Research Square

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

Condence, Calibration, and Abstention in Small Open-Weight Language Models on Psychiatric Knowledge Questions: A Cross-Domain Replication

Kunal Dhanda

kunaliitkgp09@gmail.com

IIT Kharagpur: Indian Institute of Technology Kharagpur https://orcid.org/0009-0000-3113-6902

Research Article

Keywords: large language models, psychiatry, clinical decision support, calibration, selective prediction, abstention, AUROC, on-premise inference

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11012222/v1

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

Additional Declarations: The authors declare no competing interests.

Confidence, Calibration, and Abstention in Small Open-Weight Language Models on Psychiatric Knowledge Questions: A Cross-Domain Replication

Kunal Dhanda

September 2026 (revised)

Keywords: large language models, psychiatry, clinical decision support, calibration, selective prediction, abstention, AUROC, on-premise inference

Abstract

Small, locally deployable language models are increasingly proposed for psychiatric clinical decision support on privacy and compute grounds; two contemporaneous 2026 systems (ClinMPO; PKFAR) report accuracy or reasoning-quality gains from augmenting such models, and a 120-study taxonomy of LLMs in mental healthcare lists calibration as an evaluated dimension in none of the studies it covers. A separate, contemporaneous study (Boie et al., 2026) evaluates verbalized-confidence calibration across 20 medical specialties using six *proprietary* frontier models on MedMCQA, and a third (Bentegeac et al., 2025) shows across several medical QA benchmarks that verbalized confidence is a weak predictor of correctness (AUROC 0.52–0.68) compared to token probability (AUROC 0.71–0.87). We narrow the open question this leaves for *small, open-weight* models specifically to psychiatry, using the identical four-model ladder (1.2B–9.2B parameters, Ollama, fully local) and harness as a companion general-medicine pilot, on 150 psychiatric-knowledge multiple-choice items (MedMCQA, subject = Psychiatry). Accuracy is higher than on the general-medicine benchmark (57.3–75.3% vs. 47.7–63.2%, though at $n = 150$ the per-model standard error is ±3.5–4.0 points), and the accuracy ranking partially reorders — most notably the 4.3B model falls from 3rd of 4 in general medicine to last in psychiatry. Verbalized confidence remains poorly informative about correctness as a ranking signal (AUROC 0.490–0.645; one model, the 4.3B, is statistically indistinguishable from chance) and severely miscalibrated in the aggregate sense (ECE 0.148–0.378, well above the 0.041–0.141 range Boie et al. report for proprietary frontier models on the same underlying benchmark family). A naive, unpaired "accuracy on answered items" framing suggested that offering an abstention option modestly benefited three of the four models. A same-item paired analysis (McNemar's test, bootstrap confidence intervals on paired accuracy, and an explicit A→B outcome-transition table) overturns most of that: only the smallest model (1.2B) shows a paired effect distinguishable from noise, and it is negative (−6.7 points, 95% CI $[-12.7, -0.7]$) — replicating, point estimate for point estimate, the same model's paired effect on the general-medicine benchmark (−6.7 points, 95% CI $[-10.6, -2.5]$, $n = 480$). The model with the best confidence-ranking signal of the four (7.6B, AUROC 0.645) shows the *worst-targeted* self-chosen abstentions — 8 of its 12 abstentions forfeited an item it would otherwise have answered correctly — and its paired effect reverses from a naive +0.8 to a significant −5.3 points (95% CI $[-9.3, -1.3]$). We

read the surviving evidence as: miscalibration and weak error-detection are general properties of this model family; accuracy ranking is domain-dependent and should not be assumed to transfer; and a prompted abstention option is not a free safety improvement for any of the four models once evaluated with paired, coverage-aware statistics rather than naive answered-item accuracy.

> **This is a pilot.** $n = 150$ on a single benchmark family, no clinician adjudicated any output, and MedMCQA multiple-choice measures psychiatric *knowledge recall*, not psychiatric *clinical* decision-making. Read as hypotheses for a larger, clinician-involved study, not deployment guidance.

1 Introduction

Two studies published in 2026 make the case for small, locally deployable language models as the right target for psychiatric clinical decision support rather than frontier proprietary systems: ClinMPO (Lin et al., 2026) aligns a light-parameter model's internal reasoning with evidence-based psychiatric practice via reinforcement learning and beats a cohort of 300 medical students on held-out reasoning cases; PKFAR (Wang et al., 2026) fuses a psychiatric knowledge graph with a small model's reasoning and outperforms one-shot chain-of-thought baselines from substantially larger models. Both motivate the small-model choice explicitly on privacy and compute grounds. Neither measures whether these models *know when they are wrong*.

That question has recently been studied at scale, but not for this setting. Boie et al. (2026) evaluate self-reported confidence on MedMCQA across 20 medical specialties using six proprietary frontier models (GPT-5, GPT-5-mini, GPT-5-nano, GPT-4o, Claude Sonnet 4.5, Gemini 2.5) and find substantial specialty-level variation in calibration (ECE 0.041–0.141) that does not track accuracy — Claude Sonnet 4.5 was third in accuracy but best calibrated. Bentegeac et al. (2025), across USMLE, MedMCQA, and other medical QA sets, show that verbalized confidence discriminates correct from incorrect answers only weakly (AUROC 0.52–0.68) compared with the model's own token probability (AUROC 0.71–0.87). Neither study evaluates open-weight models small enough to run locally, and neither evaluates abstention.

We narrow the remaining gap to psychiatry specifically — a domain where symptom reports are subjective and there is no laboratory value to fall back on (Gargari et al., 2024), making a confidently wrong suggestion less self-correcting than in most of general medicine — and to the small, open-weight, locally-deployable model class that privacy- and cost-constrained psychiatric settings would actually use. A companion pilot on general clinical knowledge (MedQA, MedMCQA, PubMedQA mixed) found the same four models uniformly and severely miscalibrated, and found that a naive "say I don't know if unsure" instruction *reduced* accuracy for the smallest model under a simple answered-item accuracy comparison. This paper asks whether that holds up under closer statistical scrutiny, and whether it is specific to that benchmark mix or a property of the model.

Research questions

RQ1.
Does the accuracy ranking of these four models on general medicine hold on a psychiatry-specific benchmark, or does it reorder — and how much of any reordering survives sampling noise at $n = 150$?

RQ2.
Is verbalized confidence merely miscalibrated in aggregate (ECE), or is it also a poor *ranking* signal for correctness (AUROC), and does this vary by model?

RQ3.
Does a prompted abstention option produce a genuine, same-item, paired improvement in accuracy — and does any effect found replicate on the general-medicine benchmark?

2 Related Work and Positioning

Small/local models for psychiatric reasoning. ClinMPO (Lin et al., 2026) and PKFAR (Wang et al., 2026) are the closest prior art for the model-size range studied here; both report accuracy/reasoning-quality gains from augmentation, neither reports calibration, AUROC, or abstention behavior.

Calibration across specialties, proprietary models. Boie et al. (2026) is the most directly overlapping contemporaneous study: same underlying benchmark family (MedMCQA), same broad question (does calibration vary across medical domains). It differs on every axis that matters for deployment-relevant conclusions: six *proprietary* frontier models rather than four *open-weight* models small enough to run on a single workstation; 20 general-medical specialties rather than a psychiatry-focused study with a general-medicine companion for direct comparison; no abstention analysis; and, because it evaluates only forced-choice confidence, no paired before/after comparison of any intervention. Their reported ECE range (0.041–0.141) is roughly an order of magnitude tighter than the range we find for small open-weight models on the same underlying question bank (0.148–0.378) — consistent with model scale and training investment buying calibration quality, not just accuracy, though we cannot isolate that from architecture and training-recipe differences with only four models on each side.

Confidence as an error-detection signal. Bentegeac et al. (2025) is the source of the specific metric (AUROC of confidence vs. correctness) we adopt here, and their finding — verbalized confidence is a poor discriminator relative to token probability — motivates our use of AUROC rather than ECE alone as the primary "is this signal useful" metric. They do not evaluate open-weight models below frontier scale or abstention.

Diagnostic accuracy benchmarking. Gargari et al. (2024) and Sarma et al. (2025) benchmark LLMs on DSM-5-TR case vignettes and show that integrating expert-derived diagnostic decision trees suppresses overdiagnosis; Sun et al. (2026) benchmark LLMs against 9,923 real multicenter psychiatric electronic health records and find accuracy varies by patient age. None of these measure verbalized-confidence calibration or AUROC.

Bias. Bouguettaya et al. (2025) find that of four LLMs tested for racial bias in psychiatric treatment recommendations, the one fine-tuned local model tested (a LLaMA-3 variant) showed the *most* racial bias of the four — a caution that "small and local" does not imply "safer" on every axis, complementary to this paper's calibration/abstention focus.

Taxonomy. McLeod et al. (2026) systematically reviews 120 empirical studies of LLMs in mental healthcare; calibration is not among the evaluated dimensions in any of them.

Contribution, stated narrowly. This is a pilot study of verbalized-confidence calibration, error-detection AUROC, and *paired* selective-abstention behavior in small open-weight LLMs on

a psychiatric-knowledge MCQ benchmark, with a direct cross-domain replication check against general medicine using an identical model ladder and harness. We do not claim to be first to study LLM calibration in medicine, in psychiatry broadly, or on MedMCQA — Boie et al. (2026) establishes specialty-level calibration variation on the same question bank with proprietary models — but we are not aware of a prior study combining the small-open-weight-model setting, a psychiatry focus, paired abstention analysis, and a same-harness cross-domain replication.

3 Methods

3.1 Models

Identical to the companion general-medicine pilot: `llama3.2:1b` (1.2B, digest `baf6a787fdff`), `gemma3:4b` (4.3B, digest `a2af6cc3eb7f`), `qwen2.5:latest` (7.6B, digest `845dbda0ea48`), `gemma2:latest` (9.2B, digest `ff02c3702f32`), served locally via Ollama, greedy decoding (temperature = 0, fixed seed). No external API at any point.

3.2 Data

150 items from MedMCQA (`openlifescienceai/medmcqa`, train split, filtered to `subject_name == "Psychiatry"`), a public, subject-labeled Indian postgraduate medical entrance-exam question bank. We use MedMCQA rather than DSM-5-TR case vignettes (used by Gargari et al. and Sarma et al.) because DSM-5-TR case text is copyrighted and not reproducible at scale, and because it lets us reuse the exact MedMCQA pipeline from the companion general-medicine pilot. We refer to the task throughout as **psychiatric-knowledge MCQ reasoning**, not "psychiatric diagnosis" or "psychiatric clinical decision-making" — these items test recall and applied textbook reasoning under exam conditions, not real-case clinical judgment. Random baseline: 25% (4-option MCQ).

3.3 Conditions

- **Condition A** (forced choice): model returns `{"answer": <letter>, "confidence": <0-100>}`.
- **Condition B** (abstention permitted): identical items with an added "Z — not sure / not enough information" option and an instruction to use it when unsure.

3.4 Metrics

- **Accuracy** and its standard error at $n = 150$, $\text{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$.
- **Overconfidence** = mean stated confidence − accuracy; **ECE** (10-bin expected calibration error); **Brier score**.
- **AUROC(confidence → correctness)**, condition A: does the confidence *value* rank correct answers above incorrect ones? Computed via the Mann–Whitney U / rank-sum identity, ties given average rank.

- **Risk–coverage curve**: using condition-A confidence as an abstention-ranking signal (not the model's self-reported "Z" choice), accuracy on the retained top-X% highest-confidence items at coverage 100/90/80/70%.
- **Paired A→B analysis**: for each item answered under both conditions, the outcome pair is tabulated in a 2×3 transition table. McNemar's test (with continuity correction) on the correct↔not-correct discordant pairs (abstention counted as not-correct); a paired bootstrap 95% CI (5,000 resamples) on same-item Δaccuracy, where a condition-B abstention counts as a miss.

3.5 Reproducibility

Harness (`run_pilot.py`) unchanged from the companion pilot other than the benchmark pointer. `analyze.py` reproduces the original selective-accuracy tables; `reanalyze.py` (new in this revision) reproduces every AUROC, Brier, paired-transition, McNemar, bootstrap-CI, and risk–coverage number in §4 directly from `results/raw.jsonl`, with no new model inference. Data collection used a memory-gated retry loop (`grind_gemma2.sh`) on a constrained workstation; 0 of 1,200 rows are partial, corrupted, or unparseable.

4 Results

4.1 T1 — Accuracy, cross-domain ranking, and sampling noise (RQ1)

Table 1: Psychiatry-knowledge MCQ accuracy vs. the companion general-medicine pilot (same 4 models, condition A)

Model	Params (B)	Psychiatry Acc (n=150)	SE	General-medicine macro Acc	Rank (psych.)	Rank (gen. med.)
`llama3.2:1b`	1.2	64.0	±3.9	47.7	2nd	4th (last)
`gemma3:4b`	4.3	**57.3**	±4.0	53.4	**4th (last)**	3rd
`qwen2.5:latest`	7.6	**75.3**	±3.5	63.2	**1st**	**1st**
`gemma2:latest`	9.2	72.7	±3.6	59.3	3rd	2nd

Every model scores higher here than on the general-medicine mix — consistent with a single, well-represented exam subject being easier than a heterogeneous three-benchmark mix, not with psychiatry being an easier *clinical domain*. At $n = 150$ the per-model standard error is 3.5–4.0 points, so not every rank swap should be read as a real effect: `qwen2.5:latest` vs. `gemma2:latest` (75.3 vs. 72.7, a 2.6-point gap against a combined SE of about 5.0) is within noise. The one reordering large enough to plausibly be a real effect is `gemma3:4b`'s fall: it is at least 6.7 points below every other model here, having been mid-pack (53.4, 3rd of 4) in general medicine — a gap comparable to but larger than the combined SE (∼5.6 points), and directionally consistent with the model's confidence being uninformative in this domain (§4.2). We treat RQ1 as **suggestive, not conclusive**: there is no evidence for a single "best small model" that holds across domains, but confirming any *specific* reordering needs a larger sample than this pilot's 150 items.

Table 2: Calibration (ECE, overconfidence, Brier) and error-detection (AUROC), condition A

Model	Acc A	Mean conf	Overconf.	ECE	Brier	AUROC(conf→correct)
`llama3.2:1b`	64.0	91	+27.1	0.271	0.298	0.574
`gemma3:4b`	57.3	95	+37.8	0.378	0.388	**0.490**
`qwen2.5:latest`	75.3	90	+14.8	0.148	0.198	**0.645**
`gemma2:latest`	72.7	95	+22.3	0.223	0.248	0.515

4.2 RQ2 — Confidence as calibration vs. confidence as an error-detection signal

Every model is severely overconfident in the aggregate sense (mean stated confidence 90–95% against 57–75% actual accuracy) and severely miscalibrated (ECE 0.148–0.378) relative to the 0.041–0.141 range Boie et al. (2026) report for six proprietary frontier models on the same MedMCQA question bank — roughly an order of magnitude tighter, though a four-model-vs-six-model, open-vs-proprietary comparison cannot isolate scale from architecture or training recipe as the cause.

But ECE and AUROC answer different questions, and they disagree here. `gemma3:4b` has an AUROC of 0.490 — statistically indistinguishable from 0.5, meaning its stated confidence carries **no information** about whether it is right or wrong, even though it is (like the others) globally overconfident. `qwen2.5:latest` has the best AUROC of the four (0.645) — its confidence is a genuinely useful, if imperfect, ranking signal, well within the 0.52–0.68 range Bentegeac et al. (2025) report for verbalized confidence generally (still far below their reported 0.71–0.87 for token probability, which we did not measure here — see Limitations).

Table 3: Risk–coverage curve: accuracy when abstaining on the lowest-confidence X% of items (condition-A confidence as the ranking signal, not the model's self-reported abstention choice)

Model	Acc @100% cov	@90%	@80%	@70%
`llama3.2:1b`	64.0	64.4	63.3	63.8
`gemma3:4b`	57.3	57.8	58.3	56.2
`qwen2.5:latest`	75.3	75.6	76.7	**79.0**
`gemma2:latest`	72.7	73.3	73.3	76.2

This confirms the AUROC ranking directly: discarding `qwen2.5:latest`'s bottom 30%-confidence items lifts its accuracy by 3.7 points (75.3→79.0), a real and usable selective-prediction gain; `gemma3:4b`'s curve is flat and noisy (57.3→56.2 at 70% coverage — confidence-based thresholding does not help it, consistent with AUROC ≈ 0.49).

4.3 RQ3 — Paired abstention analysis overturns the naive result

The companion pilot and an earlier draft of this one reported "Δaccuracy from abstaining" as (accuracy on whichever items condition B chose to answer) minus (accuracy on all items in condition A) — a **selective-accuracy** framing. Under that framing:

This framing is exactly the ambiguity flagged in our own Limitations in the previous draft: a rise in "accuracy on answered items" is consistent with the model abstaining disproportionately on items it would get wrong, but it is equally consistent with abstaining on a mix that happens to raise

Table 4: Selective-accuracy framing (as originally reported; kept for comparability, not relied on below)

Model	Acc A	Abstain rate B	Acc B (answered)	Δ (selective-accuracy framing)
`llama3.2:1b`	64.0	4.7%	60.1	−3.9
`gemma3:4b`	57.3	4.7%	61.5	+4.2
`qwen2.5:latest`	75.3	8.0%	76.1	+0.8
`gemma2:latest`	72.7	2.0%	74.1	+1.5

the average for other reasons, and it does not use the same denominator across conditions. **A same-item paired analysis tells a different, more specific story.**

Table 5: Paired A→B outcome transitions and McNemar's test

Model	correct→correct	correct→incorrect	correct→abst.	incorrect→correct	incorrect→incorrect	incorrect→abst.	McNemar χ^2	p
`llama3.2:1b`	80	13	3	6	44	4	3.68	0.055
`gemma3:4b`	83	1	2	5	54	5	0.12	0.724
`qwen2.5:latest`	104	1	**8**	1	32	4	4.90	**0.027**
`gemma2:latest`	108	0	1	1	38	2	0.50	0.480

Table 6: Paired bootstrap 95% CI for same-item Δaccuracy (abstention counted as a miss)

Model	Paired Δacc	95% CI
`llama3.2:1b`	**−6.7** pp	[**−12.7**, **−0.7**]
`gemma3:4b`	+1.3 pp	[−2.7, +5.3]
`qwen2.5:latest`	**−5.3** pp	[**−9.3**, **−1.3**]
`gemma2:latest`	+0.0 pp	[−2.0, +2.0]

Three things change once the analysis is paired:

1. **`gemma3:4b`'s and `gemma2:latest`'s naive gains (+4.2, +1.5) do not survive.** Both paired CIs straddle zero (McNemar $p = 0.72$ and $p = 0.48$); `gemma2:latest` abstains only 3 times in 150 items, too little to support any claim about its abstention behavior either way.
2. **`qwen2.5:latest`'s naive small gain (+0.8) reverses to a significant loss (−5.3, $p = 0.027$).** Of its 12 self-chosen abstentions, 8 (67%) were on items it would have answered *correctly* under forced choice — despite having the best confidence-ranking signal of the four models (§4.2), its behavioral decision of *when* to invoke "not sure" does not exploit that signal well.
3. **Only `llama3.2:1b` shows an effect that is negative under both framings**, and the paired estimate (−6.7 points) is close to marginal significance alone ($p = 0.055$, $n = 150$) but matches, point estimate for point estimate, the same model's paired effect on the companion general-medicine pilot: **−6.7** points, 95% CI [**−10.6**, **−2.5**], McNemar $p = 0.002$, $n = 480$ (pooled across that pilot's three benchmarks). That larger, more clearly significant general-medicine result and this psychiatry result agree exactly on the point estimate and agree in sign and rough magnitude on the interval — the strongest evidence in either paper for a real, replicated, model-specific effect.

Revised RQ3 answer: a prompted abstention option is not a free safety improvement for any of the four models once evaluated with paired, same-item statistics. The one model for which it

clearly does something is `llama3.2:1b`, and what it does is make things worse, replicated across two domains.

4.4 Accuracy by psychiatric topic (descriptive only)

The MedMCQA topic tag is present for only 75 of 150 items (50%); of the 34 distinct labeled topics, all but four have $n \leq 5$. We do not report this as a topic-capability estimate.

4.5 Latency

Model	Median s/item (condition A)
`llama3.2:1b`	0.31
`gemma3:4b`	1.33
`qwen2.5:latest`	1.25
`gemma2:latest`	1.99

All four models respond in under 2 seconds at the median on a single Apple-silicon workstation with no external API — faster than the companion pilot's numbers (0.42–3.54s), consistent with these being shorter, single-domain, no-context items.

5 Discussion

A three-layer picture, not one finding. The evidence supports separating "is the model's confidence any good" (§4.2, a property that varies enormously by model — `qwen2.5:latest`'s AUROC of 0.645 vs. `gemma3:4b`'s 0.490, indistinguishable from chance) from "does the model's *behavior* when offered abstention actually track that confidence" (§4.3, which it does not do reliably even for the model with the best confidence signal) from "does accuracy ranking transfer across clinical subdomains" (§4.1, largely no, though only one reordering — `gemma3:4b`'s fall — clears its own sampling noise). These are three different claims with three different levels of support, and collapsing them into "small LLMs are overconfident in psychiatry" would understate what is actually here and overstate what a reviewer could take at face value from the aggregate ECE numbers alone.

Relative to Boie et al. (2026) and Bentegeac et al. (2025). Boie et al. establish that calibration varies substantially by medical specialty for proprietary frontier models on this same question-bank family; we do not revisit that claim, and our ECE range (0.148–0.378) sits an order of magnitude above their proprietary-model range (0.041–0.141), consistent with — though not proof of — a scale or training-investment effect on calibration quality specifically, not just accuracy. Bentegeac et al. establish that verbalized confidence is a weaker error-detection signal than token probability across several medical QA benchmarks; our AUROC numbers (0.490–0.645) sit inside or below their reported verbalized-confidence range (0.52–0.68), and we did not measure token-probability AUROC for comparison here (see Limitations) — doing so is the most direct next experiment.

Relative to ClinMPO and PKFAR. Both report accuracy or reasoning-quality gains from augmenting small psychiatric-reasoning models. Neither reports calibration or AUROC for their

augmented models. Given that even the best-behaved model in this ladder (`qwen2.5:latest`) has an AUROC of only 0.645 and an ECE of 0.148, an augmentation method that improves accuracy without checking whether it also improves — or degrades — the confidence signal's usefulness is only reporting half of what a deployer needs to know.

Practical implication. "Add a not-sure option" is often treated as a free safety improvement. Under the paired analysis here, it is not a *reliable* improvement for any of the four models, and it is actively harmful, twice replicated, for the smallest. Any abstention-based safety layer for a small local model needs paired, same-item validation — not a before/after comparison on different denominators — before being trusted.

6 Limitations

1. **Single benchmark family, single subject,** $n = 150$**.** Per-model SE is 3.5–4.0 points on accuracy; several results described as "suggestive" rather than "conclusive" for this reason (§4.1).
2. **Knowledge-recall MCQ, not clinical decision-making.** We rename the task "psychiatric-knowledge MCQ reasoning" throughout for this reason; MedMCQA is not equivalent to DSM-5-TR case reasoning, clinical notes, or real patient encounters.
3. **No clinician adjudication.** Gold labels are exam answer keys.
4. **Token-probability AUROC not measured.** Bentegeac et al. (2025) show token probability outperforms verbalized confidence for error detection; we measured only verbalized confidence. Comparing the two for this exact model ladder is the most direct follow-up and did not require new experiment design, only new inference — deferred here specifically to avoid further heavy local-model runs on the workstation used for data collection (§3.5, `grind_gemma2.sh`).
5. **Two-domain replication, not a general law.** The `llama3.2:1b` finding replicates across two domains with matching point estimates; it is not evidence about any other domain, model, or model family.
6. **No overgeneralization by parameter count.** An earlier draft described the abstention-backfire result as characteristic of "sub-2B models"; with only one model below 2B in this ladder, the defensible claim is about `llama3.2:1b` specifically, not a parameter-count threshold.
7. **Topic breakdown underpowered** (§4.4).
8. **Prompt and decoding choices.** Single prompt, single seed, greedy decoding; both this pilot's numbers and its cross-domain comparison reflect one configuration.
9. **Model/version drift.** Ollama `:latest` tags are not permanent; digests are recorded in §3.1 and `results/`.

7 Conclusion

This is a cross-domain reliability study, not another "LLMs are overconfident" paper — that specific claim is now well established in the broader medical-LLM literature (Boie et al., 2026; Bentegeac et al., 2025) and is not this paper's contribution. What we add: (1) accuracy ranking among

four small open-weight models is domain-dependent and should not be assumed to transfer from a general benchmark to psychiatry, though only one reordering here clears its own sampling noise; (2) verbalized confidence quality is itself highly model-dependent — one model's confidence is a genuinely useful ranking signal (AUROC 0.645), another's is statistically indistinguishable from useless (AUROC 0.490) — so "the model is calibrated" and "the model's confidence can detect its own errors" are different claims that can diverge; and (3) a prompted abstention option, evaluated properly with paired same-item statistics rather than a naive answered-item comparison, is not a reliable safety improvement for any of the four models and is a replicated, statistically consistent harm for the smallest one, with matching point estimates (-6.7 points) across two independent clinical-domain benchmarks. The practical takeaway is narrow: do not trust verbalized confidence from these models without checking its AUROC specifically, and do not add a "let it abstain" instruction to a small local model without a paired before/after check.

Reproducibility artifacts

- `run_pilot.py` — data-collection harness (Ollama HTTP, deterministic, resumable)
- `grind_gemma2.sh` — memory-gated retry loop used for data collection on a constrained workstation
- `analyze.py` — original selective-accuracy tables
- `reanalyze.py` — **new**: AUROC, Brier, paired transitions, McNemar's test, bootstrap CI, risk–coverage curve (all §4.2–4.3 numbers), zero new model inference
- `data/psychmcqa.jsonl` — the exact 150 evaluated items
- `results/raw.jsonl` — all 1,200 raw generations, 0 errors
- `results/tables.md`, `results/reanalysis.md`, `results/summary.json` — generated tables backing §4

References

1. Lin, X. et al. *Evaluating an evidence-guided reinforcement learning framework in aligning light-parameter large language models with decision-making cognition in psychiatric clinical reasoning* (ClinMPO), ArXiv, 2026.
2. Wang, R. et al. *PKFAR: psychiatry knowledge-fused augmented reasoning with large language models*, Health Information Science and Systems, 2026.
3. Boie, S. D., Reis, F., Frey, N., Grünewald, E., & Balzer, F. *Calibration of Self-Reported Confidence and Accuracy of Large Language Models in Medical Question Answering*, Journal of Medical Systems, 2026. DOI: 10.1007/s10916-026-02430-0.
4. Bentegeac, R., Le Guellec, B., Kuchcinski, G., Amouyel, P., & Hamroun, A. *Token Probabilities to Mitigate Large Language Models Overconfidence in Answering Medical Questions: Quantitative Study*, Journal of Medical Internet Research, 2025. PMID: 40882190.

5. McLeod, J. et al. *Large Language Models in Psychiatry: A Systematic Review of Clinical Use Cases and Evidence*, BJPsych Open, 2026.

6. Gargari, O. K. et al. *Diagnostic accuracy of large language models in psychiatry*, Asian Journal of Psychiatry, 2024.

7. Sarma, K. V. et al. *Integrating Expert Knowledge into Large Language Models Improves Performance for Psychiatric Reasoning and Diagnosis*, Psychiatry Research, 2025.

8. Sun, M. et al. *Large Language Models for Psychiatric Diagnosis Based on Multicenter Real-World Clinical Records: Comparative Study*, JMIR Medical Informatics, 2025/2026.

9. Bouguettaya, A. et al. *Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models*, NPJ Digital Medicine, 2025.

10. Pal, A., Umapathi, L. K., & Sankarasubbu, M. *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*, 2022.

11. Kadavath, S. et al. *Language Models (Mostly) Know What They Know*, 2022.

e pilot, this author *Small Open-Weight Language Models in the Medical Field: A Four-Task Pilot on Clinical QA, Calibration, Patient Communication, and Note Summarization*, 2026.