing medical-term recall for GPT-5-mini and Qwen3-235B but reducing precision and medical-term F1. This line-based variant primarily offers computational efficiency.

Overall, JUNIORDOC is most useful for unfiltered, very long guidelines with larger frontier models, whereas LINES provides more consistent performance for shorter guidelines and two OpenAI models.

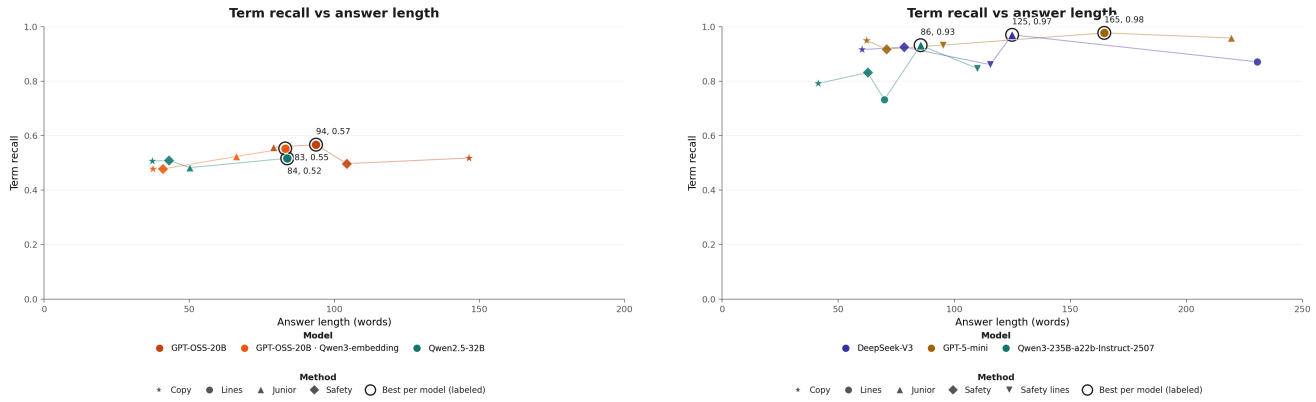
Human Evaluation Following strong automatic performance, we conducted a detailed human evaluation of model outputs generated by the GPT-OSS-120B model across 96 UCLH clinical questions, comparing both the LINES and Safety strategies (Figures 6 and 7). The Safety strategy was additionally included as a controlled contrast in output length. This contrast provides a useful lens for assessing whether metric gains reflect true clinical quality. Clinical review was performed by three clinicians (FJ, EG, NS), assessing both relevance and completeness. Relevance was rated as *irrelevant*, *relevant*, or *very relevant*, reflecting appropriateness to the clinical scenario. Completeness captured whether clinically important information was omitted or whether non-essential content was included.

Relevance. Both strategies produced highly clinically appropriate outputs. For LINES, 69.8% of responses were rated as *very relevant* and 26.0% as *relevant*, with only 4.2% classified as *irrelevant* (4 answers). Safety achieved even higher top-tier relevance, with 85.4% of outputs rated as *very relevant* and 8.3% as *relevant*, though a slightly higher proportion (6.2%) were judged *irrelevant*.

Completeness. Differences appear more clearly in completeness. LINES produced the highest proportion of *just right* answers (62.5%), with relatively few omissions (8.3% minor, 0% major), but more frequent additions (22.9% minor, 6.2% major), reflecting its tendency to include additional contextual detail. In contrast, Safety produced fewer additions (12.5% minor, 1.0% major) and a lower proportion of excessive content overall, but at the cost of increased omissions (18.8% minor, 16.7% major). Only 51.0% of Safety outputs were rated as *just right*.

Overall, LINES prioritises completeness and coverage, rarely omitting critical information but sometimes including non-essential content. Safety enforces brevity and reduces over-generation, but more frequently omits clinically relevant details, including in some cases information that may be important for safe decision-making.

Illustrative examples. The pancreatitis example (see Appendix Table A8) demonstrates a minor addition for LINES. The response appropriately prioritises urgent referral to ICU outreach, including clear clinical criteria to support severity assessment and escalation decisions. It also provides relevant contact pathways, including the general medical registrar, ICU outreach services, hepatobiliary specialists, and dietetics, all of which are appropriate for ongoing care. The inclusion of additional sections on ongoing management and imaging is not directly required to answer the referral question. However, this information



(a) Somerset guidelines

(b) NICE guidelines

Figure 5. Medical-term recall versus mean answer length for extraction from longer clinical guidelines. Colours denote models and marker shapes denote prompting strategies. Medical-term recall is the proportion of medical terms in the reference answer recovered by the extracted response.

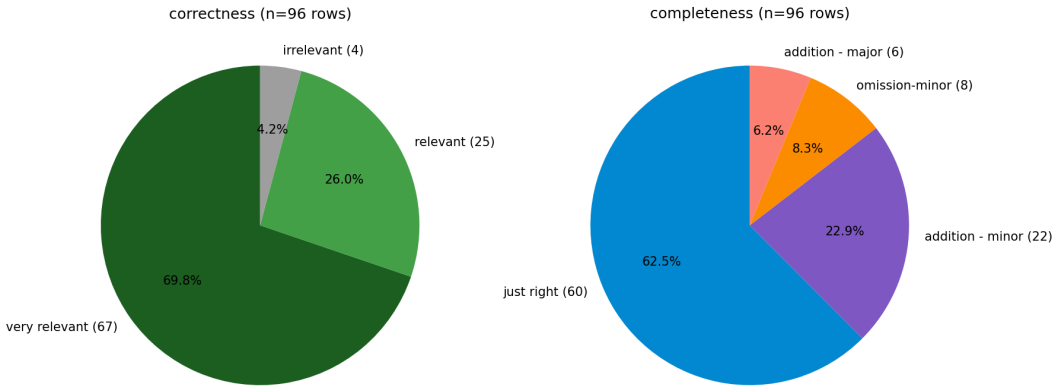


Figure 6. Clinician evaluation of LINES outputs. Distribution of relevance (left) and completeness (right) ratings across 96 clinical questions. LINES achieves high clinical relevance (95.8% at least relevant) and the highest proportion of "just right" answers (62.5%), indicating strong alignment with source guidance. No major omissions were observed, suggesting that LINES reliably preserves critical clinical content.

remains clinically relevant and does not detract from decision-making.

The ureteric colic example illustrates a minor omission. The response provides a clear and clinically useful set of discharge criteria that are easy to interpret and apply. One element of the guideline, relating to stone size and anatomical location, is not explicitly included and would require reference to an associated table. This omission does not significantly affect clinical decision-making, as it can be readily clarified if needed and does not compromise patient safety.

Inter-annotator agreement. Inter-annotator agreement was assessed on a subset of 12 shared questions (Pericarditis, Acute Kidney Injury, and Atrial Fibrillation). For relevance, unanimous agreement was observed in 3 cases, with the remaining 9 showing agreement between two annotators. Disagreements were limited to adjacent categories (e.g. *relevant* vs *very relevant*). For completeness, full agreement was observed in 6 cases, with the remaining 6 showing partial agreement. Disagreements again occurred primarily between neighbouring categories, most commonly *just right* and *addition–minor*. Majority voting was used to assign final labels.

These findings confirm that the LINES strategy provides a strong mechanism for clinically evidence extraction. The consistently high relevance (95.8%) and absence of major omissions indicate that LINES effectively preserves critical safety-relevant information. Minor omissions tend to affect peripheral details, confirming that LINES avoids unsafe under-specification.

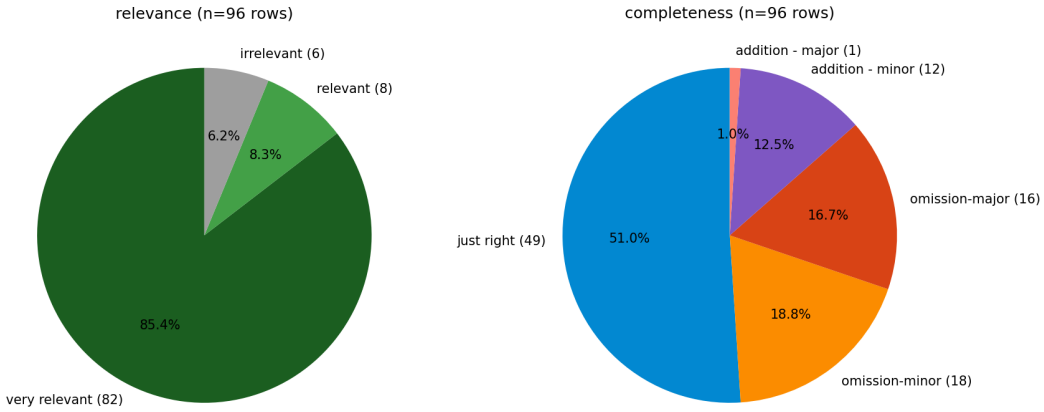


Figure 7. Clinician evaluation of Safety outputs. Distribution of relevance (left) and completeness (right) ratings across 96 clinical questions. Safety produces highly relevant outputs (93.7% at least relevant), with a higher proportion of "very relevant" ratings compared to LINES, reflecting concise and focused responses. The strategy presents increased omissions (35.5%, including 16.7% major omissions).

Discussion

Statement of key findings Our results show that prompting strategy effectiveness depends on document length, model and retrieval setting. For shorter protocol-like guidelines, LINES consistently achieves the highest medical-term recall, improving it by approximately 9 percentage points over COPY on average, although with longer answers. The restricted search space appears to support effective line-level evidence selection.

Document length alone, however, does not determine the optimal strategy. For the longer Somerset guidelines, LINES still achieves the highest medical-term recall across all model configurations. For the substantially longer NICE guidelines, JUNIORDOC performs best with Qwen3-235B and DeepSeek-V3, increasing medical-term recall over COPY by 14.0 and 5.4 percentage points, respectively. GPT-5-mini remains better suited to LINES, indicating that the benefit of draft-first retrieval tends to be model-dependent.

Overall, the results reveal a trade-off between completeness and precision. LINES and JUNIORDOC generally favour recall, whereas COPY and SAFETY produce shorter, more precise answers but may omit relevant information. Returning line numbers can reduce API-generated tokens and guarantee verbatim reconstruction, but does not itself improve extraction: SAFETY-LINES broadly retains Safety performance. Thus, direct line extraction is preferable for shorter or retrieval-constrained contexts, while sentence-based JUNIORDOC can improve completeness in long guidelines when applied to suitable frontier models.

Strengths and limitations This study introduces an approach to clinical question answering in which LLMs are constrained to operate as evidence extractors rather than free-text generators. By enforcing reproduction of guideline content, the approach directly addresses one of the key risks of LLM deployment in healthcare, the risk of hallucination, while enabling transparent and auditable outputs.

A further strength is the systematic comparison of prompting strategies across different document-length regimes. This reveals that performance is not solely a function of model capacity, but of the interaction between prompting strategy and input structure. In particular, the results demonstrate that increasing architectural complexity (e.g. multi-agent pipelines) may not always be beneficial.

However, several limitations should be mentioned. First, human evaluation was conducted on a relatively small set of clinical questions, and although designed to reflect realistic scenarios, larger-scale validation is required. Second, the study focuses on text-based guidelines and does not include common tabular or visual elements (e.g. flowcharts). Extending the approach to multimodal guideline formats remains an important direction.

Third, evaluation was performed in a simulated setting rather than in real-time clinical workflows. While this enables controlled comparison, it does not fully capture real-world usage constraints. Finally, although we evaluated several models, results may further vary with different architectures and context lengths. Nevertheless, the consistency of observed patterns across models suggests that the findings are primarily driven by prompting strategy rather than model-specific effects.

Finally, Somerset reference answers were freely written by clinicians, whereas UCLH references were selected directly from guideline lines. Consequently, lower lexical and term overlap on Somerset may partly reflect differences in annotation style rather than extraction difficulty alone.

Implications and next steps These findings have important implications for the deployment of LLMs in clinical settings. First, they suggest that retrieval and extraction should be jointly designed. In our experiments, draft-based approaches such as JUNIOR showed some advantage over longer contexts, whereas extraction methods such as LINES performed comparably over shorter ones.

Second, the results emphasise the need to explicitly manage the precision–recall trade-off. In safety-critical applications, prioritising recall may be necessary to avoid omission of clinically relevant information, even at the cost of longer outputs. Future systems may benefit from adaptive strategies that dynamically adjust verbosity based on task complexity.

Future work will focus on real-world deployment, including integration with electronic health record systems and evaluation in live clinical environments. In addition, interactive capabilities, such as clarifying missing information, may enhance both completeness and relevance.

Conclusion

This study demonstrates that LLM-based systems can provide reliable access to guideline information when constrained to evidence extraction. The optimal strategy depends on the retrieval setting: line-number extraction is particularly effective for shorter, protocol-like documents, whereas draft-first extraction can improve completeness in very long documents when used with suitable models.

Beyond reducing hallucination, our findings emphasise that preserving completeness is essential for patient safety. By combining extraction, drafting and retrieval techniques, LLMs can move towards becoming trustworthy clinical assistants.

Code Availability

The code used to reproduce the experiments and evaluation pipeline is publicly available at https://github.com/julia-ive/safeLLM-extraction.

Data Availability

The clinical guideline data used in this study are proprietary and cannot be publicly shared due to confidentiality requirements. NICE Q&A dataset is publicly available online: https://github.com/julia-ive/guidelines_qa

Ethics approval

This study was approved by the Digital Research and Innovation Board (DRB), the governance and oversight body for the Advanced Research Computing Centre at University College London (UCL ARC), under reference TRE-467283. It was also approved as a clinical service audit at the National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust (reference: 130-202425-SE).

No patients were involved, and the guideline data contained no patient-level or personally identifiable information. All materials were derived from public or institutional clinical guidelines intended for professional use. All study procedures were conducted in accordance with relevant institutional guidelines and regulations.

Consent to participate

Not applicable. This study involved no patients or patient data. All annotations were performed by co-authors acting as investigators, not as recruited research participants.

Consent for publication

Not Applicable.

Competing interests

We have no competing interests to declare.

Funding

Not applicable

Author Contributions Statement

JI: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – Final draft preparation, Writing - Reviewing and Editing. FJ: Conceptualisation, Methodology, Validation, Formal Analysis, Investigation, Writing – Early draft preparation, Writing - Reviewing and Editing. EG, NS: Conceptualisation, Methodology, Formal Analysis, Investigation, Writing – Early draft preparation, Writing - Reviewing and Editing. NJ: Conceptualisation, Writing - Reviewing and Editing. CSH, PB, EC: Writing - Reviewing and Editing. RD: Conceptualisation, Methodology, Writing - Reviewing and Editing. All authors approved the manuscript.

References

1. Guerra-Farfan, E. *et al.* Clinical practice guidelines: The good, the bad, and the ugly. *Injury* **54**, S26–S29 (2023).
2. Scraggs, E. *et al.* Factors that encourage or discourage doctors from acting in accordance with good practice final report prepared for the general medical council. Tech. Rep. (2012).
3. Kresevic, S. *et al.* Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit. Med.* **7**, 102 (2024).
4. Dean, N. C. *et al.* Decreased mortality after implementation of a treatment guideline for community-acquired pneumonia. *The Am. journal medicine* **110**, 451–457 (2001).
5. VanWagner, L. B. *et al.* Blood pressure control according to clinical practice guidelines is associated with decreased mortality and cardiovascular events among liver transplant recipients. *Am. J. Transplantation* **20**, 797–807 (2020).
6. Arnold, F. W. *et al.* Improving outcomes in elderly patients with community-acquired pneumonia by adhering to national guidelines: Community-acquired pneumonia organization international cohort study results. *Arch. internal medicine* **169**, 1515–1524 (2009).
7. Monti, G. *et al.* Effect of a quality improvement program on compliance to the sepsis bundle in non-icu patients: a multicenter prospective before and after cohort study. *Front. Medicine* **10**, 1215341 (2023).
8. Coba, V. *et al.* Resuscitation bundle compliance in severe sepsis and septic shock: improves survival, is better late than never. *J. intensive care medicine* **26**, 304–313 (2011).
9. Qumseya, B. *et al.* Barriers to clinical practice guideline implementation among physicians: a physician survey. *Int. J. Gen. Medicine* 7591–7598 (2021).
10. Tsiga, E., Panagopoulou, E., Sevdalis, N., Montgomery, A. & Benos, A. The influence of time pressure on adherence to guidelines in primary care: an experimental study. *BMJ open* **3**, e002700 (2013).
11. Freedman, S., Golberstein, E., Huang, T.-Y., Satin, D. J. & Smith, L. B. Docs with their eyes on the clock? the effect of time pressures on primary care productivity. *J. health economics* **77**, 102442 (2021).
12. Van der Vegt, A., Zuccon, G., Koopman, B. & Deacon, A. How searching under time pressure impacts clinical decision making. *J. Med. Libr. Assoc. JMLA* **108**, 564 (2020).
13. Torjesen, I. Doctors are under more work pressure than during height of covid-19 pandemic in 2020. *BMJ: Br. Med. J. (Online)* **375** (2021).
14. Daniels, J., Robinson, E., Jenkinson, E. & Carlton, E. Perceived barriers and opportunities to improve working conditions and staff retention in emergency departments: a qualitative study. *Emerg. Medicine J.* **41**, 257–265 (2024).
15. Kann, B. H., Johnson, S. B., Aerts, H. J., Mak, R. H. & Nguyen, P. L. Changes in length and complexity of clinical practice guidelines in oncology, 1996-2019. *JAMA Netw. Open* **3**, e200841–e200841 (2020).
16. Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
17. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 314–334 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2023).
18. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv [cs.CL]* (2020).
19. Ng, K. K. Y., Matsuba, I. & Zhang, P. C. RAG in health care: A novel framework for improving communication and decision-making by addressing LLM limitations. *NEJM AI* **2** (2025).
20. Lv, A., Xie, R., Sun, X., Kang, Z. & Yan, R. Language models "grok" to copy. *arXiv [cs.CL]* (2024).

21. Chen, Q. *et al.* Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat. Commun.* **16**, 3280 (2025).

22. Lewis, M. *et al.* Grounding large language models in clinical evidence: A retrieval-augmented generation system for querying UK NICE clinical guidelines. *arXiv [cs.CL]* (2025).

23. Ding, Q., Zhang, E. H. Q., Jozsa, F. & Ive, J. Building a silver-standard dataset from NICE guidelines for clinical LLMs. *arXiv [cs.CL]* (2025).

24. OpenAI *et al.* gpt-oss-120b & gpt-oss-20b model card. *arXiv [cs.CL]* (2025).

25. Yang, A. *et al.* Qwen3 technical report. *arXiv [cs.CL]* (2025).

26. Gemma Team *et al.* Gemma: Open models based on gemini research and technology. *arXiv [cs.CL]* (2024).

27. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (2004).

28. Kraljevic, Z. *et al.* Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.* **117**, 102083, DOI: 10.1016/j.artmed.2021.102083 (2021).