BIT.UA-AAUBS at ArchEHR-QA 2026: Evaluating Open-Source and Proprietary LLMs via Prompting in Low-Resource QA

Richard A. A. Jonker[*†], **Alexander Christiansen**[*], **Alexandros Maniatis**[*], **Rúben Garrido**[†], **Rogério Braunschweiger de Freitas Lima**[*], **Roman Jurowetzki**[*], **and Sérgio Matos**[†]

[*]Aalborg University Business School
Fibigerstræde 2, Aalborg East 9220, Denmark
{raaj, ach22, amania24, rlima23, roman}@business.aau.dk

[†]IEETA, DETI, LASI, University of Aveiro
Campus Universitário de Santiago, Aveiro 3810-193, Portugal
{richard.jonker, rubengarrido, aleixomatos}@ua.pt

Abstract

This paper presents the joint participation of the BIT.UA and AAUBS groups in the ArchEHR-QA 2026 shared task, which focuses on clinical question answering and evidence grounding in a low-resource setting. Due to the absence of training data and the strict data privacy constraints inherent to the healthcare domain (e.g. GDPR), we investigate the capabilities of Large Language Models (LLMs) without weight updates. We evaluate several state-of-the-art proprietary models and locally deployable open-source alternatives using various prompt engineering strategies, including task decomposition, Chain-of-Thought, and in-context learning. Furthermore, we explore majority voting and LLM-as-a-judge ensembling techniques to maximize predictive robustness. Our results demonstrate that while proprietary models exhibit strong resilience to prompt variations, domain-adapted open-source models (such as MedGemma 3 27B) achieve highly competitive performance when paired with the right prompt. Overall, our prompt-based approach proved highly effective, securing 1st place in Subtask 4 (evidence citation alignment) and 3rd place in Subtask 3 (patient-friendly answer generation). All code, results, and prompts are available on our GitHub repository: https://github.com/bioinformatics-ua/ArchEHR-QA-2026.

Keywords: Clinical Question Answering, Electronic Health Records, Large Language Models, Prompt Engineering, Biomedical NLP

1. Introduction

Patients increasingly seek to understand their health conditions and clinical course by reviewing their electronic health records (EHRs). However, clinical notes are notoriously complex, lengthy, and filled with medical jargon, making it difficult for patients to extract clear, accurate answers to their questions. The ArchEHR-QA 2026 (Soni and Demner-Fushman, 2026b,a) Shared Task addresses this by challenging systems to perform grounded question answering (QA) directly from patient-specific EHRs. Unlike general health QA, this task emphasizes that clinical grounding ensures generated answers are explicitly linked to supporting evidence within the clinical notes.

The ArchEHR-QA 2026 challenge is divided into four sequential Subtasks: interpreting a verbose patient question into a concise clinical query (Subtask 1), identifying relevant evidence sentences from the EHR (Subtask 2), generating a patient-friendly answer (Subtask 3), and aligning the generated answer with the underlying clinical evidence (Subtask 4). Developing supervised models for these highly specialized steps is difficult due to the lack of annotated data.

To overcome the lack of extensive training data, our team participated in all four Subtasks by exploring a purely generative approach. We evaluated a wide variety of Large Language Models (LLMs), encompassing both open-source and proprietary architectures, relying entirely on prompting strategies. Our goal was to determine the extent to which off-the-shelf LLMs can bridge the semantic gap between patient inquiries and EHR data across the entire QA pipeline. Specifically, this study addresses two core research questions:

- **RQ1:** To what extent can open-source LLMs bridge the performance gap to state-of-the-art proprietary models on complex clinical reasoning tasks?
- **RQ2:** To what extent does prompt engineering impact model performance and how does this vary between models.

Our comprehensive evaluation revealed substantial variability across the four Subtasks, highlighting the uneven capabilities of current LLMs in clinical contexts. While our systems achieved top results in the generation and alignment tasks, securing 1st place in Subtask 4 and 3rd place in Subtask

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 441–454
12 May 2026.

3, performance dropped considerably in the pure extraction and query formulation tasks. The remainder of this paper is structured as follows: Section 2 outlines the background and related work. Section 3 details our methodology, including our experimental setup and the diverse prompting techniques deployed. Section 4 presents our validation and official competition results, with Section 5 presenting some error analysis. Section 6 provides a discussion with Section 7 concluding our findings.

2. Background

Large Language Models in Clinical NLP. Until recently, state-of-the-art clinical natural language processing relied heavily on domain-specific, encoder-only architectures, such as ClinicalBERT (Huang et al., 2019), which required extensive supervised fine-tuning. Recently, the paradigm has shifted toward generative LLMs. Models such as GPT-4, Gemini, and Claude have demonstrated remarkable zero-shot and few-shot capabilities across complex medical tasks, including clinical summarization, diagnostic reasoning, and patient-friendly translation (Thirunavukarasu et al., 2023; Singhal et al., 2023). However, applying these models to EHRs introduces significant challenges regarding hallucination (Huang et al., 2025) and factual consistency, necessitating strict grounding mechanisms to ensure patient safety.

Prompt Engineering and In-Context Learning. While supervised fine-tuning remains the conventional standard for clinical NLP tasks, it is highly susceptible to overfitting in extreme low-resource settings and often impractical without extensive synthetic data generation (Meng et al., 2023). Consequently, advanced prompt engineering has emerged as the primary method for steering LLM behavior without weight updates (Sahoo et al., 2024; Chen et al., 2025). Foundational techniques like few-shot In-Context Learning (ICL) (Brown et al., 2020) and strict lexical constraints (Schall and de Melo, 2025) help anchor model outputs to specific source texts. For complex clinical reasoning tasks, decomposing instructions via Task Decomposition (Zhou et al., 2023) and Chain-of-Thought (CoT) prompting (Wei et al., 2022) have proven highly effective at imporving general performance. Furthermore, ensemble strategies, such as "LLM-as-a-judge" framework (Zheng et al., 2023), are increasingly utilized to smooth out the variance inherent to generative models, improving both the robustness and precision of structured outputs.

Privacy Constraints and Open-Source Alternatives. A primary barrier to deploying proprietary LLMs in real-world clinical environments is patient data privacy. Regulations such as the General Data Protection Regulation (GDPR) (European Union, 2016) and the Health Insurance Portability and Accountability Act (HIPAA) (104th United States Congress, 1996) often strictly prohibit the transmission of EHR data to external third-party APIs (Thirunavukarasu et al., 2023). Consequently, there is a growing imperative to develop and evaluate locally deployable, open-weights models. Recent releases such as MedGemma (Sellergren et al., 2025), attempt to bridge the performance gap between open-source and closed-source systems specifically in the medical domain. Evaluating the true viability of these open-weights models against state-of-the-art proprietary baselines remains a critical area of ongoing research.

3. Methodology

Given the extreme low-resource constraints of this competition, comprising a development set of only 20 samples, our methodology strictly utilizes prompt engineering over traditional fine-tuning. Across the pipeline, each of the four Subtasks incorporates an LLM component. To quantify the performance gap between state-of-the-art proprietary models and locally deployable open-source alternatives, we systematically evaluated varying configurations of prompts and model architectures. Broadly, our prompts fall into three general categories: constraint-based instructions that enforce output structure and terminology, extraction-focused steps that isolate key information prior to generation, and rephrasing strategies that guide the model toward the desired output style. Depending on the complexity of the Subtask, we applied zero-shot prompting, few-shot ICL, lexical constraints, Task Decomposition, and CoT pipelines. We frequently merged these techniques to ensure robust adherence to task guidelines and mitigate hallucination. All models were executed with a temperature of 0.0 and a top-p of 0.95 to maximize deterministic behavior, though we acknowledge this may not be optimal for all model architectures evaluated.

3.1. Subtask 1

Our objective for Subtask 1 was to transform verbose patient-authored questions into concise clinical questions strictly under a 15-word limit. Each prompt received the patient question as input and instructed the model to generate a clinician interpreted query representing the medical information needed. The generated question was required to follow a strict JSON format containing a single field *("query")*. To find the optimal balance between natural language generation and clinical accuracy, we systematically evaluated ten distinct prompt templates divided into two methodological strategies:

- **Direct Generation with Constraints**

(Prompts 1-7): This utilized a direct generation approach where the model was instructed to read the narrative and immediately synthesize a question. Prompt 1 established baseline zero-shot constraints (15-word limit, third-person perspective). Prompt 2 introduced lexical constraints to preserve exact medical terminology, to avoid paraphrasing, and Prompt 3 tested a few-shot alternative, providing 2 examples. Prompts 4 and 5 enforced verbatim extraction, explicitly forbidding the correction of misspellings to prevent model over-correction. Prompts 6 and 7 added structural rules enforcing the inclusion of medications/anatomical terms and a single continuous sentence format, respectively. However, stacking these negative and structural constraints within a single instruction frequently induced additional interference, causing the model to sporadically ignore rules or omit key terms.

- **Task Decomposition (Prompts 8-10):** To mitigate constraint overload, we transitioned to an extract-then-generate pipeline. Prompt 8 forced the model to execute an intermediate step: explicitly identifying consecutive two- and three-word medical phrases. Prompt 9 refined this to isolate the primary clinical concern prior to formulating the query. Prompt 10 combined this sequential extraction with few-shot ICL to maximize terminology retention while adhering to brevity limits.

3.2. Subtask 2

Subtask 2 evaluates a system's ability to identify clinical note sentences that provide evidence for answering a patient's question. Due to limited development data, we framed this as a prompt-based inference task, instructing models to return the relevant sentence identifiers in a structured JSON format. Each prompt received the clinician question along with the numbered sentences of the clinical note excerpt, and the model was asked to select the identifiers of sentences that provide supporting evidence. We evaluated the following prompt formulations and ensemble strategy:

- **Sentence-Level Classification (Prompts 1-2):** Models evaluated each sentence independently, assigning labels (essential, supplementary, not relevant). This approach introduced instability across sentences and required complex post-hoc aggregation.
- **Case-Level Evidence Selection (Prompts 3-8):** Prompts 3-6 received the entire excerpt and were instructed to output identifiers of relevant sentences directly, emphasizing a "minimal and sufficient" set. However, strict minimality often reduced recall by omitting partial supporting evidence. Prompts 7 and 8 introduced additional reasoning constraints (CoT) to improve transparency, but the increased prompt complexity destabilized output formatting.
- **Recall- and Precision-Oriented Selection (Prompts 9-10):** To explore the trade-off between coverage and specificity, we tested two simplified formulations. Prompt 9 instructed the model to prefer precision over recall, selecting only the most directly relevant sentence identifiers, while Prompt 10 favoured recall over precision, selecting all sentence identifiers that could help answer the clinician's question to reduce the risk of missing relevant evidence.
- **Ensemble Aggregation:** To finalize predictions, we employed a majority voting ensemble across multiple LLMs (Dietterich, 2000). Each model independently generated predictions, and sentence identifiers selected by a minimum of 2/3 models were included in the final output, reducing spurious selections while preserving consistently identified evidence.

3.3. Subtask 3

For Subtask 3, the goal was to synthesize a patient-friendly, evidence-grounded answer (maximum 75 words). Each prompt received the patient question, clinician question, and numbered clinical note sentences as input. Because this Subtask heavily evaluates factual consistency and adherence to the source text, our prompt engineering focused on preventing hallucination (Huang et al., 2025) through the following strategies:

- **Few-Shot with Explicit Constraints (Prompts 1-5, 8, 10):** We utilized a 1-shot learning approach, providing an annotated example from the task guidelines. We systematically varied the strictness of length instructions (e.g., "exactly 3 to 5 sentences" vs. "maximum 75 words") and paraphrasing allowances.
- **Multi-Shot Demonstration (Prompt 6):** Expanded ICL by providing two distinct clinical examples, testing whether exposure to diverse specialties improved the model's ability to synthesize answers without exceeding the strict word count.
- **Zero-Shot Abstractive Summarization (Prompt 7):** Removed all in-context examples to test baseline summarization capabilities, relying purely on instructional constraints

to generate faithful, evidence-grounded answers.

- **Implicit Chain-of-Thought / Pre-computation (Prompt 9):** Introduced a cognitive constraint requiring the LLM to silently identify relevant numbered sentences prior to generation. This anchored the model to an extractive mindset without polluting the final output with reasoning traces.
- **High-Density Persona & Negative Constraints (Prompt 11):** Shifted the system persona to a strict "clinical auditor" and densely stacked negative constraints (e.g., "Do NOT infer, generalize, explain, justify, speculate") to strongly discourage creative text generation.
- **LLM-as-a-Judge:** To better generalize results across model-specific biases, we employed an LLM-as-a-Judge (Zheng et al., 2023) to select the single best candidate answer per case from a pool of outputs with the strongest model-prompt configurations. The judge evaluated candidates according to the following criteria, in priority order: (1) faithfulness: every claim must be supported by the clinical note sentences. (2) medical completeness: covering the key clinical information that answers the question. (3) terminology retention: preserving exact medical terms, procedure names, and diagnoses from the note. (4) conciseness: no filler or hedging, within the 75-word limit.

3.4. Subtask 4

Subtask 4 requires systems to link each sentence of a generated answer back to supporting sentences in the clinical note excerpt. Each prompt received the patient question, clinician question, the numbered clinical note sentences, and the generated answer sentences. To ensure accurate sentence-level labels, we preprocessed the texts by explicitly labeling them with structural identifiers prior to prompting. Specifically, each sentence in the source clinical note was prepended with a unique tag (e.g., [1], [2]), and each sentence in the generated answer was similarly tagged. In a variant, answer sentences were additionally prefixed with e.g. [N1], [N2] and [S1], [S2] tags to further disambiguate source and answer identifiers. Some configurations also employed a two-step verification pass, prompting the model to revisit any unaligned answer sentences after the initial output. We framed the model as a clinical evidence alignment expert tasked with generating strictly valid JSON under four core rules: only explicitly supported facts may be cited, references must contain the minimal set that contains all relevant information, unsupported sentences must return empty lists, and all answer sentences must be processed.

- **Zero-Shot Alignment (Prompts 1, 10):** Prompt 1 established the baseline with markdown-style rules and explicit-support constraints. Prompt 10 utilized a stricter formulation, explicitly forbidding inference and generalization. However, zero-shot models struggled to consistently differentiate direct evidence from loosely related context.
- **Progressive Few-Shot Scaling (Prompts 2-9):** Prompt 2 introduced a 2-shot setting with clean alignment structures. Prompt 3 expanded to a 4-shot setting incorporating complex cases with high citation density and unsupported answers. Prompt 4 added a fifth conservative example to pull the model toward precise evidence linking. Prompt 5 tested a reordering of the Prompt 4 set to investigate and mitigate recency effects. Prompts 6 and 7 shifted back to 4-shot with alternative cases, while Prompts 8 and 9 extended this to 5-shot configurations, systematically varying example ordering and composition to identify the most effective sequence.
- **Ensemble Aggregation:** Because all few-shot examples were drawn from the 20-case development set, individual prompt configurations risked overfitting. To improve generalization, we searched over candidate model-prompt configurations to construct an ensemble.

4. Results

We employed a two-stage evaluation methodology across all four Subtasks. In the initial stage, we conducted extensive validation on the 20-case development set, evaluating a representative, though non-exhaustive, pool of state-of-the-art proprietary models (e.g., Gemini 3 Flash (Doshi and The Gemini Team, 2025), Gemini 2.5 Flash (Comanici et al., 2025), Claude Sonnet 4.5 (Anthropic, 2025), GPT 4.1 (OpenAI, 2025), Grok 4.1 Fast (xAI, 2025)) alongside open-source alternatives (e.g., Llama 3.1 8B Instruct (Llama Team, AI @ Meta, 2024), Qwen 3 (Qwen Team, 2025) and various domain finetunes (Tran et al., 2024)[1][2], MedGemma-27B (Sellergren et al., 2025). We iteratively adjusted this baseline pool to incorporate newer, higher-performing models as they became available. This validation phase identified our top-performing configurations,

[1] https://huggingface.co/khazarai/Bio-8B-it

[2] https://huggingface.co/Echelon-AI/Med-Qwen2-7B

which were subsequently applied to the hidden test set (comprising 47 samples for Subtasks 1-3 and 147 samples for Subtask 4) to generate the official results. To maximize competitive performance, our final submissions featured proprietary models rather than open-source alternatives; we acknowledge this reliance as a limitation of our work. Regarding computational resources, smaller open-source models (<70B parameters) were executed on a SLURM cluster utilizing up to two NVIDIA L4 (24GB) GPUs. All other models were accessed via API endpoints (OpenRouter), incurring a total cost of $109.45 USD for the duration of the competition. The following subsections detail the validation findings and official performance for each Subtask, with a a complete set of validation experiments for Subtasks 1, 2, 3, and 4, including a brief evaluation of very large open-source models, is provided in the Appendix, for Subtask 1, 3, 4.

4.1. Subtask 1

Performance in Subtask 1 is evaluated using an average of four automatic text generation metrics: ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). Our validation results are presented in Figure 1. Across the Direct Generation phase (Prompts 1-7), performance gradually improved as constraints were refined, though scores remained generally lower than those in the Task Decomposition phase (Prompts 8-10). Task Decomposition consistently achieved the highest average scores across model architectures, suggesting that breaking the task into sequential extraction steps is a highly effective strategy. Based on these validation results, Prompts 8, 9, and 10 were identified as the strongest configurations, with prompt 10 offering the best performance for most proprietary models, indicating the advantage of using few shot examples. In general we note that most proprietary models offer alot more performance and robustness than most open source alternatives.

The official leaderboard results for Subtask 1 are presented in Table 1. Our primary submission utilized Claude Sonnet 4.5 with Prompt 10, combining the sequential extraction pipeline with 2-shot in-context learning (ICL) to maximize terminology retention within the 15-word limit. This approach achieved a score of 19.0, ranking 13th on the leaderboard. A secondary submission using GPT-5.2 with Prompt 8 (extract-then-generate) achieved a score of 16.6, indicating that the in-context examples in Prompt 10 provided a meaningful advantage. Notably, the substantial discrepancy between our internal validation scores and the official test results suggests a potential distribution shift in the test data, an overfitting effect to the highly constrained 20-case development set, or some internal errors

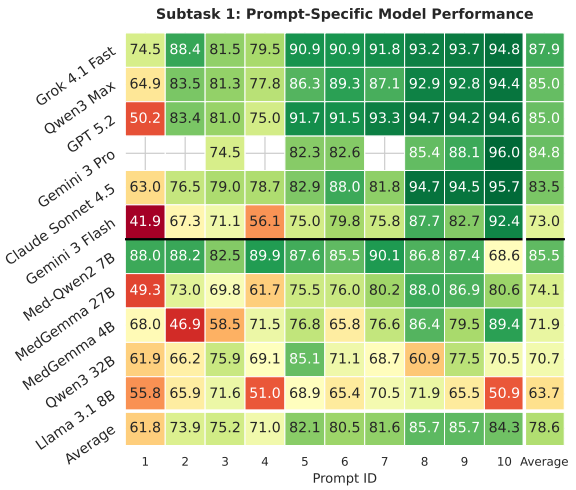

Figure 1: Subtask 1 Validation Results showing the different prompts over several different open and close source models. The scores represent the official evaluation metrics of the average of ROUGE, BERTScore, AlignScore, and MEDCON.

on our part.

Model / Team	Rank	Score
Our Submissions		
Sonnet-4.5 Prompt 10	13	19.0
GPT-5.2 Prompt 8	–	16.6
Leaderboard		
Best Competitor	**1**	**31.2**
Median	7	25.6

Table 1: Subtask 1 Official Results. The score represents the average of ROUGE, BERTScore, AlignScore, and MEDCON.

4.2. Subtask 2

Performance in Subtask 2 is evaluated using the Strict Micro F1 score to measure a system's ability to accurately identify the minimal and sufficient set of relevant evidence sentences. Validation results (Figure 2) highlight a distinct performance gap between proprietary and open-source models. Proprietary models consistently achieved the highest scores, typically exceeding 50.0 Micro F1 and demonstrating relative stability across various prompt configurations. In contrast, open-source models produced substantially lower scores and exhibited higher sensitivity to prompt variations. Further we note many cases of complete failure (F1 score of zero), where the model/prompt configuration would generate responses which were unparseable. In terms of our prompting approaches, we noticed no clear winners in terms of prompts, with various models having various

optimal prompts, with the exception of prompt 10, obtaining top performance for most propriety models.

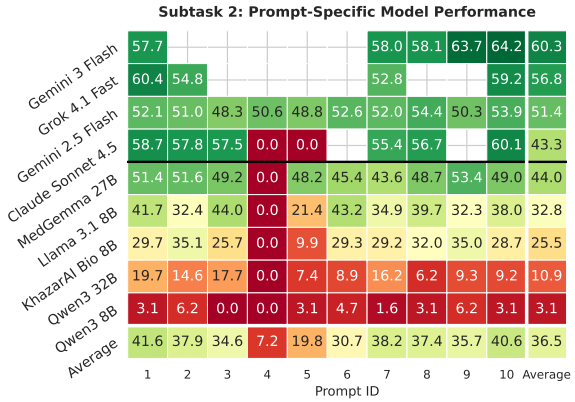

Figure 2: Subtask 2 Validation Results over the Strict Micro F1 metric

Consequently, our final system relied on an ensemble of the strongest proprietary models. The primary submission, a majority voting ensemble comprising Gemini 3 Flash Preview, Grok 4.1 Fast, and Claude Sonnet 4.5 (all using prompt 10), achieved a Strict Micro F1 score of 58.8, ranking 11th on the leaderboard (Table 2). This performance is near the median system score (59.8) and remains competitive with the top-performing system (63.7). An alternative single-model submission (Grok 4 Fast Prompt 10) achieved a score of 56.0. The minimal variance between the ensemble and single-model runs demonstrates the stability of the prompting strategy, suggesting that a single model is highly reliable and rendering more computationally expensive ensembling strategies unnecessary.

Model / Team	Rank	Strict Micro F1
Our Submissions		
Ensemble Top 3	11	58.8
Grok 4 Fast (Prompt 10)	–	56.0
Leaderboard		
Best Competitor	**1**	**63.7**
Median	7/8	59.8

Table 2: Subtask 2 Performance: Strict Micro F1 on ArchEHR-QA.

4.3. Subtask 3

Performance in Subtask 3 is evaluated using an average score over the metrics: BLEU, ROUGE, SARI (Xu et al., 2016), BERTScore, AlignScore, and MEDCON. Validation results (Figure 3) revealed that performance narrowly varied across all eleven prompts. This indicates that for patient-friendly answer generation, models are relatively robust to prompt variation, making model selection the primary driver of performance. Gemini 2.5 Flash and Claude Sonnet 4.5 consistently outperformed other models, while open-source models like MedGemma-27B and a domain fine-tuned Qwen 3-8B[3] performed competitively, narrowing the performance gap observed in other Subtasks.

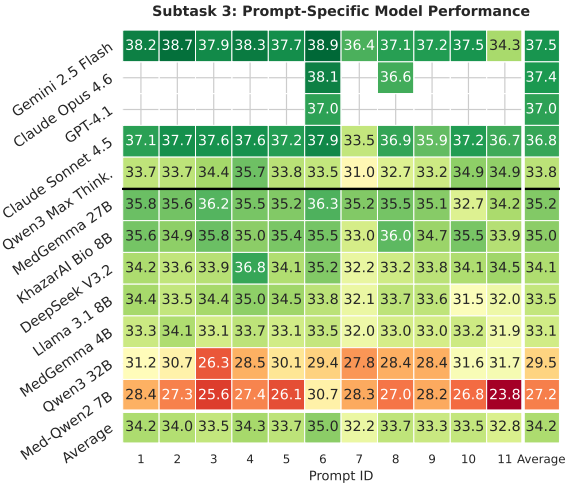

Figure 3: Subtask 3 Validation Results using an average score of the metrics: BLEU, ROUGE, SARI, BERTScore, AlignScore, and MEDCON

Given the marginal prompt-based variance, our submission strategy prioritized model diversity and high-quality demonstrations. Our first submission utilized an LLM-as-a-Judge (GPT-4.1) over the strongest model-prompt configurations: Gemini 2.5 Flash, Claude Sonnet 4.5, and Claude Opus 4.6 on prompt 6. This approach ranked 3rd on the official leaderboard with a score of 35.6 (Table 3), closely trailing the top competitor (36.3). Single-model submissions also demonstrated strong performance: Gemini 2.5 Flash (Prompt 6, multi-shot) achieved 35.5, while Claude Sonnet 4.5 (Prompt 2, 1-shot explicit constraints) scored 33.9. This suggests that for this generation task, the LLM-as-a-Judge does not add much value relative to the overhead. The narrow 0.7-point gap between our best ensemble and the winning submission underscores the viability of prompt-based approaches for simplified medical text generation.

4.4. Subtask 4

Performance in Subtask 4 is evaluated using sentence-level precision, recall, and Micro F1 metrics over predicted alignment links. Validation results (Figure 4) indicated that using few shot prompt-

[3] https://huggingface.co/khazarai/Bio-8B-it

Model / Team	Rank	Score
Our Submissions		
LLM-as-a-Judge 4 Models	3	35.6
Gemini-2.5-Flash (Prompt 6)	–	35.5
Sonnet-4.5 (Prompt 2)	–	33.9
Leaderboard		
Best Competitor	**1**	**36.3**
Leaderboard Median	7	32.9

Table 3: Subtask 3 Performance Comparison. The Overall metric is a composite score evaluating simplified medical text generation, incorporating BLEU, ROUGE-Lsum, SARI, BERTScore, AlignScore, and MEDCON (UMLS).

ing (Prompts 2-9), often gave the open source models a clear benefit, with most models obtaining their best scores between prompts 5-7. However, because all few-shot examples were drawn from the evaluation development set, we believed that this might reflect partial overfitting. In general, proprietary models significantly outperformed open-source alternatives. Google's MedGemma-27B was a notable exception (average 81.0), though it still trailed the top proprietary models.

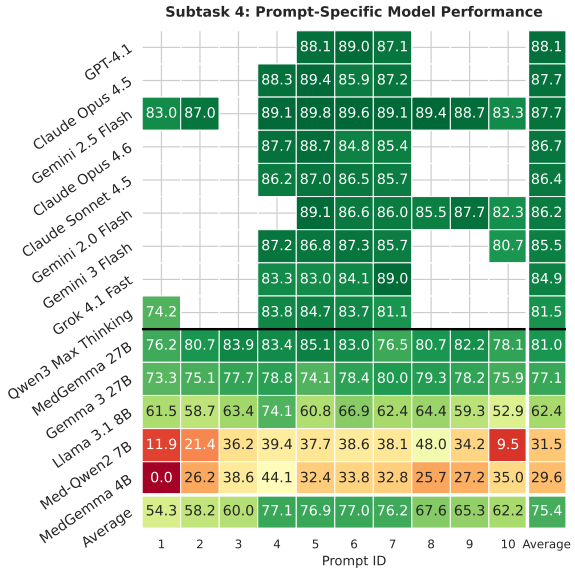

Figure 4: Subtask 4 Validation Results

To maximize robustness, we utilized an ensemble search over the strongest model-prompt combinations (Prompts 5-7), with the final submissions using Prompts *5* and *6*. The final ensemble composition was selected based solely on Micro F1 performance on the 20-case development set, with no subsequent adjustment made prior to or following test set submission. Our primary submissions, a 3-model majority voting ensemble (Gemini 2.5 Flash, Claude Opus 4.6, Gemini 2.0 Flash) with *Prompt 5*, achieved 1st place on the leaderboard with a Micro F1 of *81.5* (Table 4). Our second submission used a 4-model ensemble combining Gemini 2.5 Flash and GPT-4.1, each evaluated with *Prompts 5 and 6*, resulting in four model-prompt configurations and achieving *81.3*. Finally, a strictly constrained zero-shot single-model configuration (Gemini 2.5 Flash, Prompt 10) achieved *81.2*, demonstrating that carefully engineered zero-shot prompting can perform competitively with few-shot ensembles. Although the performance differences between our approaches is minimal, the margin between the 1st and 4th place submission was 1.2 Micro F1, indicating that these marginal changes affect the ranking of the systems. Furthermore, despite earlier concerns about potential overfitting due to few-shot examples drawn from the development set, the similarity between validation and official results suggests the overfitting was minimal.

Model / Team	Rank	Micro F1
Our Submissions		
Ensemble 3 models	**1**	**81.5**
Ensemble 4 models	–	81.3
Gemini 2.5 Flash (Prompt 10)	–	81.2
Leaderboard		
Best Competitor	2	81.3
Median	8/9	77.5

Table 4: Subtask 4 Performance: Micro F1 on ArchEHR-QA.

5. Error Analysis

Subtask 1: Question Interpretation Error Analysis. The model occasionally failed to transform patient narratives into professional clinical queries. In Case 129, it generated an informal, query with lots of spelling errors (*"Why is he not eatin feeling weak and shakey loosin weight..."*) rather than the required concise clinical formulation (*"What explains his weakness, weight loss, and poor appetite..."*). Upon reflection this is a model directly following our instructions where we ask specifically to not rephrase or correct spelling errors.

Subtask 2: Evidence Retrieval Error Analysis. Retrieval primarily suffered from over-selection and missed evidence. In over-selection cases (e.g., Case 19), the system successfully retrieved the gold annotations but included numerous false positives containing irrelevant procedural background due to an over-reliance on broad semantic similarity. Notably, our attempts to mitigate this using stricter filtering and ensemble methods have not fully resolved the issue. Conversely, in missed evidence

cases (e.g., Case 20), the model hyper-focused on explicit diagnostic terminology, completely overlooking relevant sentences describing treatments. Addressing this requires non-trivial improvements in semantic reasoning.

Subtask 3: Answer Generation Error Analysis. Generation errors typically involved medical hallucinations or context misinterpretation. In Case 21, the model correctly identified a patient's cirrhosis but hallucinated plausible yet entirely unsupported complications (*portal gastropathy, Budd-Chiari syndrome*). In Case 29, it ignored the primary clinical note regarding deep vein thrombosis treatment, instead generating an unrelated discussion about Coumadin and alcohol use. These issues emphasize the need for stronger grounding mechanisms to tightly couple generation with the retrieved evidence.

Subtask 4: Evidence Alignment Error Analysis. The model demonstrated difficulty with multi-evidence reasoning, where multiple context sentences jointly support a single claim. In Case 3, the model correctly cited one supporting sentence but missed a second complementary citation detailing warning symptoms, largely because the missing sentence utilized different terminology. Mitigating these incomplete citations will require prompting and retrieval strategies that explicitly encourage aggregating complementary evidence.

6. Discussion

In this work, we aimed to evaluate the limits of LLMs in the extreme low-resource biomedical setting of the ArchEHR competition, relying primarily on prompt engineering and model selection rather than supervised fine-tuning. Our findings highlight several key dynamics regarding model scaling, open-source viability, and the practical trade-offs of ensemble methodologies in clinical applications.

Addressing the viability of open-weights models for secure clinical deployment (RQ1), internal validation highlighted the impressive capabilities of open-source, domain-adapted models. Notably, MedGemma 3 27B performed exceptionally well, closely trailing large proprietary counterparts despite its relatively small parameter count. This shrinking performance gap is critical for the healthcare domain, where data privacy regulations often strictly prohibit sending sensitive patient data to external API endpoints. Locally deployable domain tuned open-source models are often the only option in these settings; however, our findings confirm that more work is needed to extract optimal performance from these smaller models.

Regarding the effectiveness of prompt engineering (RQ2), our validation phase revealed that the impact of specific techniques is inversely proportional to baseline model capability. State-of-the-art proprietary models generally demonstrated high robustness, yielding less performance variance across different prompt templates while maintaining higher relative performance, with several prompts often achieving similarly top-tier results. Conversely, open-source LLMs exhibited significant, model-dependent performance swings (e.g., the high variance between Qwen 3-32B and MedGemma architectures in Subtask 1, though Subtask 3 proved an exception). However, even for the most robust models, careful prompt tuning remained a necessity for enforcing strict structural constraints, such as rectifying the consistent JSON formatting failures observed in Subtask 2. Furthermore, it is important to note that poorly constructed prompts will still degrade the performance of proprietary models, as evidenced by Prompt 1 in Subtask 1. In general, regarding proprietary models, once a baseline of high performance is reached, further prompt engineering yields diminishing returns. This is aligned with the existing literature (Zhuo et al., 2024).

To maximize our competitive standing, we utilized ensembling techniques across the pipeline, including majority voting and LLM-as-a-judge frameworks. While these strategies consistently improved robustness and yielded positive results, the absolute performance gains were often marginal. In a shared task competition, fractional improvements are vital for leaderboard positioning. However, from a real-world clinical deployment perspective, these techniques introduce severe operational bottlenecks. Running multiple inferences per query effectively multiplies operational costs or latency by a factor of the ensemble size. For live, scalable health systems, the marginal gains of ensembling are not worth the resulting computational inefficiency.

Broadly, the ArchEHR competition demonstrates that LLMs are highly competitive in this low-resource clinical setting, though their efficacy varies significantly by task type. For Subtask 3 (answer generation), LLMs remain the natural and dominant choice, excelling at abstractive synthesis. Surprisingly, our prompt-based ensemble also achieved state-of-the-art results in Subtask 4 (evidence citation), which we initially hypothesized would be dominated by fine-tuned systems. Conversely, the leaderboard standings for Subtask 2 suggest that other approaches may be more optimal over prompt-based LLMs for pure evidence-span retrieval, if the best competitor uses other techniques. Finally, we note that our official Subtask 1 performance was an anomaly compared to our internal validation; we believe there remains substantial, untapped potential for LLM-driven query transformation in this space.

7. Conclusion

This study investigated the efficacy of LLMs in the extreme low-resource clinical setting of the ArchEHR competition. By systematically evaluating zero-shot, few-shot, and CoT prompting strategies alongside ensemble methodologies, we demonstrated that state-of-the-art LLMs can achieve highly competitive performance across complex biomedical NLP tasks without any supervised fine-tuning. Notably, our prompt-based ensemble approach achieved first place in Subtask 4 (evidence citation alignment) and third place in Subtask 3 (patient-friendly answer generation), proving that LLMs are naturally suited for both abstractive synthesis and strict evidence-linking. Conversely, our lower performance in Subtasks 1 and 2 indicates that pure evidence-span retrieval and strict query formulation remain challenging for generative, prompt-only architectures.

Our findings also highlight critical dynamics for the real-world deployment of clinical AI. While massive proprietary models exhibited strong robustness to prompt variations, open-source models, particularly MedGemma 3 27B, demonstrated highly competitive capabilities with certain prompts. This narrowing performance gap is a crucial development for healthcare environments bound by strict data privacy frameworks like GDPR, where local deployment is mandatory. Finally, while ensembling techniques such as majority voting and LLM-as-a-judge provided the marginal gains necessary for competitive leaderboard positioning, their associated computational costs and latency make them inefficient for scalable, real-world clinical deployment.

8. Limitations

While our work showed good results, this study has several notable limitations. First, due to the extreme low-resource nature of the shared task (only 20 development cases), there is a persistent risk that our few-shot prompts and ensemble configurations are partially overfit to the development distribution, despite our efforts to utilize varying case structures. This impacts the conclusions drawn from the validation results. Additionally, our uniform use of temperature 0.0 and top-p of 0.95 across all models, while intended to maximize deterministic behavior, may not be optimal for all architectures evaluated, particularly reasoning-oriented models that are designed to benefit from some degree of stochasticity.

Second, our evaluation of proprietary LLMs was not exhaustive, and was not run on all prompt configurations. The financial constraints associated with commercial API costs limited our ability to perform evaluate every prompt variation across all available proprietary models. Consequently, the optimal prompt-model combinations reported may reflect a local maximum rather than absolute peak performance.

Furthermore, our system heavily relies on proprietary, closed-weights models (e.g., Gemini and Claude) accessed via API endpoints to achieve top-tier performance. This introduces reproducibility challenges, as the underlying model weights may be silently updated by the providers. And in GDPR strict scenarios this cannot be permitted. Finally, the evaluation was restricted to English clinical notes; the efficacy of these prompting strategies on multilingual or non-English EHRs remains unverified.

9. Ethical Considerations

The deployment of generative LLMs in clinical settings carries profound ethical implications, primarily concerning patient safety and data privacy. Generative models are inherently prone to hallucination. While our system achieved high precision in evidence alignment (Subtask 4) and grounding (Subtask 3), the risk of clinical hallucination cannot be entirely eliminated through prompt engineering alone. Therefore, systems like ours must strictly be deployed as "human-in-the-loop" assistive tools to reduce clinician cognitive load, rather than as autonomous diagnostic or patient-facing decision-makers.

Additionally, our highest-performing submissions rely on sending clinical text to external APIs. In real-world healthcare environments governed by privacy frameworks such as HIPAA and GDPR, transmitting sensitive Protected Health Information to external servers is often legally prohibited. While we demonstrated that local, open-weights models like MedGemma 3 27B show strong promise, bridging the performance gap between secure local models and proprietary systems remains a critical ethical and technical imperative for the field.

10. Acknowledgments

This work was funded by FEDER - Fundo Europeu de Desenvolvimento Regional funds through Programa Regional do Centro, within project CENTRO2030-FEDER-02595400 and by the Foundation for Science and Technology (FCT) through the contract https://doi.org/10.54499/UID/00127/2025. Richard A. A. Jonker is funded by the FCT doctoral grant PRT/BD/154792/2023, with DOI identifier https://doi.org/10.54499/PRT/BD/154792/2023. Alexander Christiansen and Alexandros Maniatis are funded by the Danish

Data Science Academy (DDSA) under grant numbers 2026-6484 and 2026-6480, respectively.

11. Bibliographical References

104th United States Congress. 1996. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Public Law 104-191, 110 Stat. 1936.

Anthropic. 2025. Introducing claude sonnet 4.5. Anthropic News. Accessed: 2026-03-10.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6).

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Tulsee Doshi and The Gemini Team. 2025. Gemini 3 flash: frontier intelligence built for speed. Google The Keyword Blog. Accessed: 2026-03-10.

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 1-88.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR.

OpenAI. 2025. Introducing gpt-4.1 in the api. OpenAI Blog. Accessed: 2026-03-10.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1.

Maximilian Schall and Gerard de Melo. 2025. The hidden cost of structure: How constrained decoding affects language model performance. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1074–1084, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Sarvesh Soni and Dina Demner-Fushman. 2026a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *Scientific Data*.

Sarvesh Soni and Dina Demner-Fushman. 2026b. Overview of the archehr-qa 2026 shared task on grounded question answering from electronic health records. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

xAI. 2025. Grok 4.1 fast and agent tools api. xAI News. Accessed: 2026-03-10.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A. Additional Experiments

In this appendix, we present the full set of experiments conducted on the development set for Subtasks 1, 3, and 4 (Figures 5, 6, and 7, respectively). Notably, these extended validation results demonstrate that very large open-source models (such as Qwen 3 235B, GLM4.7, and Deepseek V3.2) performed exceptionally well, achieving results highly competitive with proprietary models. However, we chose not to emphasize these models in the main text due to their prohibitive local hardware requirements. Running these models locally requires substantial compute resources (often exceeding 140GB of VRAM), which is impractical for most standard biomedical deployments. Because of this limitation, we accessed these open-source models via API endpoints during testing; however, for API-based applications, we opted to focus our primary analysis on established proprietary models. For these practical reasons, we did not investigate these massive open-source models further in the main study.

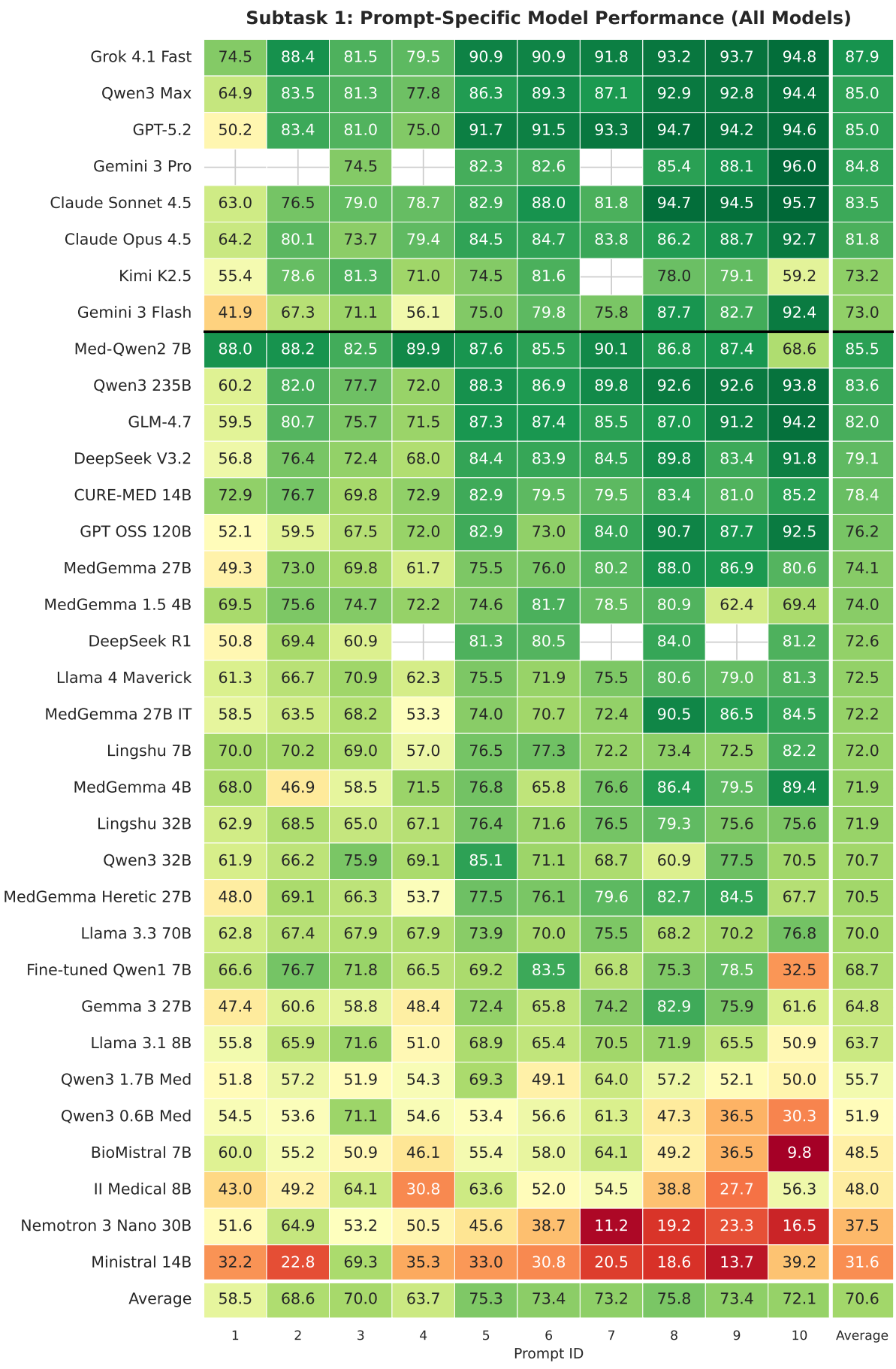

Figure 5: Subtask 1 Validation Results (All Models)

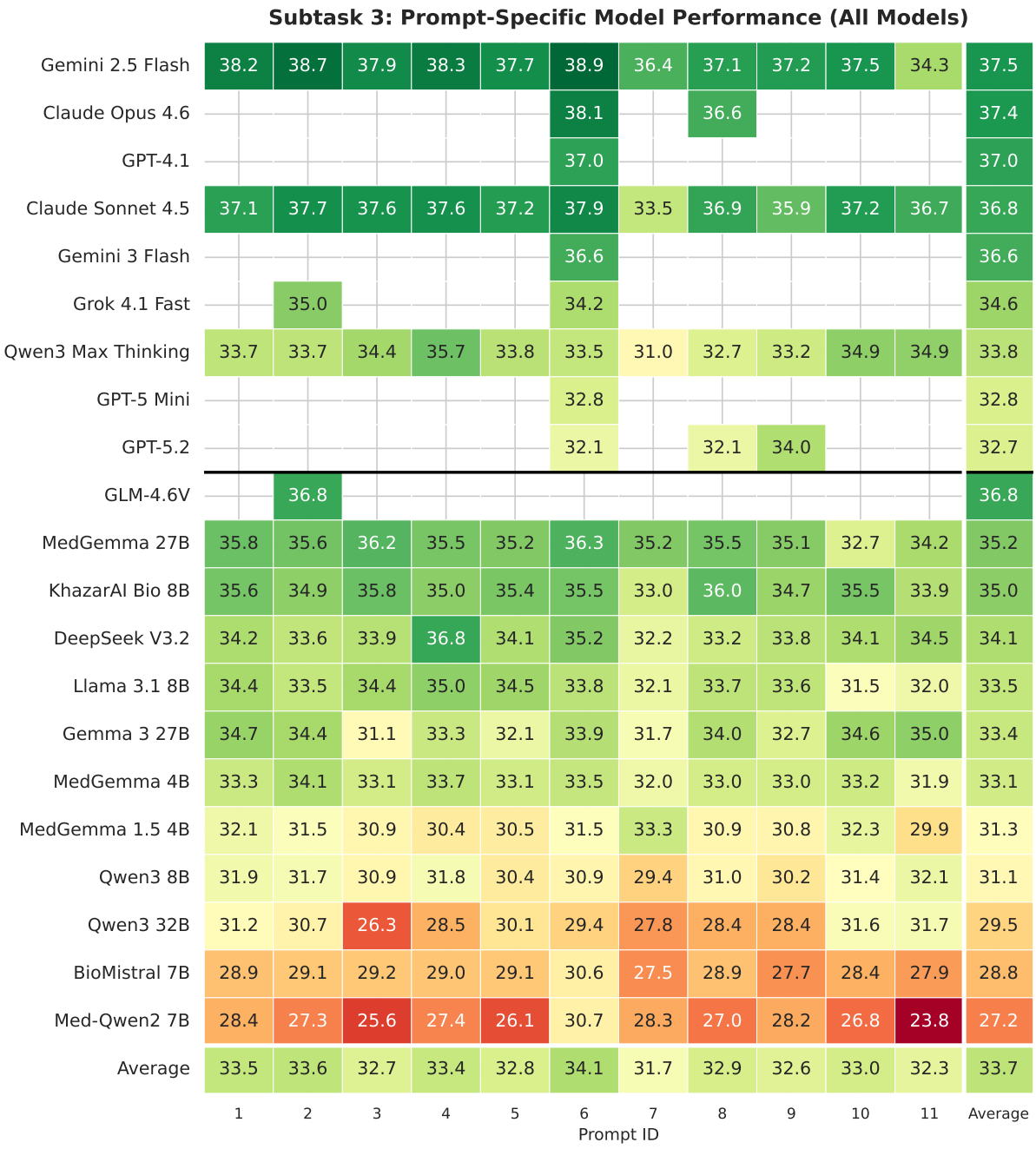

Figure 6: Subtask 3 Validation Results (All Models)

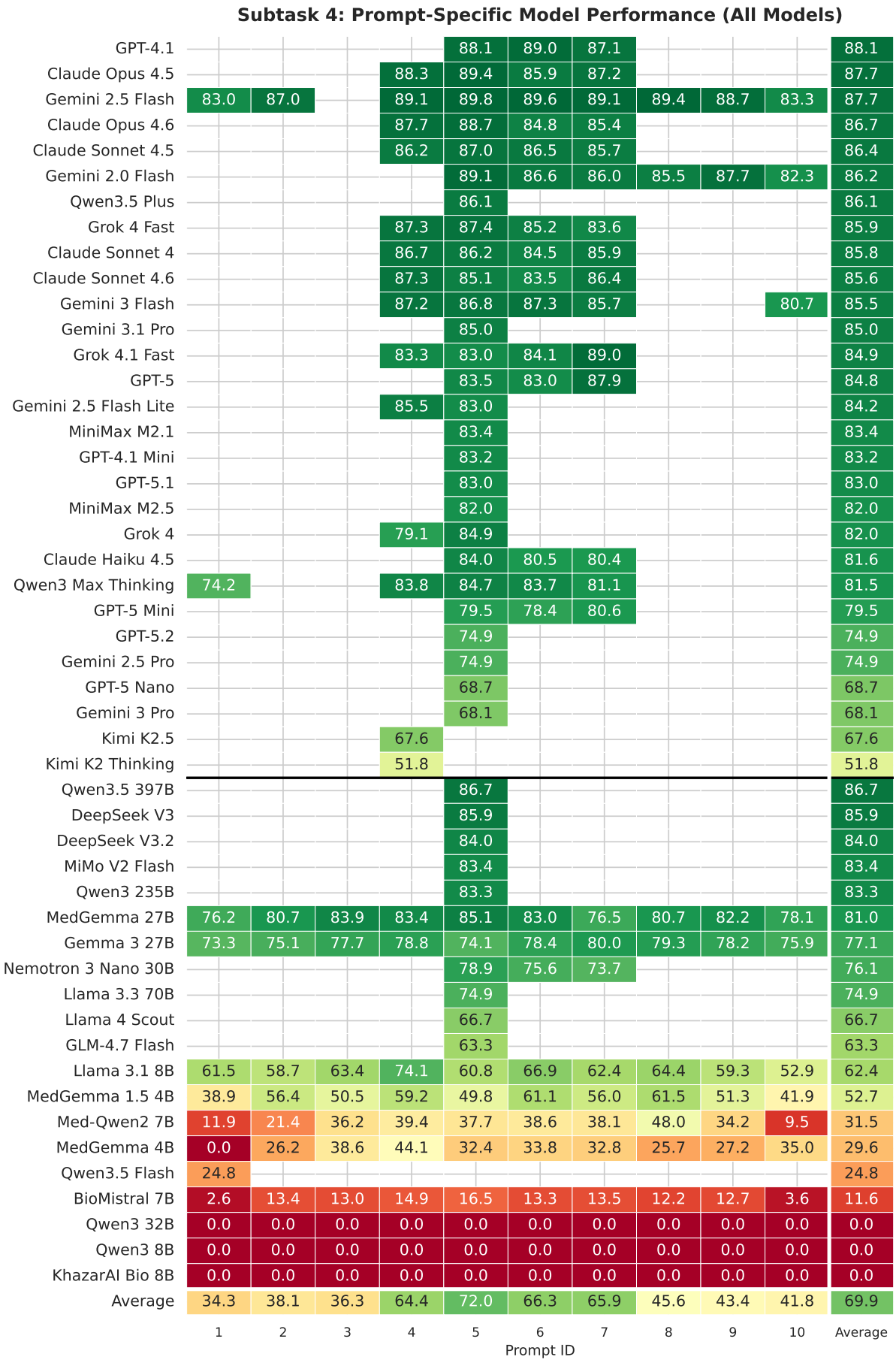

Figure 7: Subtask 4 Validation Results (All Models)