



How Close to the Frontier? Measuring the Fidelity of Deployable Open-weights LLMs to a Frontier Reference for Clinical Phenotype Extraction

Tomi Thomo¹ Alex Thomo^{2*}

¹*Albanian Ministry of Health, Tirana, Albania, and European University of Tirana, Tirana, Albania*

²*Department of Computer Science, University of Victoria, Victoria, BC, Canada*

* Corresponding author's Email: thomo@uvic.ca

Abstract: Frontier large language models read clinical narratives at near expert level, but hospitals often cannot deploy them: privacy regulation, data-residency rules, and on-premises requirements forbid sending patient text to a cloud service. The deployable alternative is a small open-weights model run locally. We ask a scoped fidelity question: how closely do such models reproduce a frontier reference for clinical phenotype extraction (cohort identification from records, not bedside diagnosis)? On a 500-case subset of PMC-Patients and six guideline-defined cardiometabolic conditions, we place seven self-hostable open-weights models between a keyword floor and a frontier ceiling, the consensus of two frontier models. The strongest reach macro-F1 of about 0.95 against the ceiling; under a realistic deployment prompt, a 14B model matches a 70B (no significant difference), and a three-small-model ensemble significantly outscore the standalone 70B, at far lower hardware cost. Once self-hosted, a case carries no per-call cost, so deployment becomes a hardware decision. We release the benchmark, reference, and code.

Keywords: Clinical NLP, phenotype extraction, Open-weights large language models, On-premises deployment, LLM-as-a-judge, Cohort identification, Model scaling, Clinical decision support.

1. Introduction

Deciding, from a patient's notes, whether they have a given condition, a task known as phenotype extraction, is the upstream step for a range of routine tasks: recruiting patients into clinical trials [1, 2], assembling retrospective cohorts, reporting quality measures, and pharmacovigilance. These are population-scale tasks, run in batch over records to assemble or audit a cohort, not point-of-care diagnosis of an individual patient, so the operative benchmark is the best available automated labeler rather than a bedside clinician. The automated method hospitals deploy for this is, very often, keyword search of the notes, sometimes augmented with synonym lists [3–5]. It is cheap, it is deterministic, and it runs anywhere, which is much of why it persists.

Large language models read clinical text far more flexibly. Frontier models now answer medical

questions at or near expert level [6, 7]. The difficulty is that a hospital usually cannot deploy a frontier model where it matters. Privacy regulation such as HIPAA and GDPR, data-residency rules, and on-premises or air-gapped requirements frequently forbid sending patient text to a cloud API, and the case for running open-weights models locally instead is by now well established [8–11]. The model a hospital can actually run is therefore not the frontier model it might want, but a smaller open-weights model hosted on its own hardware. That gap between the desirable and the deployable is the subject of this paper.

We ask a deliberately scoped fidelity question. How closely do deployable open-weights models reproduce a frontier ceiling on clinical phenotype extraction, and is the residual gap small enough for a deployable model to suffice? We treat the consensus of two frontier models as that ceiling, an operational reference rather than a claim of clinical truth, following the established practice of using a strong

model as an evaluation judge [12]; and we treat keyword search, the deployed status quo, as the floor. The deployable models are then placed between the two. Working on a 500-case subset of PMC-Patients [13] and six guideline-defined cardiometabolic conditions, we evaluate seven self-hostable open-weights models spanning roughly 4 to 400 billion parameters.

This paper makes the following contributions.

A three-tier benchmark, released. We place a keyword floor (concept matching with a validated negation and context layer), seven self-hostable open-weights models, and a frontier ceiling (the consensus of two vendors' models, cross-model $\kappa=0.93$) on a common axis, and release the benchmark, the reference, and the code.

Deployable models nearly match the frontier. The strongest self-hostable models reach macro-F1 0.955 to 0.957 against the ceiling, leaving only a small residual gap to the frontier reference.

Quality decouples from size. The curve is not monotone in size. Under a realistic deployment prompt, a 14B model matches a 70B (the difference is not statistically significant), and a three-small-model ensemble (0.919) shows no significant gain over its best member while significantly outscoring the standalone 70B, at a fraction of the hardware.

A cost reframe. On premises there is no per-call API charge, so cost is a fixed hardware question rather than a usage one. Once hardware is priced in, the most accurate model and the most deployable one come apart. The top scorer is a 400B mixture-of-experts that needs a datacenter node, while a 14B model or a small-model ensemble reaches nearly the same accuracy on a single consumer card. The right model therefore depends on the hardware a site can field, and for the many hospitals without a datacenter node it is not the most accurate one.

Why the keyword floor fails. One might expect the floor to look strong, since case reports often state the diagnosis verbatim; instead its dominant failure is under-recall of laboratory-defined diagnoses, and we show with a validated negation library that no keyword rule recovers them.

A deployment risk map. Where the two frontier models disagree, concentrated in the kidney conditions, the open-weights panel is also divided, more than twelve times as often as elsewhere, giving a usable signal for which cases to route to a human.

The remainder of the paper is organized as follows. Section 2 situates the work among related efforts on clinical language models and phenotype extraction. Section 3 describes the corpus, the floor-subjects-ceiling design, the frontier reference, the scoring, and how we treat deployment cost. Section 4

reports where the deployable models land, why the keyword floor fails, and how hardware reorders the ranking, then probes the models' reasoning and prompt sensitivity. Section 5 turns the findings into a deployment recommendation, Section 6 states the limitations and the scope of the fidelity claim, and Section 7 concludes.

2. Related Work

2.1 Clinical phenotyping methods

Identifying a patient's conditions from text is a long-standing task in clinical informatics, and the deployed tradition is rule-based. The standard approach is dictionary and concept matching over vocabularies such as MeSH or UMLS, with negation and context handling, embodied in clinical NLP pipelines such as cTAKES [14] and MetaMap [15]. The eMERGE network [16] and the PheKB algorithm catalog [17] codified these into portable, reusable "computable phenotypes." The i2b2 and n2c2 shared tasks established the benchmark lineage for exactly the conditions we study, recognizing obesity and its comorbidities in discharge summaries [18], extracting heart-disease risk factors over time [19], and selecting patient cohorts against trial criteria [20]. Our keyword floor instantiates this rule-based tradition; the open-weights models we evaluate represent the recent shift to language models, which we measure against it.

2.2 Strong models as evaluation references

Using a capable language model to judge the outputs of other models has become a standard evaluation method. Zheng et al. [12] showed that a frontier model's judgments correlate with human preference at roughly the level of human-to-human agreement, establishing the model-as-judge paradigm we adopt for the ceiling. We adopt it with a deliberately scoped claim. Our frontier consensus is an operational reference against which we measure approximation, not an assertion of clinical truth (Section 5.4). Frontier models perform at or near expert level on related clinical tasks [6, 7], which supports the choice but is not load-bearing for our argument.

2.3 Open-weights LLMs under clinical privacy constraints

The case for running open-weights rather than cloud models on patient data, driven by HIPAA, GDPR, and on-premises requirements, has been made in several recent studies [8–11]. We take that

motivation as background. What those works do not provide, and what this paper contributes, is a characterization of how good the deployable models are relative to a frontier ceiling across the open-weights size range, and of what the deployment decision costs in hardware terms.

2.4 Clinical LLM benchmarks

Broad clinical LLM benchmarks have grown rapidly. BRIDGE [21] evaluates ninety-five models across eighty-seven tasks and nine languages and reports, consistent with our finding, that open models can match proprietary ones. Its breadth is both its contribution and its limit. It is a leaderboard, without a keyword floor, a deployment-cost axis, or a hardware-feasibility analysis. Ours is a complementary study, deep on a single task, with a floor, a per-condition risk map, and procurement guidance. CaseReportBench [22] benchmarks LLMs on case reports but for a different task, dense information extraction.

2.5 LLMs for cohort selection

Closest to our setting, Tai and Tannier [23] evaluate LLMs on the n2c2 cohort-selection challenges. They score against the challenges' existing gold labels, without a keyword baseline or a deployment framing. They also note a gap we take up directly. Prior work running proprietary models on those corpora did not address whether sending the data to a cloud API is consistent with the data's terms of use. Our focus on open-weights models that a hospital can run locally is, in part, a response to that concern.

3. Methods

We frame the evaluation as a spectrum. At one end is the method hospitals deploy today, keyword search of the notes; at the other end is a frontier language model, which is accurate but, as Section 1 argued, often cannot be run on patient data in practice. Between them sit the models that matter for deployment, open-weights models a hospital can run

on its own hardware. We place those models on the spectrum and ask how far up they reach. This section describes the corpus, the three tiers (floor, subjects, ceiling), the scoring, and how we treat deployment cost.

3.1 Corpus and phenotypes

We build on PMC-Patients [13] because it is openly licensed and redistributable, unlike restricted clinical corpora, so we can release the benchmark for reproduction. It is a collection of patient summaries drawn from open-access case reports, from which we take a stratified 500-case subset (random seed 42) balanced so that each target condition has at least fifty positive examples, together with unambiguous negative cases. Working with 500 cases rather than the full corpus keeps the frontier reference (below) affordable to produce while leaving enough positives per condition for stable per-condition estimates.

We study six conditions: type 2 diabetes mellitus, chronic kidney disease, acute kidney injury, heart failure, hypertension, and obesity. We selected them on three grounds. First, they are common enough in case narratives to yield enough positives to measure. Second, each has an explicit, guideline-anchored numerical definition (Table 1), so the operational criterion for "has the condition" is reproducible rather than a matter of taste. Third, they form a coherent cardiometabolic and renal cluster, the comorbidities that co-occur in complex patients and that drive real cohort-identification work. This is a deliberate slice, and a narrow one. It is biased toward conditions that have laboratory or numerical definitions, and it excludes mental-health, oncologic, infectious, respiratory, and rare phenotypes. We return to this scope limit in Section 6. Each (case, condition) pair is an independent binary decision; the 500 cases are the unit of analysis.

3.2 Three tiers

The floor is a keyword rule, the deployed status quo. The subjects are seven self-hostable open-weights models.

Table 1. The six conditions and their operational criteria. A documented prior diagnosis in the narrative, or active disease-specific treatment (e.g., antihypertensive or glucose-lowering therapy), also satisfies the criterion for any condition. Sources: ADA [24], KDIGO [25, 26], ESC [27], ACC/AHA [28], WHO [29]

Condition	Operational criterion
Type 2 diabetes	HbA1c $\geq 6.5\%$, or fasting glucose ≥ 126 mg/dL
Chronic kidney disease	eGFR < 60 mL/min/1.73 m ² for ≥ 3 months, or a chronic damage marker for ≥ 3 months
Acute kidney injury	creatinine rise ≥ 0.3 mg/dL in 48 h, or $\geq 1.5\times$ baseline in 7 d
Heart failure	LVEF $\leq 40\%$, or LVEF $\geq 50\%$ with diastolic dysfunction
Hypertension	office BP $\geq 130/80$ mmHg
Obesity	BMI ≥ 30 kg/m ²

The ceiling is the consensus of two frontier models. The floor and the seven models are all scored against the ceiling, and we report where each model lands between the floor and the ceiling.

3.3 The frontier ceiling

We define the ceiling operationally, as the agreement of two strong models from different vendors. Using a capable model as a reference judge is by now a standard evaluation technique, validated against human preference in general-purpose settings [12]; we adopt it and scope our claims accordingly (Section 5).

The two judges are Claude (claude-opus-4-8) and GPT (gpt-5.5), two cloud-served frontier models from different vendors. Using them is permissible only because PMC-Patients is open published text with no privacy restriction; on restricted data they would be off-limits, which is precisely the constraint that motivates the rest of the paper. Each judge received a single fixed prompt per case that asked, in one turn, for a yes/no/uncertain decision on all six conditions, applying the operational criteria of Table 1 and explicitly discounting family history, ruled-out differentials, negated mentions, and conditions belonging to other people (released with the code). We ran one call per case and froze the outputs.

For each (case, condition) cell, when both judges agree we take their shared answer as the reference label; when they disagree we mark it a hard cell and report it separately rather than discarding it. Across $500 \times 6 = 3000$ cells both judges completed every case with no errors. They agreed on 2940 cells (98.0%), a Cohen's κ of 0.93 [30], leaving 60 hard cells. On the agreed cells the positive rate is 0.177. Per condition, the agreed cells carry 135 hypertension, 105 type 2 diabetes, 84 CKD, 72 AKI, 69 heart-failure, and 55 obesity positives. This cross-vendor agreement is the reference's reliability signal. A high agreement rate could in principle reflect two models sharing a blind spot, so the reference's credibility rests on more than the rate. The judges come from different vendors and training pipelines, so a common systematic error is less likely than within a single model family, and they apply the explicit operational criteria of Table 1 rather than deciding by open-ended judgment, which limits arbitrary, one-off calls (per-condition agreement is reported in Table 6). We therefore treat the resulting labels as an operational reference, adequate for measuring how closely deployable models reproduce a strong model's reading but not a substitute for clinician adjudication, and we report the

60 disagreements as hard cells rather than forcing a label.

For every model we record and release its exact identifier, access tool and version, run date, and parse-failure count alongside the frozen per-case outputs, so the reference is a fixed, citable artifact (Section 6).

3.4 The keyword floor

The floor is the kind of keyword rule used in deployed cohort-identification pipelines. Its concept layer is a case-insensitive, word-boundary search of a per-condition synonym list assembled from MeSH Entry Terms [31], the Grossman Liu medical abbreviation inventory [32], and expert-curated staging and treatment terms. A naive keyword rule, however, fires on negated and contextual mentions ("no history of diabetes," "father had CKD"), so we strengthen it with a validated clinical negation and context algorithm. Each candidate mention is passed through medspaCy's ConText implementation [33–35], which labels it negated, family-history, hypothetical, or historical. A (case, condition) is positive only if it has at least one asserted mention, that is, one that is not negated, not family history, and not hypothetical; a purely historical mention still counts, since the patient did have the condition. We hold the concept layer identical to the naive rule and do not hand-tune the ConText rules, because the point of using a validated library is precisely that its behavior is not tuned to our data. We report this ConText floor as the baseline, and in Section 4 compare it against the naive rule to show how much the added negation handling buys.

3.5 The open-weights panel

The subjects are seven open-weights models that a hospital could host on its own hardware, spanning roughly 4 to 400 billion parameters: Llama 4 Maverick [36], gpt-oss-120b [37], Qwen 2.5 72B [38], Llama 3.3 70B [39], Phi-4 14B [40], MedGemma 4B [41], and Llama 3.1 8B [42]. These were chosen to span the self-hostable size range rather than to be exhaustive, with at least one widely used open model per hardware tier, from laptop-class (4–8B) through single-GPU (14–120B) to a multi-GPU mixture-of-experts (400B), plus one biomedical-tuned model (MedGemma); other strong open families such as Mistral, Gemma 3, and DeepSeek occupy the same axes and are left to future work.

Each was prompted zero-shot at temperature 0, with `max_tokens=1024` and up to four exponential-backoff retries, to make the same per-condition

binary decision. All seven models received one identical prompt, a fixed system instruction (SHA-256 fb1e30c2ffb1; released with the code) plus a per-case user message with the case narrative, the six conditions with their operational criteria, and the required JSON schema. Only the masked-token instruction changes for the redacted runs, so no model is advantaged over another by prompt engineering. This subject prompt is deliberately lighter than the adjudication rubric the frontier judges receive (released with the code). A deploying hospital would prompt a model zero-shot, not hand-engineer the judges' full list of edge-case exclusions, so we hold the subjects to a realistic prompt and reserve the careful rubric for constructing the reference. This asymmetry is conservative by construction. The judges' extra rules almost all withhold a positive call on edge cases like family history, a ruled-out differential, or a single transient reading. A more lightly prompted subject tends to call those edges positive, and on the agreed cells that disagreement counts against it. We test this directly by re-running the panel on the judges' full rubric (Section 4.7): harmonizing the prompt removes nearly a fifth of the panel's false-positive over-calls, concentrated on exactly the edge-heavy conditions the rubric targets, and lifts every model at or above 70B, while leaving the smallest models, which cannot reliably apply the longer rubric, unchanged or slightly worse. The reported light-prompt fidelity is therefore a genuine lower bound for the capable models; for the smallest models the light prompt is their better setting.

We map a model's textual output to a binary call as true $\rightarrow 1$, false/uncertain $\rightarrow 0$ (treating abstention as a negative call, the conservative choice for cohort work); a small number of API errors (empty completions after all retries; 36 cells, 18 each for two of the seven models) are dropped from those two models' scoring rather than forced to a call, so in the main evaluation gpt-oss-120b is scored on 2923 agreed cells and Qwen 2.5 72B on 2922 (one of gpt-oss-120b's 18 error cells fell on a hard cell, already outside the scored set); the prompt-harmonization ablation (Section 4.7) reports its own, separate empty-completion counts.

Six of the seven models were served on PMC-Patients through the OpenRouter inference API; MedGemma 4B we self-hosted on a single H100 GPU. Using a hosted API for these open-access predictions is permitted because it does not affect the deployment argument, which concerns where the weights can run, not how we happened to run them here. For every case we also produced a redacted variant in which each explicit mention of a condition

is replaced by a single masked token, which drives the reasoning-versus-vocabulary probe of Section 4.

3.6 Scoring

We score each model against the frontier reference on the agreed cells, and report the hard cells separately. Predictions are joined to reference labels on the (patient, condition) key, with an assertion that the source identifiers match on every joined row, a guard against a silent misalignment that can otherwise inflate apparent agreement. Our headline metric is macro-F1, the mean of the per-condition F1 scores, which weights each condition equally and is not inflated by the large negative base rate the way raw accuracy is. To express how much of the distance from the floor a model covers, we report the fraction of the gap closed, $(\text{model} - \text{floor}) / (\text{ceiling} - \text{floor})$, where the ceiling is 1.0 by construction (the reference agrees with itself). We attach a 95% confidence interval to every macro-F1 with a case-level bootstrap (2000 resamples of the 500 cases, resampling whole cases so the within-case correlation across the six conditions is preserved), and test the closest model comparisons with the paired bootstrap difference. These comparisons are pre-specified (the consumer-card configurations against the 70B, and the top three against one another), so we report them without a multiplicity correction.

3.7 Deployment cost

Because the deployment scenario is on-premises, the per-case dollar cost of a hosted API would be the wrong axis. Once the weights run on hardware the institution owns, a case carries no metered per-call cost. The cost that remains is the hardware itself. We therefore characterize each model by the GPU class needed to host it, which we derive directly from its parameter count and architecture (dense versus mixture-of-experts) as the weight memory at the precision each model was evaluated at, noting that key-value cache adds to this floor at inference time. We size the resident weights as the total parameter count times the bytes per parameter at the evaluated precision, two bytes at bf16, one byte at fp8, and about half a byte at 4-bit (a little more for the MXFP4 format, whose block scales add metadata, so gpt-oss-120b's 120 billion parameters occupy about 64 GB rather than 60); a mixture-of-experts model keeps all of its experts resident, so routing does not lower this floor, and the key-value cache is additional. Our deployment claim is therefore specifically one of memory feasibility: whether a model's weights, plus its key-value cache, fit within the memory of a given hardware tier, which is the constraint that decides

whether a site can host the model at all. We make no claim about the other axes of serving performance. Latency, throughput, batch size, maximum context length, power draw, quantization degradation, failure rates, and the choice of serving stack all depend on deployment engineering and are out of scope here; they govern how well a feasible model runs, not whether it can be hosted.

4. Results

4.1 How far up the spectrum the deployable models reach

The keyword floor reaches a macro-F1 of 0.797 against the frontier reference; the ceiling is 1.0 by construction. Table 2 places the seven open-weights models between them. The three largest, Llama 4 Maverick, gpt-oss-120b, and Qwen 2.5 72B, cluster at 0.955 to 0.957, closing roughly four-fifths of the distance from the floor to the ceiling. In absolute terms they sit just under the ceiling, within roughly four macro-F1 points of perfect agreement with the two frontier models. Their pairwise differences are not statistically significant (Table 2), so the ordering among the top three is not statistically resolved and should not be read as a ranking.

Quality does not, however, track size cleanly. Phi-4, at 14 billion parameters, scores 0.906, in fact higher than the five-times-larger Llama 3.3 70B's 0.888, though the 0.018 gap is not statistically significant (paired difference in Phi-4's favour, 95% CI [-0.01,0.05]). The lesson is not that the smaller model is reliably the better of the two, but that five times the parameters, and the hardware to run them, buy no measurable gain here. The smallest two models, a 4B and an 8B, sit together near 0.85, closing under a third of the gap. The practical reading is that there is a usable knee in the curve around 14B. Below it, quality drops noticeably; above it, returns to scale are small.

4.2 The floor it replaces fails by missing, not over-calling

It is worth examining what the keyword floor actually gets wrong, because the common assumption is that keyword search over-recruits, firing on conditions the patient does not have. Across the full benchmark the opposite is true. Against the frontier reference the ConText floor has a precision of 0.931 but a recall of only 0.723. Its errors are dominated by 144 false negatives against 28 false positives. The keyword rule's real weakness is that it misses patients, specifically those whose diagnosis is established by laboratory values, medications, or numerical criteria without the condition being named. Per condition (Fig. 1), the floor recovers only about half of the true type 2 diabetes and chronic kidney disease cases (recall near 0.5 for each), the two conditions most often defined by a lab value (HbA1c, eGFR) rather than a word.

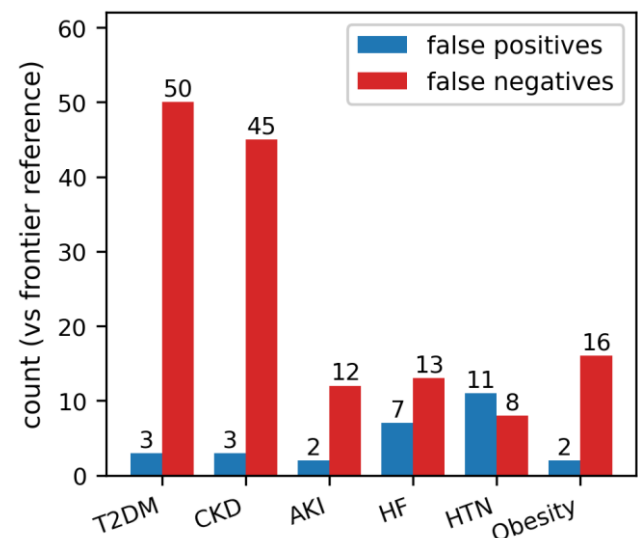


Figure. 1 The keyword floor's errors against the frontier reference, by condition. False negatives (missed diagnoses) dominate false positives (over-calls), most heavily for the laboratory-defined conditions, type 2 diabetes and chronic kidney disease

Table 2. Open-weights panel against the frontier ceiling (macro-F1 on n=2940 agreed cells; gpt-oss-120b and Qwen 2.5 72B are scored on 2923 and 2922 agreed cells respectively; 95% CIs from a case-level bootstrap). %-gap is the fraction of the keyword-floor-to-frontier distance closed, computed from unrounded macro-F1 with a floor of 0.7968 (shown rounded as 0.797). The floor is the validated ConText keyword rule

Model	macro-F1	95% CI	%-gap closed
Llama 4 Maverick	0.957	[.943,.970]	79.1
gpt-oss-120b	0.956	[.940,.968]	78.1
Qwen 2.5 72B	0.955	[.941,.968]	78.0
Phi-4 14B	0.906	[.886,.924]	53.8
Llama 3.3 70B	0.888	[.860,.914]	45.0
MedGemma 4B	0.852	[.826,.877]	27.3
Llama 3.1 8B	0.850	[.827,.871]	26.1
keyword floor (ConText)	0.797	—	0.0

Over-calling is real but secondary, and it is concentrated in hypertension, where the term appears in non-patient contexts. Adding negation handling helps, but only at the margin, and we can say exactly how much. A naive concept-match floor scores macro-F1 0.790; a homemade NegEx-style rule lifts it to 0.800; the validated ConText rule, which also handles family-history and hypothetical mentions, lands at 0.797. ConText removes 26 of the naive rule's 54 false positives (most through negation, a handful through family history), leaving 28 that come from contexts no negation layer addresses, ruled-out differentials, drug-class mentions, and risk-score components. It also over-suppresses 11 true positives, which is why its slightly more aggressive scope trades a little recall for the cleaner precision. The conclusion is robust across all three floors and is the point that matters. Negation and context account for a minority of the keyword rule's error budget, and none of them touch the 144 false negatives (Fig. 1). Better string matching cannot recover diagnoses the string never contains.

4.3 Once self-hosted, the choice is about hardware

On premises, running a case carries no per-call API cost for every model in the panel, so the deployment decision is not which model is cheapest per call but which model a site can host at all. Table 3 reports the GPU class each model needs, derived from its parameter count and architecture. Overlaying this on the quality ranking reorders it (Fig. 2).

Llama 4 Maverick, which tops Table 2 at 0.957, is a 400-billion-parameter mixture-of-experts model. Although only a few of its experts activate per token, a serving stack must keep the full expert set resident in memory, so the footprint is set by the total parameter count, not the active one: its weights occupy roughly 400 GB at the fp8 precision we evaluated, and about 200 GB even quantized to 4-bit, so hosting it needs a multi-GPU node either way; for

most institutions it is not an on-premises option at all. Its 0.001 advantage over gpt-oss-120b therefore buys nothing deployable. gpt-oss-120b, at the same quality, fits on a single 80 GB accelerator because its 120 billion total parameters occupy only about 64 GB in native 4-bit (MXFP4), making it the strongest single-GPU choice.

The sharpest case against size is Llama 3.3 70B, which needs the same single 80 GB GPU as Qwen yet scores 0.888, below the 0.906 that Phi-4 14B reaches on a 32 GB consumer card. It is both more expensive to host and less accurate. On that consumer card the best options are Phi-4 14B and, as the next subsection shows, an ensemble of small models.

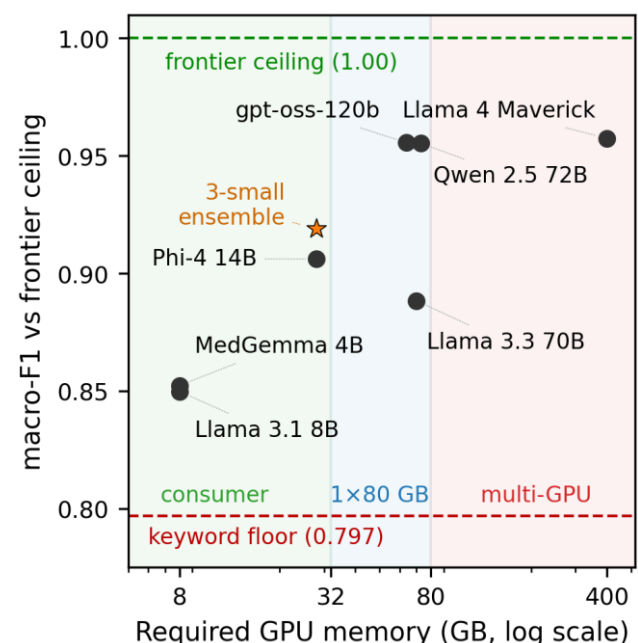


Figure. 2 Deployable quality against hardware footprint. Each open-weights model is placed by the GPU memory it needs at its evaluated precision (horizontal axis, log scale) and its macro-F1 against the frontier ceiling (vertical axis); the keyword floor and ceiling are dashed lines, and the star marks the three-small-model ensemble

Table 3. Hardware feasibility. Each row is the model at the listed precision: its macro-F1 (Table 2) and its VRAM footprint both reflect that precision, so they describe one deployable artifact. Key-value cache adds to the weight memory at inference. Models are ordered by footprint, the axis a site provisions against

Model	macro-F1	VRAM (prec.)	Host tier
Llama 4 Maverick	0.957	~400 GB (fp8)	multi-GPU node
Qwen 2.5 72B	0.955	~73 GB (fp8)	single 80 GB GPU
Llama 3.3 70B	0.888	~70 GB (fp8)	single 80 GB GPU
gpt-oss-120b	0.956	~64 GB (fp4)	single 80 GB GPU
Phi-4 14B	0.906	~28 GB (bf16)	single 32 GB consumer GPU
Llama 3.1 8B	0.850	~8 GB (fp8)	laptop / consumer
MedGemma 4B	0.852	~8 GB (bf16)	laptop

4.4 Do the models reason, or match keywords?

To separate clinical reasoning from keyword spotting we re-scored each model on the redacted narratives, in which every explicit mention of a condition is masked, against the same full-text reference. Table 4 reports the result. The key finding is that the models keep working without the word. With every disease name removed, they still reach 0.56 to 0.77 macro-F1, retaining most of their full-text accuracy. They are recognizing the diagnosis from its evidence, a lab value, a medication, a described complication, rather than from the literal term, and that inference is precisely what lets them catch the cases a keyword search misses.

The models do also use the name when it is present. The masked score falls by 0.17 to 0.30, the part of their accuracy that rides on vocabulary rather than reasoning. The size of that gap is our best available predictor of transfer to real records, where conditions are phrased differently than in case reports, a model that holds up when the canonical term is removed should degrade less under that distribution shift. Llama 4 Maverick reasons most independently of the keyword, holding 0.774 with the highest full-versus-redacted agreement (0.935); the smaller models lean on it more.

One number needs a caveat. gpt-oss-120b shows the largest drop (0.303) despite its size, because it abstains heavily and masking pushes it further toward abstention, which our conservative scoring counts as a negative; its drop reflects that behavior as much as lost reasoning. The same convention applies on the full text, so gpt-oss's leaderboard score is, if anything, a slight underestimate of its underlying accuracy.

4.5 Where the models are weakest, and which cases to hand back

The per-condition view (Fig. 3) localizes where the smaller models are weakest, and the pattern is clinically coherent rather than random. Their soft spot

is concentrated on the two kidney conditions. Llama 3.1 8B drops to 0.73 on acute kidney injury, and chronic kidney disease is the next weakest across the small models. Both require reasoning over time, an acute creatinine rise relative to baseline, or chronicity over months, the hardest signal to extract from a narrative. That coherence is what makes the limitation manageable rather than worrying. The small models do not fail unpredictably, they miss in one identifiable place, so a deployment can route the kidney cases to a clinician and rely on the model everywhere else.

The same conditions dominate the cells where the two frontier models disagree. Of the 60 hard cells, 24 are acute kidney injury and 13 are chronic kidney disease, so the kidney conditions account for 62% of all disagreements, against only three for hypertension. These disagreements reflect genuine ambiguity rather than error. In 43 of the 60, one frontier model hedges to uncertain while the other commits, and only 17 are outright yes-versus-no contradictions.

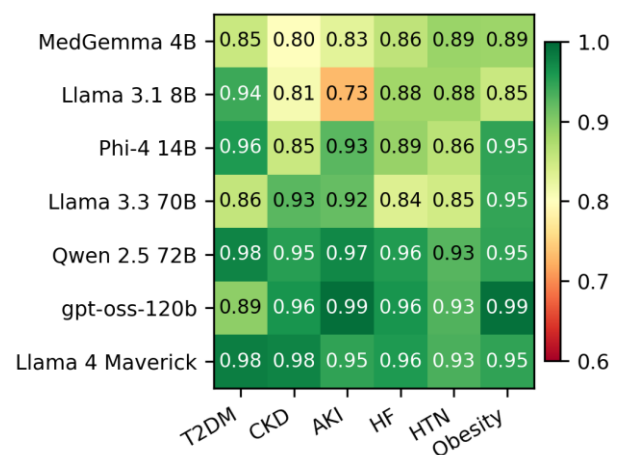


Figure. 3 Per-condition F1 against the frontier reference, models ordered from smallest (top) to largest (bottom). The smaller models are weakest on the kidney conditions (CKD, AKI), where Llama 3.1 8B reaches only 0.73 on AKI; the larger models handle them well

Table 4. Reasoning-versus-vocabulary probe. Full and redacted F1 are macro-F1 against the full-text reference on the agreed cells; the redacted run masks every condition mention. Agreement is the full-versus-redacted prediction agreement, measured across all cells

Model	full	redacted	drop	agree
Llama 4 Maverick	0.957	0.774	0.183	0.935
Qwen 2.5 72B	0.955	0.748	0.207	0.928
Llama 3.3 70B	0.888	0.722	0.167	0.920
Phi-4 14B	0.906	0.696	0.210	0.899
Llama 3.1 8B	0.850	0.680	0.169	0.920
gpt-oss-120b	0.956	0.653	0.303	0.902
MedGemma 4B	0.852	0.564	0.289	0.904

Most importantly, frontier disagreement marks cases the whole field finds hard, not just the two judges. On the hard cells the seven open-weights models are themselves divided, that is, between a fifth and four-fifths of them call the cell positive, 55% of the time; on the agreed cells they are divided only 4% of the time, more than a twelvefold difference (Fig. 4). This turns the reference's by-product into an operating procedure. A deployment cannot run the frontier judges, but the open panel's own division flags the same hard cells, kidney conditions above all, and those are the cells to route to a human rather than trust to any single model.

4.6 A committee of small models

Finally, because small models are cheap to run on modest hardware, we asked whether several of them voting can do better than any one. A majority vote of the three small models, MedGemma 4B, Llama 3.1 8B, and Phi-4 14B, which run on a single 32 GB card when loaded one at a time, reaches macro-F1 0.919 and closes 60% of the gap. It does not significantly improve on its strongest member (Phi-4 at 0.906; the paired difference falls within the bootstrap CI), so voting adds no significant gain over the best single small model. Its distinct value is elsewhere. The ensemble is the one consumer-card option that significantly clears the standalone Llama 3.3 70B (0.888; difference 0.031, 95% CI [0.002,0.062]), which Phi-4 alone does not significantly beat, and on a fraction of that model's hardware. The ensemble does not match the large-model cluster at 0.956, but it closes most of the remaining distance using only hardware a clinic can own.

4.7 The prompt asymmetry is conservative, not inflating

Because the subjects are scored on a lighter prompt than the frontier judges received, one might worry that the reported gap to the ceiling is partly an artifact of under-prompting. To check, we re-ran all seven models on the judges' full adjudication rubric, the same operational criteria and edge-case exclusions the judges saw, at temperature 0 with every other setting held fixed, and rescored against the same reference. Table 5 reports the result.

The error profile moves exactly as the conservativeness argument predicts. Pooled across the panel, harmonizing the prompt cuts false-positive over-calls from 490 to 397, a 19% reduction, while recall is essentially unchanged (0.939 to 0.938); the entire gain is in precision (0.874 to 0.895). The reduction concentrates on the two conditions whose edges the rubric explicitly excludes: hypertension

false positives fall from 183 to 117 once the rubric rules out a single transient reading during an acute event, and acute-kidney-injury false positives fall from 72 to 28 once it supplies the numeric creatinine criterion. This is the over-calling the light prompt was expected to induce, now measured.

The effect on macro-F1, however, tracks model capability. Every model at or above 70 billion parameters improves, the weakest of them, Llama 3.3 70B, by the most (+0.078, to 0.966), because a larger model can apply the thresholds it is handed. The heavier rubric also raised Llama 3.3 70B's empty-completion rate to 96 of 3000 cells (none under the light prompt); 95 of them are agreed cells, so its harmonized figure is scored on 2845 of the 2940 agreed cells. gpt-oss-120b likewise left 24 cells empty under the harmonized rubric (21 agreed), so its harmonized figure is scored on 2919; no other model left any.

The three smallest models do not benefit: Llama 3.1 8B is flat (+0.003), while Phi-4 14B and MedGemma 4B edge down (−0.045 and −0.047), because a rubric a small model cannot follow reliably trades its removed over-calls for new mistakes.

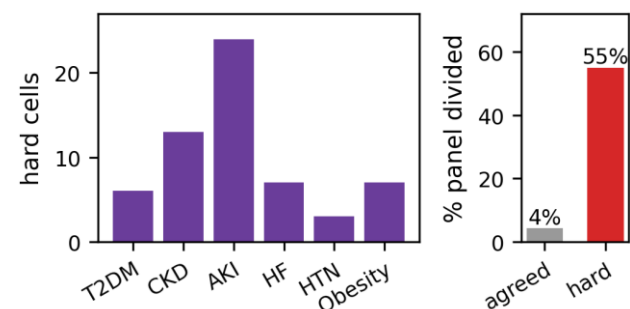


Figure. 4 Left: the 60 frontier-disagreement hard cells by condition, dominated by acute kidney injury and chronic kidney disease. Right: the open-weights panel is itself divided on 55% of the hard cells against 4% of the agreed cells, so frontier disagreement flags the cases the whole field finds hard

Table 5. Prompt-harmonization ablation: macro-F1 when each model is re-run on the frontier judges' full adjudication rubric instead of the lighter deployment prompt (up to n=2940 agreed cells, fewer where a model left cells empty; temperature 0, all other settings fixed). Δ is harmonized minus light

Model	Light	Harmonized	Δ
Llama 4 Maverick	0.957	0.968	+0.011
gpt-oss-120b	0.956	0.984	+0.028
Qwen 2.5 72B	0.955	0.968	+0.013
Phi-4 14B	0.906	0.861	−0.045
Llama 3.3 70B	0.888	0.966	+0.078
MedGemma 4B	0.852	0.805	−0.047
Llama 3.1 8B	0.850	0.853	+0.003
Panel mean	0.909	0.915	+0.006

No harmonized score exceeds the ceiling. The reading for the leaderboard is therefore twofold: for the capable models the light-prompt scores are a genuine lower bound that a fuller prompt only raises, and for the small models the light prompt is already their better setting, so in neither regime does the reported fidelity overstate what an open model delivers. This also scopes the size-independence result: a 14B matching a 70B is specific to the light deployment prompt (under the matched rubric the 70B, 0.966, separates from Phi-4 14B, 0.861), which is the regime a hospital would actually run.

5. Discussion

5.1 What a hospital should deploy

The deployment in view is cohort identification over records, not bedside diagnosis. The results translate into a procurement ladder rather than a single winner, because the binding constraint is hardware, not accuracy. A site with no GPU budget can run MedGemma 4B or Llama 3.1 8B on a laptop and recover roughly 0.85 macro-F1 of the frontier ceiling. A site with one consumer 32 GB card can run Phi-4 14B (0.906) or a majority vote of three small models (0.919, loaded one at a time), statistically no better than Phi-4 alone yet the one consumer-card option that significantly clears the standalone 70B; this is the sweet spot, because it closes most of the gap on hardware a clinic can actually own. A site with one datacenter accelerator should run gpt-oss-120b (0.956), the strongest model that fits on a single 80 GB GPU. The model that tops the raw quality ranking, Llama 4 Maverick, is the one we would not recommend deploying. Its marginal advantage over gpt-oss-120b is 0.001, and claiming it requires a multi-GPU node most institutions will not field. Feasibility, not peak score, should decide. And because on-premises inference carries no per-case API fee, the decision is a one-time capital choice about which accelerator to buy, not a recurring per-case cost.

5.2 Hand the hard cells back

A deployment does not have to treat every case the same way. Frontier disagreement is a reliable flag for cases the whole model field finds hard. On those cells the open-weights panel is divided more than twelve times as often as on the agreed cells. The disagreements concentrate on the kidney conditions, which need temporal reasoning that current small models do not reliably perform. The operating procedure that follows is concrete. Let the model make the call on the easy majority of cells, and route

to human review the cells where an ensemble of models splits, which fall overwhelmingly on the kidney conditions. This converts the benchmark from a score into a triage rule, and it keeps the human effort where it changes outcomes rather than spread thinly across cases the model already handles.

5.3 Keyword search under-recruits more than it over-recruits

Our floor analysis corrects a common framing. The standard concern about keyword cohort identification is over-recruitment. A chart mentions a condition the patient does not have, and the rule fires. Across the full benchmark that is the smaller problem. The keyword rule's dominant failure is under-recall, it misses patients whose diabetes or kidney disease is documented through a lab value or a medication but never named, missing about half of those two conditions. A validated negation and context library removes fewer than half of the false positives and cannot recover any of the false negatives, because no amount of string matching finds a diagnosis the string does not contain. The value of moving from a keyword rule to a language model is therefore mostly in recall, in catching the cases the keyword misses, and that value is robust to a smarter keyword baseline. The redaction probe identifies the mechanism. The gain comes from models reasoning past the literal term, which is exactly what is needed to recognize a diagnosis from its evidence.

5.4 On using a frontier model as the reference

We have been deliberate about what the frontier reference is and is not. It is an operational ceiling, the agreement of two strong models from different vendors, and our claim is about how closely deployable models approximate that ceiling, not about clinical truth.

The ceiling is not an arbitrary target. Frontier models perform at or near expert level on medical question answering [6, 7, 43], and in head-to-head studies they have matched or exceeded practicing physicians on diagnostic reasoning and consultation [44, 45]. That is exactly what makes approximating them worth measuring. We treat their consensus as a strong operational reference, not as infallible clinical ground truth, and the fidelity question does not require it to be.

The agreement is robust in a stronger sense than the headline rate. Because most cells are negative (fewer than one in five is positive), a high overall figure could in principle be carried by the easy negatives; it is not. Agreement on the positive calls alone is 0.97, and per-condition Cohen's κ stays

between 0.92 and 0.99 across all six conditions (Table 6), including the temporally hardest, acute kidney injury. The two judges agree not only often but on the cells that carry information and on the conditions that are hard to call.

5.5 Relation to prior clinical LLM benchmarks

A natural question is how these numbers compare with the recent wave of clinical LLM benchmarks. Because this is a measurement and deployability study of one task, not a method competing for rank on a shared benchmark, the relevant comparison is not a cross-benchmark leaderboard but an internal, like-for-like one: seven open-weights models scored on one common set of cases, labels, prompt, and metric, against a keyword floor and a frontier ceiling. A cross-benchmark accuracy comparison would instead mix different tasks, corpora, and label definitions, and would not be interpretable. What is comparable is the study's shape, and Table 7 places it against the closest prior work on the axes that matter for deployment. Two features are ours alone in this set: a deployed keyword floor, without which an accuracy number has no baseline to be judged against, and an explicit hardware-feasibility axis that turns a quality ranking into a procurement decision. The broad leaderboards [21] and the case-report and cohort-selection benchmarks [22, 23] contribute breadth or human-gold rigor that we do not; none of them pairs a floor with a per-model hardware constraint, which is the specific gap this paper fills. A head-to-head on those corpora's gold labels is a natural extension of this work.

6. Limitations

6.1 Fidelity, not clinical accuracy

The estimand of this study is fidelity to a frontier-model reference, not clinical accuracy against clinician-adjudicated labels; we report how closely deployable models reproduce that reference by construction, and human adjudication would answer a distinct, complementary question rather than a missing part of the one posed here.

6.2 Phenotype scope

The six conditions are a deliberate cardiometabolic slice with explicit numerical definitions (Section 3.1); they exclude mental-health, oncologic, infectious, respiratory, and rare phenotypes. The benchmark therefore speaks to guideline-defined cardiometabolic phenotype extraction, not to clinical phenotyping in general.

6.3 Case reports are not electronic health records

PMC-Patients consists of published case narratives, which are cleaner and more complete than the fragmented, abbreviated, and noisier notes of a real record. The redaction-robustness axis is only an indirect proxy for how models would transfer to such notes, and it does not capture the misspellings, templated boilerplate, and missing context that characterize real EHR text. We therefore do not claim these fidelity numbers transfer to clinical notes unchanged; validating them on real records or shared-task clinical corpora (for example n2c2 or MIMIC) is future work.

6.4 Threshold choices

The operational criteria of Table 1 encode specific guideline thresholds (for example, office BP $\geq 130/80$ mmHg); a different but defensible cutoff would move some labels, so absolute scores should be read against this rule set rather than as universal.

Table 6. Per-condition inter-judge agreement (the two frontier judges) on the present/absent call, over the 500 cases per condition. κ is Cohen's κ ; positive agreement is the base-rate-robust agreement on positive calls. The overall present/absent κ is 0.96 (0.93 across the three yes/no/uncertain categories, Section 3.3); overall positive agreement is 0.97

Condition	Positives	κ	Pos. agr.
T2DM	105	0.97	0.98
CKD	84	0.96	0.97
AKI	72	0.92	0.93
HF	69	0.98	0.98
HTN	135	0.99	0.99
Obesity	55	0.95	0.96

Table 7. Positioning against the closest prior clinical LLM benchmarks. The last column marks whether a study pairs a deployed keyword floor (a baseline to measure accuracy against) with a per-model hardware-feasibility analysis (what deployment costs); only this study provides both

Study	Scope	Data	F+H
This work	6 conditions (depth)	PMC-Patients (open)	Yes
BRIDGE [21]	87 tasks (breadth)	Real-world clinical	No
CaseReportBench [22]	Dense extraction	Case reports	No
Tai & Tannier [23]	Cohort selection	n2c2 (restricted)	No
i2b2/n2c2 tasks [18–20]	Phenotyping	Discharge notes (restricted)	No

7. Conclusion

We asked how closely a model a hospital can actually host reproduces a frontier ceiling on clinical phenotype extraction. For six guideline-defined cardiometabolic conditions, it comes close. The strongest self-hostable open-weights models nearly match the frontier reference, and quality does not track parameter count. Once the binding cost is on-premises hardware rather than per-call price, the model a site should run is rarely the largest. We also found that the keyword rule these models replace fails mainly by missing laboratory-defined diagnoses, which no negation layer repairs. The cases where two frontier models disagree, concentrated in the kidney conditions, form a usable queue for human review. We release the benchmark, the frontier reference, and all code. Validating the reference against human-annotated gold and extending beyond cardiometabolic phenotypes are the natural next steps.

Data and Code Availability

The 500-case PMC-Patients subset and its stratification seed, the frozen frontier-reference labels and the 60 hard cells, the ConText keyword-floor predictions, the per-model outputs (full and redacted variants), and all analysis and figure-generation code are available at <https://github.com/thomouvic/pmc-llm>.

Generative AI use

We used Claude Code to assist with experiment scripting and prose editing. All contributions are authors' own. The authors reviewed all AI-assisted content and take full responsibility for the accuracy and integrity of the paper.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, A. Thomo and T. Thomo; methodology, software, and formal analysis, A. Thomo and T. Thomo; clinical criteria and validation, T. Thomo; writing—original draft, A. Thomo; writing—review and editing, T. Thomo and A. Thomo; supervision, A. Thomo.

Acknowledgments

We thank the maintainers of PMC-Patients, the Grossman Liu 2021 Medical Abbreviation Meta-

Inventory, and the medspaCy project for releasing those artifacts under reuse-permissive licenses.

References

- [1] P. J. Embi, A. Jain, J. Clark, S. Bizjack, R. Hornung, and C. M. Harris, "Effect of a clinical trial alert system on physician participation in trial recruitment", *Archives of Internal Medicine*, Vol. 165, No. 19, pp. 2272–2277, 2005, doi: 10.1001/archinte.165.19.2272.
- [2] A. Sertkaya, H.-H. Wong, A. Jessup, and T. Beleche, "Key cost drivers of pharmaceutical clinical trials in the United States", *Clinical Trials*, Vol. 13, No. 2, pp. 117–126, 2016, doi: 10.1177/1740774515625964.
- [3] C. Shivade, *et al.*, "A review of approaches to identifying patient phenotype cohorts using electronic health records", *Journal of the American Medical Informatics Association*, Vol. 21, No. 2, pp. 221–230, 2014, doi: 10.1136/amiajnl-2013-001935.
- [4] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (EHRs): A survey", *ACM Computing Surveys*, Vol. 50, No. 6, p. 85, 2018, doi: 10.1145/3127881.
- [5] D. A. Hanauer, Q. Mei, J. Law, R. Khanna, and K. Zheng, "Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE)", *Journal of Biomedical Informatics*, Vol. 55, pp. 290–300, 2015, doi: 10.1016/j.jbi.2015.05.003.
- [6] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems", *arXiv Preprint*, arXiv:2303.13375, 2023, doi: 10.48550/arXiv.2303.13375.
- [7] K. Singhal, *et al.*, "Large language models encode clinical knowledge", *Nature*, Vol. 620, No. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.
- [8] I. C. Wiest, D. Ferber, J. Zhu, *et al.*, "Privacy-preserving large language models for structured medical information retrieval", *npj Digital Medicine*, Vol. 7, No. 1, p. 257, 2024, doi: 10.1038/s41746-024-01233-2.
- [9] L. Builtjes, J. Bosma, M. Prokop, B. van Ginneken, and A. Hering, "Leveraging open-source large language models for clinical information extraction in resource-constrained settings", *JAMIA Open*, Vol. 8, No. 5, p.

- ooaf109, 2025, doi: 10.1093/jamiaopen/ooaf109.
- [10] D. J. Spaanderman, *et al.*, “Evaluating open-weight large language models for structured data extraction from narrative medical reports across multiple use cases and languages”, *arXiv Preprint*, arXiv:2511.10658, 2025, doi: 10.48550/arXiv.2511.10658.
 - [11] S. Nowak, J.-F. Laß, N. Mesrobian, *et al.*, “Secure on-premise deployment of open-weights large language models in radiology: An isolation-first architecture with prospective pilot evaluation”, *arXiv Preprint*, arXiv:2604.22768, 2026, doi: 10.48550/arXiv.2604.22768.
 - [12] L. Zheng, *et al.*, “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena”, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 46595–46623, 2023, doi: 10.52202/075280-2020.
 - [13] Z. Zhao, Q. Jin, F. Chen, T. Peng, and S. Yu, “A large-scale dataset of patient summaries for retrieval-based clinical decision support systems”, *Scientific Data*, Vol. 10, No. 1, p. 909, 2023, doi: 10.1038/s41597-023-02814-8.
 - [14] G. K. Savova, *et al.*, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications”, *Journal of the American Medical Informatics Association*, Vol. 17, No. 5, pp. 507–513, 2010, doi: 10.1136/jamia.2009.001560.
 - [15] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: Historical perspective and recent advances”, *Journal of the American Medical Informatics Association*, Vol. 17, No. 3, pp. 229–236, 2010, doi: 10.1136/jamia.2009.002733.
 - [16] O. Gottesman, *et al.*, “The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future”, *Genetics in Medicine*, Vol. 15, No. 10, pp. 761–771, 2013, doi: 10.1038/gim.2013.72.
 - [17] J. C. Kirby, *et al.*, “PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability”, *Journal of the American Medical Informatics Association*, Vol. 23, No. 6, pp. 1046–1052, 2016, doi: 10.1093/jamia/ocv202.
 - [18] Ö. Uzuner, “Recognizing obesity and comorbidities in sparse data”, *Journal of the American Medical Informatics Association*, Vol. 16, No. 4, pp. 561–570, 2009, doi: 10.1197/jamia.M3115.
 - [19] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner, “Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2”, *Journal of Biomedical Informatics*, Vol. 58, pp. S67–S77, 2015, doi: 10.1016/j.jbi.2015.07.001.
 - [20] A. Stubbs, M. Filannino, E. Soysal, S. Henry, and Ö. Uzuner, “Cohort selection for clinical trials: n2c2 2018 shared task track 1”, *Journal of the American Medical Informatics Association*, Vol. 26, No. 11, pp. 1163–1171, 2019, doi: 10.1093/jamia/ocz163.
 - [21] J. Wu, *et al.*, “BRIDGE: Benchmarking large language models for understanding real-world clinical practice text”, *Nature Biomedical Engineering*, 2026, doi: 10.1038/s41551-026-01719-2.
 - [22] X. Y. C. Zhang, C. R. Ferreira, F. Rossignol, R. T. Ng, W. Wasserman, and J. Zhu, “CaseReportBench: An LLM benchmark dataset for dense information extraction in clinical case reports”, In: *Proc. of The Sixth Conference on Health, Inference, and Learning*, pp. 527–542, 2025.
 - [23] C.-E. A. Tai and X. Tannier, “Clinical trial cohort selection using large language models on n2c2 challenges”, *arXiv Preprint*, arXiv:2501.11114, 2025, doi: 10.48550/arXiv.2501.11114.
 - [24] American Diabetes Association Professional Practice Committee, “Standards of care in diabetes—2024”, *Diabetes Care*, Vol. 47, No. Suppl. 1, pp. S1–S321, 2024.
 - [25] KDIGO CKD Work Group, “KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease”, *Kidney International*, Vol. 105, No. 4S, pp. S117–S314, 2024.
 - [26] KDIGO AKI Work Group, “KDIGO clinical practice guideline for acute kidney injury”, *Kidney International Supplements*, Vol. 2, No. 1, pp. 1–138, 2012.
 - [27] T. A. McDonagh, M. Metra, M. Adamo, *et al.*, “2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure”, *European Heart Journal*, Vol. 42, No. 36, pp. 3599–3726, 2021, doi: 10.1093/eurheartj/ehab368.
 - [28] P. K. Whelton, R. M. Carey, W. S. Aronow, *et al.*, “2017 ACC/AHA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults”, *Journal of the American College of Cardiology*, Vol. 71, No. 19, pp. e127–e248, 2018, doi: 10.1016/j.jacc.2017.11.006.
 - [29] World Health Organization, “Obesity and overweight (BMI classification)”, 2024,

- [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
- [30] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
- [31] C. E. Lipscomb, "Medical subject headings (MeSH)", *Bulletin of the Medical Library Association*, Vol. 88, No. 3, pp. 265–266, 2000.
- [32] L. Grossman Liu, *et al.*, "A deep database of medical abbreviations and acronyms for natural language processing", *Scientific Data*, Vol. 8, No. 1, p. 149, 2021, doi: 10.1038/s41597-021-00929-4.
- [33] H. Eyre, *et al.*, "Launching into clinical space with medspaCy: A new clinical text processing toolkit in Python", In: *Proc. of AMIA Annual Symposium*, pp. 438–447, 2021.
- [34] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, "ConText: An algorithm for determining negation, experienter, and temporal status from clinical reports", *Journal of Biomedical Informatics*, Vol. 42, No. 5, pp. 839–851, 2009, doi: 10.1016/j.jbi.2009.05.002.
- [35] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries", *Journal of Biomedical Informatics*, Vol. 34, No. 5, pp. 301–310, 2001, doi: 10.1006/jbin.2001.1029.
- [36] Meta AI, "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation", 2025, [Online]. Available: <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- [37] OpenAI, "gpt-oss-120b & gpt-oss-20b model card", 2025, [Online]. Available: <https://openai.com/index/gpt-oss-model-card/>.
- [38] Qwen Team, "Qwen2.5 technical report", *arXiv Preprint*, arXiv:2412.15115, 2024, doi: 10.48550/arXiv.2412.15115.
- [39] Meta AI, "Llama 3.3 70B Instruct model card", 2024, [Online]. Available: <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>.
- [40] M. Abdin, *et al.*, "Phi-4 technical report", *arXiv Preprint*, arXiv:2412.08905, 2024, doi: 10.48550/arXiv.2412.08905.
- [41] Google Research and Google DeepMind, "MedGemma technical report", *arXiv Preprint*, arXiv:2507.05201, 2025, doi: 10.48550/arXiv.2507.05201.
- [42] A. Grattafiori, A. Dubey, *et al.*, "The Llama 3 herd of models", *arXiv Preprint*, arXiv:2407.21783, 2024, doi: 10.48550/arXiv.2407.21783.
- [43] K. Singhal, T. Tu, J. Gottweis, *et al.*, "Toward expert-level medical question answering with large language models", *Nature Medicine*, Vol. 31, No. 3, pp. 943–950, 2025, doi: 10.1038/s41591-024-03423-7.
- [44] E. Goh, R. Gallo, J. Hom, *et al.*, "Large language model influence on diagnostic reasoning: A randomized clinical trial", *JAMA Network Open*, Vol. 7, No. 10, p. e2440969, 2024, doi: 10.1001/jamanetworkopen.2024.40969.
- [45] T. Tu, M. Schaeckermann, A. Palepu, *et al.*, "Towards conversational diagnostic artificial intelligence", *Nature*, Vol. 642, No. 8067, pp. 442–450, 2025, doi: 10.1038/s41586-025-08866-7.