The Verbose Context Problem in Medical Records

Anonymous Authors[1]

Abstract

The verbose context problem occurs when structured concepts have token-inefficient textual representations. This bottleneck is acute in population health: cohort-level analysis of longitudinal patient records requires reasoning over thousands of medically-coded events, often exceeding 400K tokens. We present PopMedQA, a benchmark isolating this problem through computational tasks on groups of longitudinal patient records. We construct the benchmark using `neopatient`, a new library for language-controlled generation of artificial patient records. Through extensive ablations—including prompting strategies, prompt compression, and agentic decomposition—we find that domain-independent methods fail to alleviate the verbose context problem. There remains significant opportunity to exploit domain-specific structure in language model inputs for population-scale reasoning.

1. Introduction

Population health analytics focuses on identifying patterns, detecting anomalies, and quantifying disease burden across large groups of individuals. While large language models (LLMs) offer a flexible alternative to traditional rule-based risk adjustment systems, their application is hindered by the *verbose context problem*. This problem arises when structured concepts—such as medical codes in electronic health records (EHRs)—have token-inefficient textual representations that inflate context lengths. For example, "ICD-10 I21: Acute myocardial infarction" is a textual representation $\mathrm{str}(c)$ of the underlying concept c of a heart attack. The usual process is to tokenize the string, lookup the token embeddings, and process the resulting sequence of vectors by the model f. In this notation, the language model's output is: $f \circ \mathrm{emb} \circ \mathrm{tok} \circ \mathrm{str}(c) := y$. When such concepts pervade the context, as they do in longitudinal patient records, the total context length N becomes too long for practical reasoning.

We capture the verbose context problem in a new benchmark called PopMedQA. Each of its questions involves reasoning over groups of 10-50 synthetic longitudinal patient records, which typically amount to 64K-256K tokens in a textual representation. As described in Section 2, it differs from existing long-context benchmarks in two ways. First, it specifically isolates how verbosity, rather than the presence of irrelevant "hay" context, affects performance. Second, it involves multi-hop reasoning over population-scale cohorts that cannot be decomposed into questions about individuals. Constructed with clinician and expert review, PopMedQA reflects the real-world priorities of population health, such as identifying latent clinical clusters or detecting sparse anomalies across disparate patient trajectories.

To construct the large volume of synthetic data required for PopMedQA, we introduce `neopatient`, a new software library for language-controlled generation of artificial patient records. Unlike rule-based generators (like Synthea (Walonoski et al., 2018)), `neopatient` trajectories are controlled through natural language descriptions, allowing for the creation of complex clinical cohorts without custom simulation code. Details on `neopatient` are provided in Appendix B and Related Work in Appendix A.

In Section 4, we conduct a thorough evaluation of a range of language models on PopMedQA. These confirm the claimed design characteristics of PopMedQA, such as decomposition resistance. For ablations, we conduct a meta-analysis of multiple families of techniques for improving long-context performance, including prompting strategies, prompt compression, and agentic decomposition.

Our analysis reveals several generalizable insights on the nature of population-scale EHR reasoning: (1) both clinical competence and long-context capability are required; (2) generic prompt compression is fragile; (3) medical pretraining does not substantially improve performance; and (4) agentic decomposition is ineffective and/or cost-prohibitive. Overall, we find that domain-independent methods fail to alleviate the verbose context problem, exposing a significant unrealized opportunity to exploit domain structure for population-scale reasoning.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. The Verbose Context Problem

We consider heterogeneous prompts $x : \text{list}(\text{string} + \mathcal{C})$ for LLMs. $\mathcal{C}$ is a set of concepts distinguished from the other unstructured parts of the prompt. As the canonical example in this paper, we take $\mathcal{C}$ to be the set of all medical codes. The verbose context problem arises when converting $c \in \mathcal{C}$ to strings. Let $\text{str}(c)$ be a string which conveys c to a target language model with the desired level of precision. For a medical code, it may not be sufficient to pass just the abbreviated code; instead, much of the extended description may be needed as well. To be processed by the language model, the string must be tokenized, and then each token's embedding is looked up in the embedding matrix. This produces a variable-length sequence of vectors $\text{emb} \circ \text{tok} \circ \text{str}(c) \in \text{list}\ \mathcal{E}$, where $\mathcal{E}$ is the LLM embedding space, which is typically a few thousand dimensions. Given a probability distribution over prompts x, the baseline (overall) expected context length is N. M is the context length that originates from $\mathcal{C}$.

$$M = \mathbb{E}_x \sum_{c \in x \cap \mathcal{C}} \text{Length}(\text{emb} \circ \text{tok} \circ \text{str}(c))$$
$$N = \mathbb{E}_x\ \text{Length}(\text{emb} \circ \text{tok} \circ \text{str}(x))$$

(In the second line, we slightly abuse notation to have $\text{str}(x)$ operate on all parts of the prompt). Informally, the verbose context problem is that N is too long to be practical.

Definition 2.1 (Verbose Context Problem)**.** This occurs when M/N is large, i.e. $\Omega(1)$ as $N \to \infty$.

3. PopMedQA

To embody the verbose context problem, we present a new benchmark. PopMedQA consists of questions about groups (of size between 10 and 50) of longitudinal patient records. These records are drawn from over 25 thousand synthetic patients generated specifically for the benchmark. These records comprise over 14M medically-coded events. When a typical code is conveyed as text, it uses between 8 and 20 tokens. A code accompanies each timestamped event. There are typically thousands of such events in each record. This makes most of PopMedQA's questions approximately 64K, 128K, or 256K tokens in length, when represented as text.

An important design feature of PopMedQA is *decomposition resistance*. Long context can often be circumvented by partitioning it into separate chunks, process the chunks separately, and aggregating an answer from multiple rounds of processing. The purpose of PopMedQA is to more specifically reward verbose-context capability rather than this more routine context engineering. Accordingly, most of PopMedQA's questions are designed so that they cannot be readily solved by asking a series of individual patient-level questions. Each question poses one of nine computational tasks (see Figure 7 and Appendix C). Some of these, such as planted clique and clustering, are especially challenging without holistic in-context processing. We see quantitative evidence of such decomposition resistance in our experimental results, presented in Section 4.

PopMedQA presents realistic questions from population health. Whereas clinical medicine and biomedicine focus on interventions upon individual patients, population health concerns questions about groups of patients. PopMedQA's questions were reviewed by both clinicians and population health experts for pertinence and validity.

3.1. PopMedQA Pipeline

Since the questions in PopMedQA are very long, they are generally not possible for humans (or computers) to solve or verify. Thus, a correct pipeline cannot generate questions and then (independently) generate labels. To ensure answers are correct, the questions and the data involved in the answers must be jointly generated. The overall pipeline is as follows.

1. Task Definition This involves specifying the schema of the answer (e.g. a list of lists of patient IDs) as well as the scoring function between the answer and the truth (see Appendix C for details on each task's metric). It also involves specifying the cohorts that should be generated so that the answer can be synthesized from them. For example, planted clique is the task of finding the most cohesive or similar size-k subset of patients. The task indicates two cohorts should be created: the size-k clique, and the $N - k$ remaining patients.

2. Question Ideation Given a task, abstract question ideas (or template) are generated. The idea does not have a concrete question statement, nor does it have individual patients generated. Instead, the idea has a question template parameterized by N, e.g. "among these N ED and urgent care records from the last 72 hours in our metro area, is there a subset of patients presenting with an unusual combination of symptoms...". The idea also gives question-specific descriptions of the cohorts that need to be generated. For example:

1. Syndromic subset of size $\min(12, \text{int}(N/4))$: a subset of patients from recent ED and urgent care records in a specific zip code over the last 72 hours. All have a documented chief complaint that includes a combination of 'severe gastrointestinal distress (vomiting/diarrhea)' AND 'unusual rash'.

2. Other population of size $N - \min(12, \text{int}(N/4))$: all other ED and urgent care records from the last 72 hours,

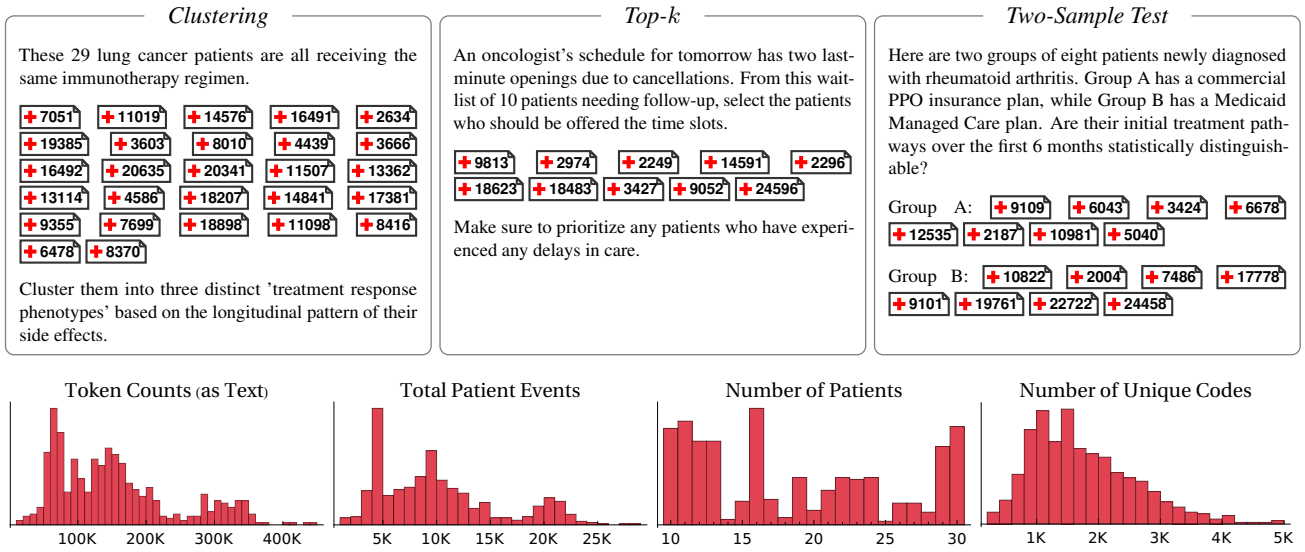

Figure 1. Example questions from PopMedQA (above), as well as statistics of all its questions (below). Within the questions, each patient record is visually depicted as a boxed patient ID number. Since these records may contains thousands of coded events, and codes have verbose textual representations, PopMedQA is a long-context benchmark.

showing typical presentations like chest pain, respiratory infections, and minor trauma, without the combination of severe GI distress and rash.

3. Question Sampling For each question idea, cohorts are generated at a large size $N_{\max}$. To create concrete questions of size N, the idea's cohorts are subsampled at the specified value of N.

4. Clinician Validation To ensure the quality and practical relevance of the benchmark, questions undergo a clinician validation step. Clinician reviewers (M.D.s) score each question from 1–10 on three metrics: (1) *Realism* (how likely or commonly the questions would arise in practice), (2) *Difficulty* (of solving them by hand), and (3) *Coherence* (whether the cohorts are well-defined and distinct). Questions are retained in PopMedQA only if they score at least 5 on all three metrics.

4. Experiments

4.1. Setup

We conduct a thorough evaluation of a range of language models on PopMedQA, testing models that range from 7B parameters to frontier scale, with context lengths from 128K to 2M tokens. On top of baseline models, we examine four families of interventions or ablations designed to improve long-context performance:

Prompting strategies We consider three ways of representing patient records as text: (1) a baseline method that replaces codes with truncated descriptions, (2) a naive method using only raw codes, and (3) a "codebook" method that maps unique codes to IDs to reduce redundancy (see Figures 5 and 6).

Prompt compression We evaluate two methods: rendering text to images using the Glyph pipeline (Cheng et al., 2025) and LLMLingua-2 text chunk compression using the `microsoft/llmlingua-2-xlm-roberta-large-meetingbank` model (Jiang et al., 2023).

Reasoning We expand inference-time compute using chain-of-thought prompting (Wei et al., 2022). "Think" models were run with default temperatures and no upper bound on answer length, while non-reasoning models were run with temperature 0 and a 2048 token limit.

Multi-turn interaction We utilize agentic context engineering and decomposition via systems like MARS, LongCEPO, and Claude Code. These systems were allowed to perform multiple queries to solve a single question, with Claude Code having a $5 USD limit per answer.

4.2. Results and Discussion

The primary experimental results on PopMedQA are presented in Figures 2 and 3. Overall, the leaderboard results align with expectations, as frontier models demonstrate superior performance, followed by large open-source models with strong long-context capabilities. The tasks in PopMedQA effectively stress frontier-level capabilities,

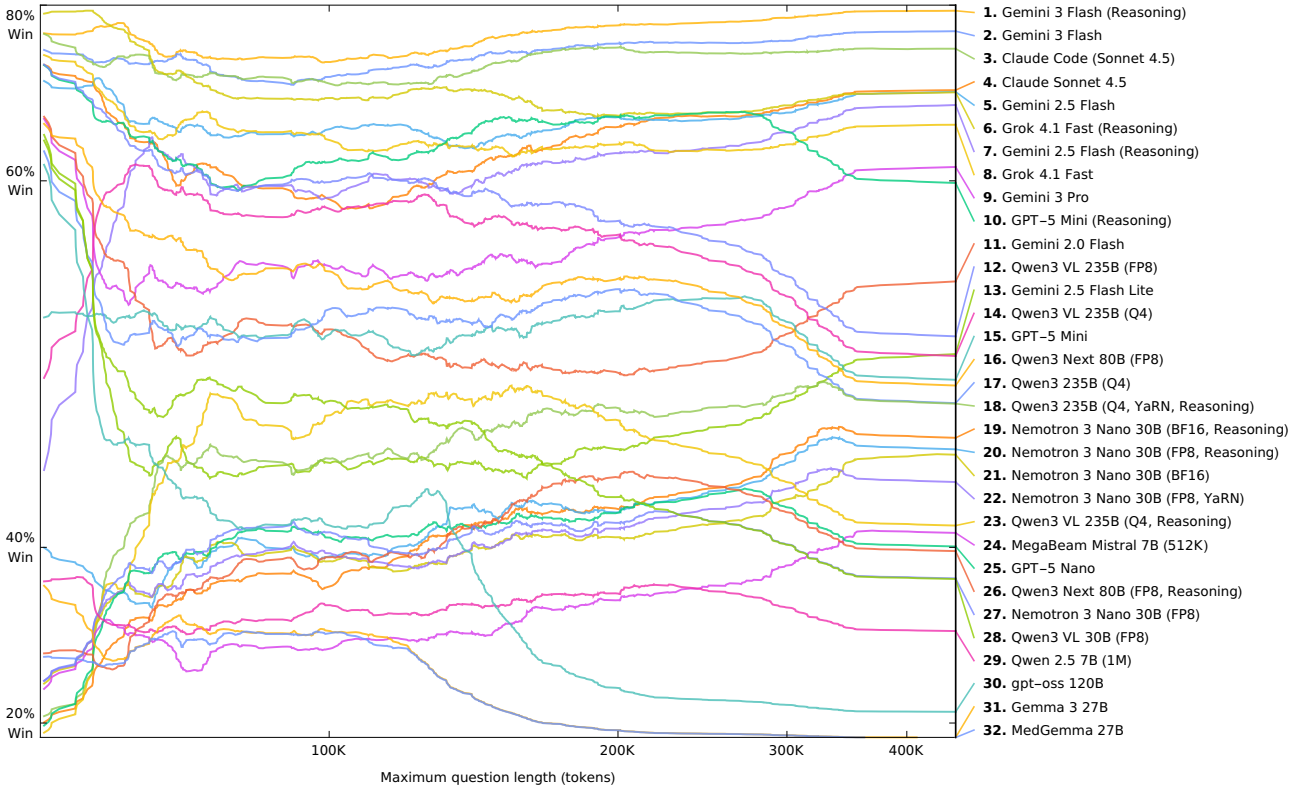

Figure 2. Leaderboard performance of models across all tasks on PopMedQA. The y-axis is the percentage of comparisons the model won against all other models. The x-axis restricts these comparisons to questions up to a given token length. We observe that performance stresses frontier-level capabilities.

with performance declining as context length increases. Detailed task-specific scoreboards are provided in Figure 4.

Clinical competence and long-context capability are required. Maintaining a high rank at increased context lengths is not guaranteed. While models like Gemini 3 Flash are consistent, smaller models like Nemotron 3 Nano only rise in rank as the context expands, indicating that both medical domain knowledge and long-context processing are essential for PopMedQA.

Generic prompt compression is fragile. Compressed prompts, such as those generated by rendering text to images, significantly degrade model performance, particularly in instruction following. Domain-independent compression methods appear to strip away critical information needed for complex medical reasoning.

Medical pretraining does not substantially improve performance. General-purpose frontier models often outperformed specialized medical models on PopMedQA. This suggests that the primary challenge is not a lack of clinical knowledge but rather the ability to perform robust, multi-hop reasoning over the verbose contexts characteristic of longitudinal EHR data.

Agentic (multi-turn) decomposition is ineffective and/or cost-prohibitive. While agentic systems showed relative strength, they failed to achieve absolute performance gains that justify their high computational and financial costs. Systems like MARS and LongCEPO frequently harmed instruction-following and absolute performance compared to monolithic baselines. Furthermore, Claude Code did not surpass the efficiency of frontier monolithic models. This confirms that PopMedQA's tasks are resistant to simple decomposition and require holistic in-context reasoning.

5. Conclusion

The verbose context problem inhibits population-level EHR reasoning. Our results demonstrate that domain-independent methods—including prompt compression and agentic decomposition—fail to alleviate performance degradation. These findings reveal a significant unrealized opportunity to exploit domain-specific structure in language model inputs to enable robust population-scale reasoning.

References

Arnrich, B., Choi, E., Fries, J., McDermott, M., Oh, J., Pollard, T., Shah, N., Steinberg, E., Wornow, M., and van de Water, R. Medical event data standard (meds): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series For Health (TS4H)*, 2024. URL https://openreview.net/forum?id=IsHy2ebjIG.

Artificial Analysis. Artificial analysis long context reasoning benchmark leaderboard (aa-lcr). Online leaderboard https://artificialanalysis.ai/evaluations/artificial-analysis-long-context-reasoning, 2025. Independent benchmark evaluating language models' ability to extract, reason about, and synthesize information from long-form documents (10k–100k tokens, cl100k_base tokenizer). Continuously updated as of February 2026. Part of the Artificial Analysis Intelligence Index.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

Cheng, J., Liu, Y., Zhang, X., Fei, Y., Hong, W., Lyu, R., Wang, W., Su, Z., Gu, X., Liu, X., et al. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*, 2025.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. Comorbidity measures for use with administrative data. *Medical care*, pp. 8–27, 1998.

Eyuboglu, S., Ehrlich, R., Arora, S., Guha, N., Zinsley, D., Liu, E., Tennien, W., Rudra, A., Zou, J., Mirhoseini, A., et al. Cartridges: Lightweight and general-purpose long context representations via self-study. *arXiv preprint arXiv:2506.06266*, 2025.

Fleming, S. L., Lozano, A., Haberkorn, W. J., Jindal, J. A., Reis, E., Thapa, R., Blankemeier, L., Genkins, J. Z., Steinberg, E., Nayak, A., Patel, B., Chiang, C.-C., Callahan, A., Huo, Z., Gatidis, S., Adams, S., Fayanju, O., Shah, S. J., Savage, T., Goh, E., Chaudhari, A. S., Aghaeepour, N., Sharp, C., Pfeffer, M. A., Liang, P., Chen, J. H., Morse, K. E., Brunskill, E. P., Fries, J. A., and Shah, N. H. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22021–22030, 2024. doi: 10.1609/aaai.v38i20.30205.

Grolleau, F., Alsentzer, E., Keyes, T., Chung, P., Swaminathan, A., Aali, A., others, and Chen, J. H. Medfacteval and medagentbrief: A framework and workflow for generating and evaluating factual clinical summaries. In *Pacific Symposium on Biocomputing*, volume 31, pp. 388–399, 2026.

Huang, S.-C., Huo, Z., Steinberg, E., Chiang, C.-C., Langlotz, C., Lungren, M., Yeung, S., Shah, N., and Fries, J. Inspect: A multimodal dataset for patient outcome prediction of pulmonary embolisms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74141–74163, 2023.

Jahanian, A., Chai, L., and Isola, P. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HylsTT4FvB.

Jiang, H., Wu, Q., Lin, C.-Y., Yang, Y., and Qiu, L. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023.

Jiang, Y., Black, K. C., Geng, G., Park, D., Zou, J., Ng, A. Y., and Chen, J. H. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144.

Kindig, D. and Stoddart, G. What is population health? *American journal of public health*, 93(3):380–383, 2003.

Lee, G., Hwang, H., Bae, S., Kwon, Y., Shin, W., Yang, S., Seo, M., Kim, J.-Y., and Choi, E. Ehrsql: A practical text-to-sql benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15589–15601, 2022.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

OpenAI. Graphwalks: Multi-hop reasoning long-context benchmark. Introduced in https://openai.com/index/gpt-4-1, April 2025a. Dataset available at https://huggingface.co/datasets/openai/graphwalks (MIT licensed); evaluates breadth-first search and parent retrieval in large directed graphs.

OpenAI. Openai-mrcr: Multi-round coreference resolution benchmark. Introduced in https://openai.com/index/gpt-4-1, April 2025b. Dataset available at https://huggingface.co/datasets/openai/mrcr.

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., Ingber, M. J., Levy, J. M., and Robst, J. Risk adjustment of medicare payments using the cms-hcc model. *Health Care Financing Review*, 25(4):119, 2004.

Subramani, N., Suresh, N., and Peters, M. E. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Vodrahalli, K., Ontanon, S., Tripuraneni, N., Xu, K., Jain, S., Shivanna, R., Hui, J., Dikkala, N., Kazemi, M., Fatemi, B., et al. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*, 2024.

Walonoski, J., Klaus, M., Granger, B., Hall, D., Gregor, C., Neyarapally, T., Watson, A., and Scanlon, J. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.

Wei, H., Sun, Y., and Li, Y. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Weiner, M. G. et al. The johns hopkins adjusted clinical group (acg) system: A guide for researchers. *Health Services Research*, 2012.

Wornow, M., Thapa, R., Steinberg, E., Fries, J., and Shah, N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Yen, H., Gao, T., Hou, M., Ding, K., Fleischer, D., Izsak, P., Wasserblat, M., and Chen, D. Helmet: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhang, X., Chen, Y., Hu, S., Xu, Z., Chen, J., Hao, M., Han, X., Thai, Z., Wang, S., Liu, Z., et al. ∞-bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024.

Zheng, M., Feng, X., Si, Q., She, Q., Lin, Z., Jiang, W., and Wang, W. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9102–9124, 2024.

A. Related Work

Long-Context Benchmarks. Since the advent of modern language models, dozens of long-context benchmarks have been developed (Tay et al., 2021; Bai et al., 2023; Zhang et al., 2024; Yen et al., 2025; Bai et al., 2025). Context lengths under evaluation have increased from 4K to above 1M. Different evaluations attempt to isolate different failure modes. Early diagnostic tests focus on retrieval failures and positional bias (e.g. "lost in the middle"). Newer evaluations target reasoning degradation in multi-step tasks (OpenAI, 2025a). In these benchmarks, difficulty and context length are driven primarily by (1) the amount and density of irrelevant distractors, and (2) the number of reasoning hops required to bridge dispersed information (Vodrahalli et al., 2024; OpenAI, 2025b). PopMedQA emphasizes a different cause of context bloat (verbosity) in order to expose different failure modes.

While some benchmarks isolate long-context reasoning through abstract tasks (OpenAI, 2025a), others prioritize naturalistic, real-world inquiries (Bai et al., 2023; Artificial Analysis, 2025). Our work belongs to the latter category. We encourage more situated, concrete study of long-context problems by recognizing and addressing their domain-specific aspects. We aim to further close the gap between benchmark performance and real-world utility.

AI on EHRs. Existing benchmarks for AI in Electronic Health Records (EHRs) evaluate a wide range of clinical and administrative capabilities. For structured data, EHRSHOT (Wornow et al., 2023) and INSPECT (Huang et al., 2023) assess few-shot clinical prediction and algorithmic fairness within individual patient timelines. EHRSQL (Lee et al., 2022) and MedAgentBench (Jiang et al., 2025) extend these evaluations to cohort-level queries; however, these frameworks primarily test the model's ability to translate natural language into structured queries (SQL or FHIR) that delegate computational aggregation to an external database engine. For unstructured text, MedAlign (Fleming et al., 2024) and MedFactEval (Grolleau et al., 2026) focus on instruction-following and factuality within clinical notes. PopMedQA diverges from these approaches by shifting the analytical focus to population health, requiring models to perform holistic, in-context reasoning across the raw longitudinal records of cohorts of 10 to 50 patients simultaneously. This framework unlocks complex use cases in population-level pattern discovery, such as identifying latent clinical clusters or detecting sparse anomalies across disparate patient trajectories, that cannot be readily addressed by standard query-based aggregation or individual-level processing.

Population Health Analytics. Population health analytics focuses on the health outcomes of groups of individuals and the distribution of these outcomes within the group (Kindig & Stoddart, 2003). Its primary objectives are to quantify disease burden and guide resource allocation to ensure equitable and efficient healthcare delivery. To achieve this, established risk adjustment systems like the Johns Hopkins Adjusted Clinical Group (ACG) System (Weiner et al., 2012), the CMS Hierarchical Condition Category (CMS-HCC) model (Pope et al., 2004), and comorbidity indices such as Charlson (Charlson et al., 1987) and Elixhauser (Elixhauser et al., 1998) are widely employed. These tools primarily rely on rule-based aggregation of structured diagnosis codes and pharmacy data to perform retrospective financial risk stratification and predict healthcare utilization. However, these statistical frameworks are often limited by fragile or manual feature engineering that cannot capture the complex dependencies within a patient's history, leading to low individual-level predictive accuracy (e.g., R^2 values frequently below 15% for prospective cost prediction) (Pope et al., 2004). PopMedQA evaluates a more flexible, yet still code-centric, alternative where modern language models perform comprehensive longitudinal reasoning over groups of patient records across entire cohorts.

Alternative Concept Representations. Are there more succinct ways to convey concepts to language models than language itself? Multiple lines of work support this general idea. The common strategy is to inject concepts as vectors at different model layers, in lieu of providing more input text. In prefix tuning (Li & Liang, 2021), the output embeddings of the earlier part of a prompt are truncated and directly optimized to improve the accuracy of subsequent generation. Cartridges (Eyuboglu et al., 2025) also condense the prefix by optimizing the contents of key-value caches in attention layers. Steering vectors (Jahanian et al., 2020; Subramani et al., 2022) are added to activations to control generation in an input-agnostic manner. Rendering text to images and using vision language models can be effective for long-context inference (Zheng et al., 2024; Cheng et al., 2025; Wei et al., 2025).

B. neopatient: Language-Controlled Generation of Patient Records

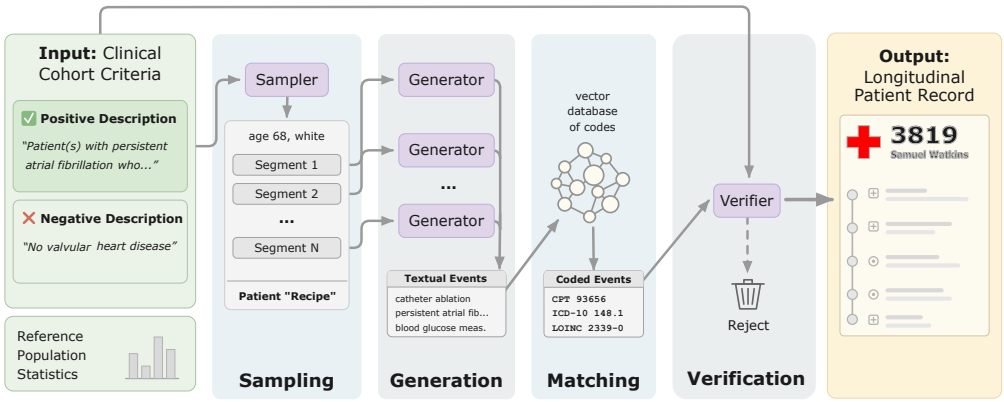

Figure 8. The neopatient architecture for language-controlled artificial patient generation. The pipeline transforms natural language criteria describing a cohort into a set of coded longitudinal records in the MEDS format. The architecture consists of four primary stages: (1) Sampling, where an LLM generates a "patient recipe" that defines demographics and divides the patient's life trajectory into discrete temporal segments to ensure coherence; (2) Generation, where longitudinal medical events are produced for each segment; (3) Matching, where a vector database maps natural language descriptions to standardized medical codes; and (4) Verification, a final correctness check where an LLM validates that the completed record strictly satisfies the original cohort specification.

To generate the large volume of synthetic data required for PopMedQA, we implemented a new software library, `neopatient`, for language-controlled generation of artificial patient records. Unlike rule-based generators (like Synthea (Walonoski et al., 2018)), patient trajectories in `neopatient` are controlled through natural language descriptions, allowing for the creation of complex clinical cohorts without the need for custom simulation code or state machines.

Generating longitudinal patient records with LLMs presents several inherent challenges. First, **output length** constraints are significant; LLM generations are typically limited to less than 64K tokens, and they often become unreliable when producing long, structured outputs at that scale. Second, **coding knowledge** is limited, as LLMs do not precisely know the vast and frequently updated medical ontologies. Third, maintaining **clinical plausibility** requires managing complex sequential dependencies, such as ensuring a prescription follows an appropriate diagnosis. Finally, the need for **batching for efficiency** when generating large cohorts of hundreds of patients restricts the use of complex, multi-turn agentic loops. The `neopatient` pipeline consists of four primary stages designed to address these challenges:

1. Sampling An LLM generates individualized "patient recipes" that define demographics and divide the patient's life trajectory into discrete temporal segments. This high-level blueprint ensures long-term clinical coherence—such as maintaining consistent medication dosages—while segmentation allows the system to bypass LLM context limits and avoid the "drifting" common in long-form generation.

2. Generation For each recipe, an LLM produces longitudinal medical events across the temporal segments. For each event, the LLM generates a natural language description as well as a target coding system (e.g., ICD-10 or SNOMED). At this stage, the descriptions are medically plausible but may not yet match official ontology strings exactly.

3. Matching Because LLMs are prone to hallucinating invalid codes or using imprecise language, the system uses a precomputed vector database (e.g., ChromaDB) to map free-text descriptions to standardized medical codes (SNOMED, ICD-10, LOINC, RxNorm, etc.).

4. Verification A final correctness check is performed where an LLM validates each completed record against the original input specifications. This acts as an automated quality gate, filtering out records that failed to follow the recipe or accidentally triggered exclusion criteria.

The resulting records are produced in the Medical Event Data Standard (MEDS) format (Arnrich et al., 2024). `neopatient` is designed to be scalable, using LLM batch APIs and state-tracking for resumability, enabling the cost-effective generation

of tens of thousands of records.

Static Resources The `neopatient` pipeline relies on two primary static resources. First, a vector database of medical codes ensures that natural language descriptions are accurately mapped to standardized ontologies. This database was constructed by embedding code descriptions (using Qwen 3 8B to produce 4096-dimensional vectors) and storing them in ChromaDB. Second, the library maintains a set of reference statistics derived from real-world EHR data. these statistics define the typical length and density of both inpatient and outpatient longitudinal records, ensuring that the generated synthetic trajectories reflect realistic clinical patterns.

Note on Realism The primary goal of `neopatient` is to generate records that strictly adhere to provided clinical specifications rather than to exhaustively replicate every facet of real-world patient records. This design choice serves the core objective of PopMedQA: to isolate and evaluate the specific computational challenges of the verbose context problem in a controlled setting.

C. Task Scoring

Task	Metric	Details
Planted Clique	Precision	$\lvert\text{Predicted} \cap \text{True}\rvert/k$
In-Context Classification	Accuracy	Binary classification accuracy on the test set
Clustering	Rand Index	Proportion of correctly identified pairs (same vs. different)
Inverse-Propensity Weighting	Absolute Error	Absolute difference between predicted and true ATE (Hajek)
Outlier Detection	F_1 Score	Harmonic mean of precision and recall
Two-Sample Test	Accuracy	0–1 accuracy in identifying if distributions differ
Similarity Search	Precision@k	Proportion of model's top-k that are in the true top-k
Top-k	Precision@k	Proportion of model's top-k that are in the true top-k
Sorting	Kendall's τ	Restricted to pairs from different ground-truth cohorts
Classification	Accuracy	Multi-class accuracy across all categories

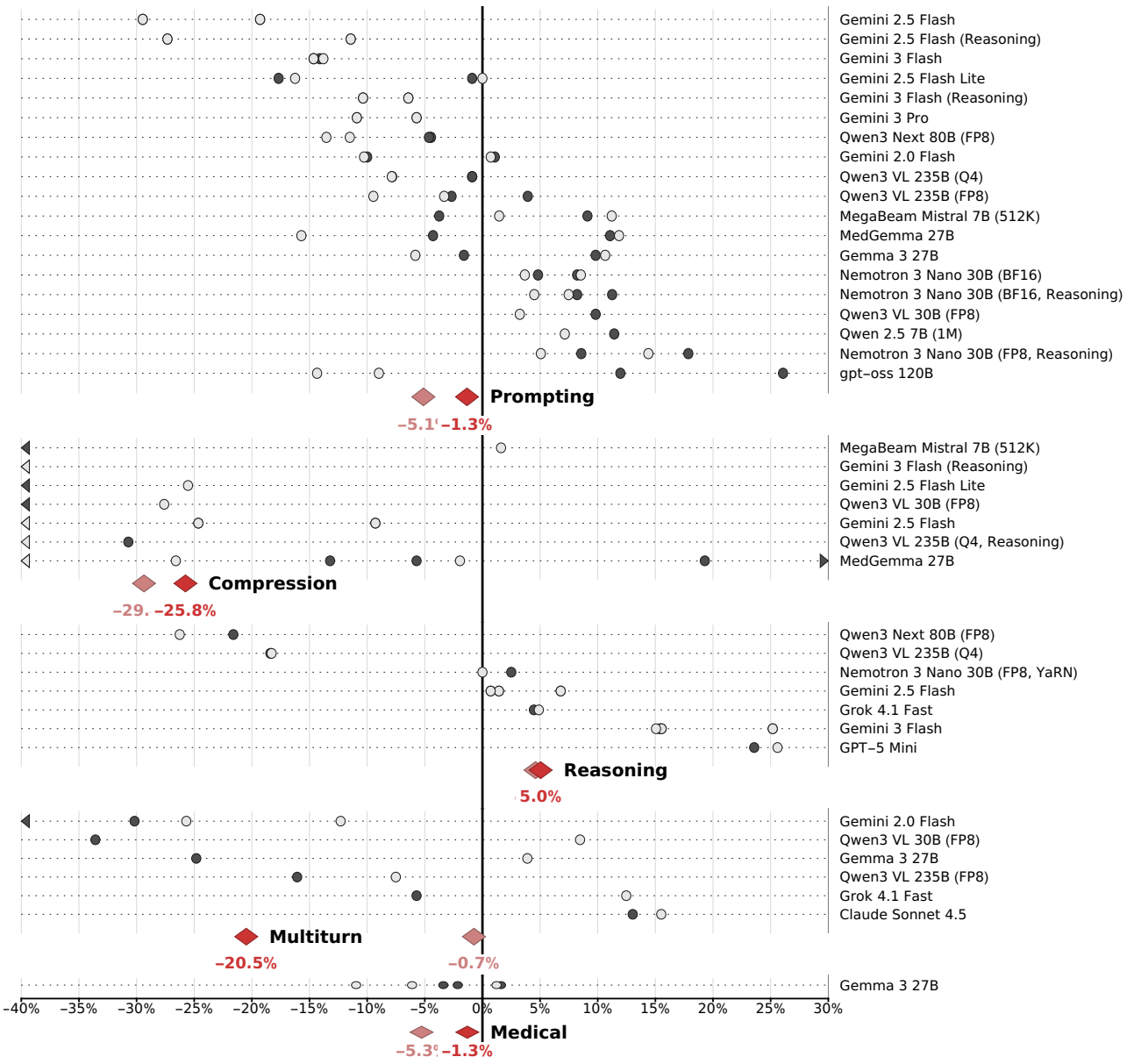

Figure 3. Meta-analysis of ablations on PopMedQA. Each dot compares two runs on PopMedQA: a baseline and an ablation. The model's name is on the right; different families of ablations are presented. A dot's position quantifies the effect of the ablation. A dark dot at -5% indicates that the baseline model won 55% of head-to-head comparisons, and therefore the ablation had a negative effect. A light dot restricts the scoring to examples where both models gave an answer; this distinction is important when the ablation affects the model's capability to return correctly-formatted answers. For each family, the mean ablation effects are shown as diamonds. Overall, we find that most families of ablations, besides reasoning, are ineffective on PopMedQA.

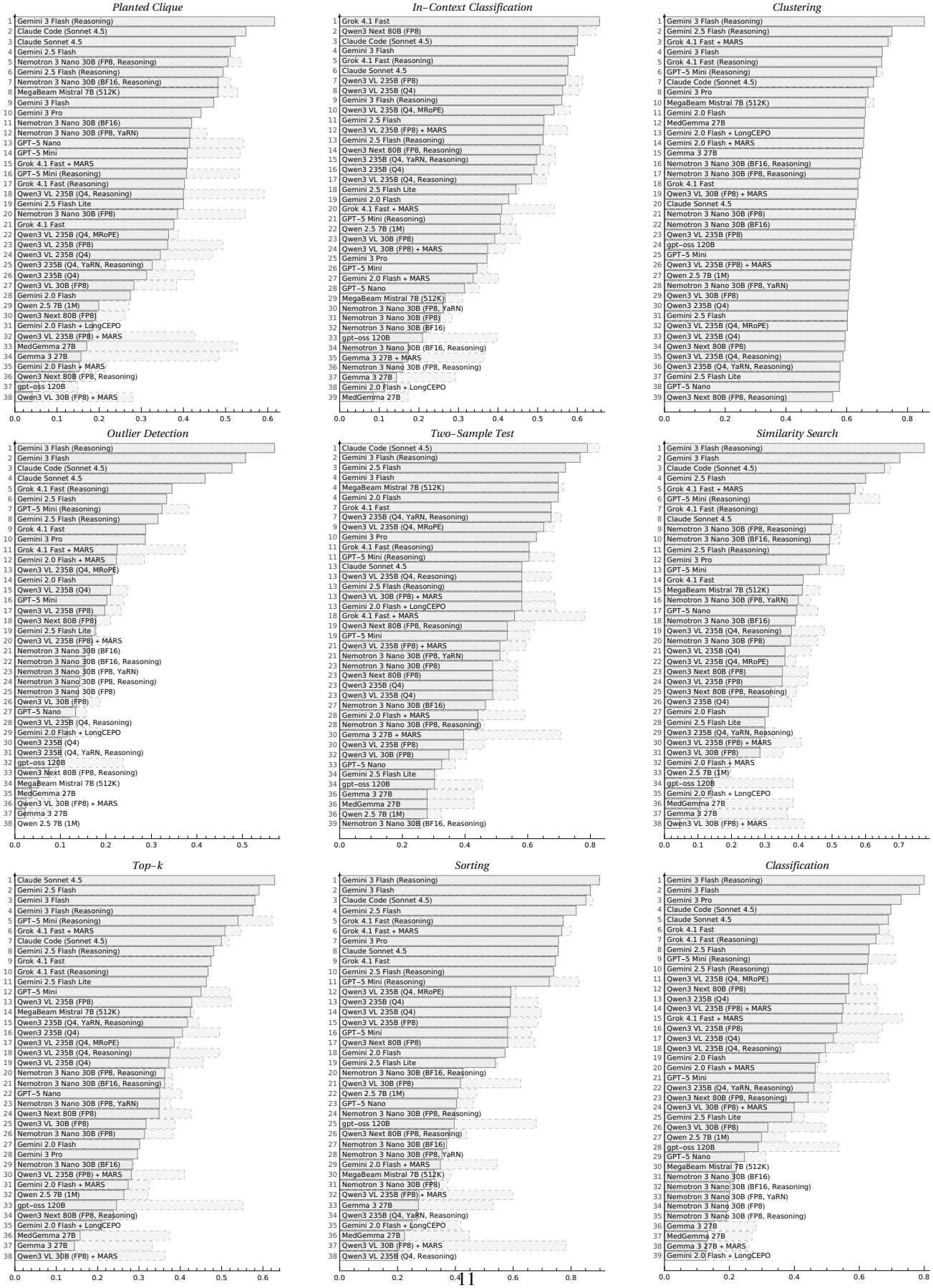

Figure 4. Task-specific scoreboards. The y-axis indicates the model's rank, and the x-axis denotes its mean score on the task. The filled-in bars with solid borders indicate the mean over all the questions in the task. The faint bars with dashed borders indicate the mean over just the questions the model answered properly (i.e. in the correct format).

```
=== Patient ID: 1182 ===
Time: 1968-07-25
Birth
Time: 2023-08-15
4256F - Anesthesia administration documented for l
CPT Code 44140 - Colectomy, partial; with anastomo
Excision of Head, Open Approach - This surgical pr
Diverticulosis of Large Intestine without Perforat
Partial Intestinal Obstruction (K56600) - This con
Cefazolin - Cefazolin Sodium, manufactured by Gene      1 vial
Morphine Sulfate Injection - Morphine Sulfate, man      10 ML
Sodium Chloride Injection Solution - Sodium Chlori      1 bag
Heart rate by Noninvasive       95 beats/min
Systolic blood pressure mean    135 mmHg
Diastolic blood pressure mean   82 mmHg
Oxygen saturation in Arterial blood by Pulse oxime      97 %
Thyrotropin [Units/volume] in Serum or Plasma --ba      37 C
Postoperative state - This term refers to the peri
General anesthesia - A medical procedure used to i
Postoperative monitoring - This refers to the syst
<... elided for figure ...>
Time: 2023-08-16
Heart rate by Noninvasive       105 beats/min
Systolic blood pressure mean    110 mmHg
Oxygen saturation in Arterial blood by Pulse oxime      94 %
Thyrotropin [Units/volume] in Serum or Plasma --ba      38.1 C
Prostate Tumor Incidental Histologic Finding - Thi
Postoperative period - This term refers to the rec
Bowel Sounds Quiet - This condition refers to the
<... elided for figure ...>
```

```
=== Patient ID: 1182 ===
Time: 1968-07-25
Birth
Time: 2023-08-15
cpt//4256F
cpt//44140
icd10_proc//0WB00ZZ
icd10//K57.31
icd10//K56.600
ndc//52584-924  1 vial
ndc//0409-1134  10 ML
ndc//0264-1800  1 bag
loinc//76477-9  95 beats/min
loinc//96608-5  135 mmHg
loinc//96609-3  82 mmHg
loinc//59408-5  97 %
loinc//14999-7  37 C
snomed//19585003
snomed//50697003
snomed//182775008
<... elided for figure ...>
Time: 2023-08-17
loinc//76477-9  112 beats/min
loinc//96608-5  105 mmHg
loinc//59408-5  92 %
loinc//14999-7  38.8 C
snomed//406127006
snomed//101379003
snomed//207206002
<... elided for figure ...>
```

Figure 5. Different prompting methods on the same patient record. (Left): the standard prompting method used as a baseline in this paper. It replaces the code altogether (since those are often not recognized by language models) by a truncated description of the code. (Right): a less verbose (and less informative) prompting method which includes the code but not the description.

```
Codebook (integer IDs assigned to medical codes):
#1: Birth of patient
#2: (cpt//0513T) 0513T - External shockwave, integrated wound healing, each additional wound.
#3: (cpt//1021887) Skilled Nursing Facility
#4: (cpt//1111F) CPT Code: 1111F - Discharge medication reconciliation with current medication review.
<... elided for figure ...>
#313: (snomed//91251008) Physical Therapy Procedure - A therapeutic intervention aimed at alleviating
     physical impairments and enhancing mobility and fun
=== Patient ID: 1182 ===
Time: 1968-07-25
#1
Time: 2023-08-15
#17
#20
#111
#77
#76
#171    1 vial
#160    10 ML
#157    1 bag
#144    95 beats/min
#152    135 mmHg
#153    82 mmHg
#137    97 %
#124    37 C
#222
#295
<... elided for figure ...>
Time: 2023-08-17
#144    112 beats/min
#152    105 mmHg
#137    92 %
#124    38.8 C
#280
#187
#225
<... elided for figure ...>
```

Figure 6. An alternative prompting method which attempts to eliminate redundancy across groups of patient records. It defines a succinct ID numbers for all unique codes (across all patients), and then references those IDs within the subsequent patient records.

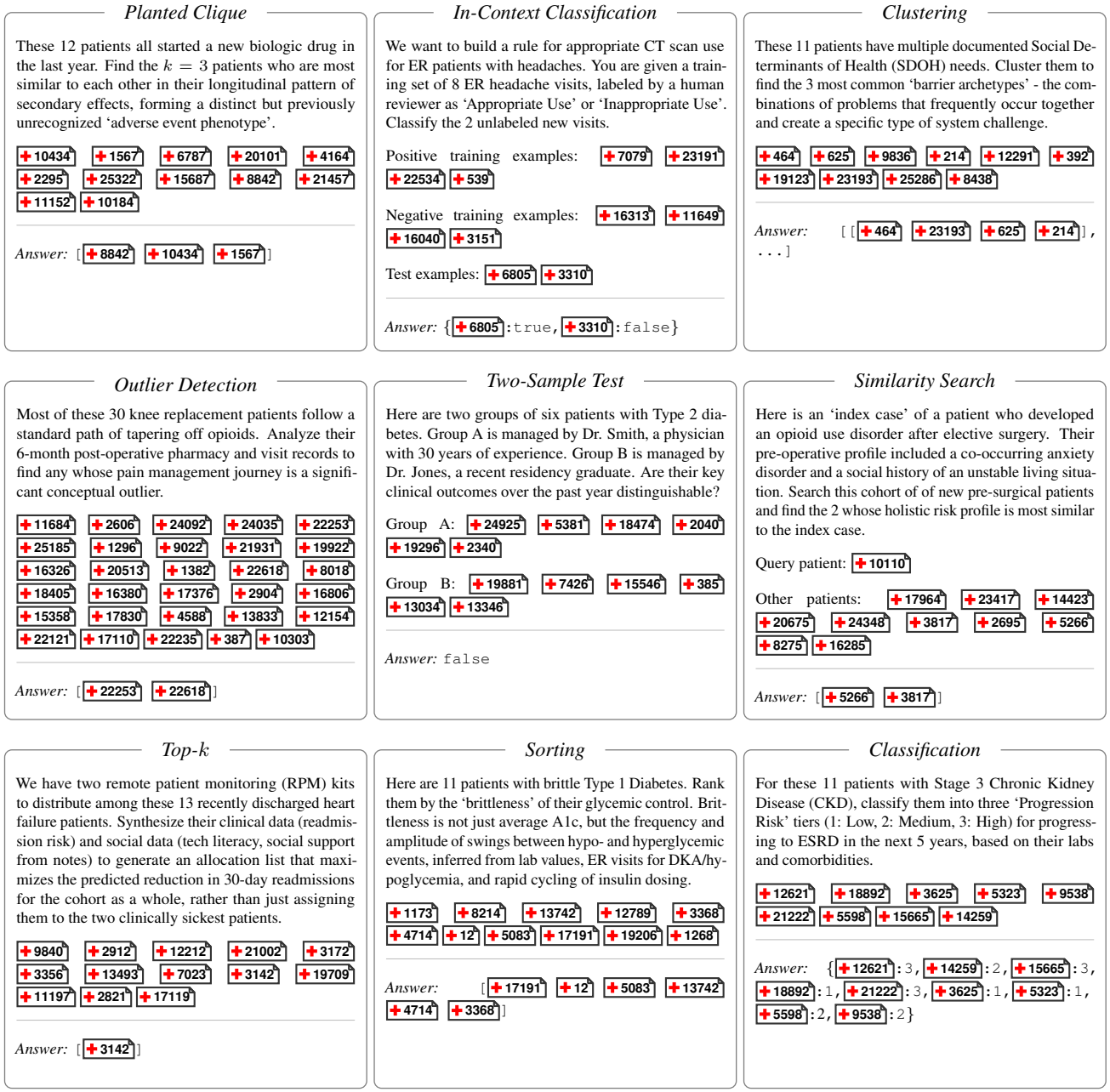

Figure 7. Example questions from PopMedQA. Each question poses one of nine computational tasks. Each patient record is visually depicted as a boxed patient ID number.