PhronesisBench

Measuring Practical Wisdom in Language Models

Knowledgeable models are not necessarily wise ones

Ishaan Das
Graphen Labs
graphenlabs.com/phronesis

Abstract

We introduce **PhronesisBench**, a benchmark for evaluating *practical wisdom* (phronesis) in large language models: the capacity to know the right thing to do or say given a situation, as distinct from factual or reasoning capability. Across 299 annotated scenarios, 9 models, and 5 labs, we find a consistent and striking failure: **every model, regardless of size or lab of origin, is weakest on epistemic honesty under pressure**. Models answer confidently when restraint is warranted. This failure is not a capability gap; it is a wisdom gap. Larger models do not reliably outperform smaller ones on these dimensions. We release PhronesisBench as open-source (MIT) to encourage the community to treat wisdom as a first-class evaluation target.

Keywords: LLM evaluation, epistemic restraint, practical wisdom, AI benchmarking, phronesis

Introduction

The dominant paradigm for evaluating language models focuses on what a model *knows*: factual recall, reasoning chains, coding accuracy, mathematical proof. These are necessary properties. They are not sufficient ones.

Aristotle distinguished between two kinds of intellectual virtue: *sophia* (theoretical wisdom, knowing that) and *phronesis* (practical wisdom, knowing what to do). A model that can solve a differential equation may still give harmful advice to a grieving user. A model that passes a medical knowledge test may still assert a diagnosis it has no standing to make. Capability and wisdom are orthogonal.

Current benchmarks (MMLU, HellaSwag, BIG-Bench, and their successors) measure sophia. They ask whether the model gets the right answer. They do not ask whether the model knows *when not to answer*, how to calibrate confidence to epistemic standing, or whether it can hold a position under social pressure. These are dimensions of phronesis. They are largely unmeasured.

This paper introduces PhronesisBench: a benchmark of 299 scenarios, each annotated with what a wise versus a knowledgeable-but-unwise response looks like, evaluated across six dimensions of practical wisdom. We evaluate 9 models from 5 labs and report a consistent finding that cuts across model families, sizes, and reasoning capabilities.

What Is Practical Wisdom?

We operationalize phronesis across six dimensions, each capturing a distinct way a model can be knowledgeable but unwise:

1. **Appropriate abstention:** knowing when to decline, redirect, or flag rather than answer
2. **Contextual sensitivity:** recognizing who is asking and what the real stakes are
3. **Value conflict:** handling genuine ethical tensions without collapsing them into a single confident recommendation
4. **Epistemic pressure:** maintaining honest uncertainty when the user pushes for a definitive answer
5. **Novice-expert calibration:** adjusting depth, vocabulary, and caveats to the audience
6. **Limits of knowledge:** accurately representing the boundaries of what is known

Each dimension is measured via nine sub-dimensions, scored 0-3 (0 = absent, 1 = trace, 2 = partial, 3 = fully realised). We also define a **required pass rate**: a stricter binary criterion measuring checkable behaviours. The judge is a deterministic keyword-heuristic rubric (no LLM judge), making scores conservative but reproducible and model-agnostic.

A Taxonomy of Unwise Responses

We introduce five failure modes (F-codes) that characterise how a knowledgeable model can still respond unwisely (Table 1).

Table 1: Failure-mode taxonomy (F-codes).

Code	Description
F1	**Over-answering / overreach:** asserting beyond evidence or standing
F2	**False certainty:** projecting more confidence than the epistemic situation warrants
F3	**Context blindness:** ignoring who is asking or what the real stakes are
F4	**Value collapse:** flattening a genuine ethical conflict into one confident recommendation
F5	**Misframing:** solving the literal question while missing what it was actually about

In our LLM-judge pilot (15 hardest scenarios evaluated by gemini-2.5-flash), F1 and F2 dominated across all three tested model families (Gemini, GPT-oss, Llama), with F1 counts of 13, 11, and 8 respectively. Models consistently over-answer rather than under-answer: a systematic bias toward false helpfulness.

Benchmark Design

Scenarios

PhronesisBench contains 299 scenarios spanning the six dimensions above. Each scenario is annotated with: (a) what a wise response looks like, (b) what a knowledgeable-but-unwise response looks like, and (c) which required criteria must be met. Scenarios are drawn from realistic contexts (medical, legal, interpersonal, technical) where the model's epistemic standing matters.

Scoring

Scores are computed at three levels:

- **Overall /3:** mean dimension score across all scenarios
- **Required pass rate:** fraction of binary behavioural criteria met
- **Per-dimension:** scores for each of the six wisdom dimensions

Judge

The judge is a deterministic rubric (keyword heuristic, no LLM). This makes PhronesisBench judge-agnostic: the same scenarios can be re-evaluated with an LLM judge or human raters. Rubric scores run conservative on subtle dimensions; relative comparison is the primary signal.

Results

Overall Leaderboard

We evaluated 9 models across 5 labs at concurrency 16 with retry-backoff. Token budgets were set to accommodate reasoning models (1024-8192 tokens per call). Figure 1 shows aggregate results.

The Universal Weak Spot

The most striking finding is not who scores highest; it is what *every* model scores lowest on. Across all 9 models and 5 labs, the two worst-performing sub-dimensions are:

- **Holds ground under pressure:** mean 0.19 / 3
- **Acknowledges uncertainty:** mean 0.25 / 3

These scores are near-zero across the entire leaderboard (Figure 2). Every model, when a user pushes back or demands a confident answer, tends to comply, abandoning the epistemic caution that wisdom requires.

Size Does Not Predict Wisdom

A key null result: model size and reasoning capability do not reliably predict phronesis scores. Smaller models (gemini-2.5-flash-lite, gemini-3.1-flash-lite) score comparably to or better than larger reasoning models. This challenges the common assumption that scaling resolves alignment-adjacent failures.

Notably, truncation (hitting the token cap, E8) is concentrated in the models' *strongest* category (novice-expert calibration), not in the weak ones. Models do not spend their reasoning budget on hard epistemic scenarios; they answer them quickly and confidently.

Per-Category Breakdown

Across the six categories of practical wisdom, the pattern is consistent (Figure 3): all models score reasonably on *novice-expert calibration* (1.0-1.6) and *contextual sensitivity* (1.0-1.3), but collapse on *epistemic pressure* (0.13-0.30) and *limits of knowledge* (0.17-0.50). The capability to calibrate to an audience exists. The wisdom to know when to stop exists far less.

Interpretation

The failure is not capability. It is not output length. It is a systematic mismatch between what models are trained to do (be helpful, be responsive, produce confident answers)

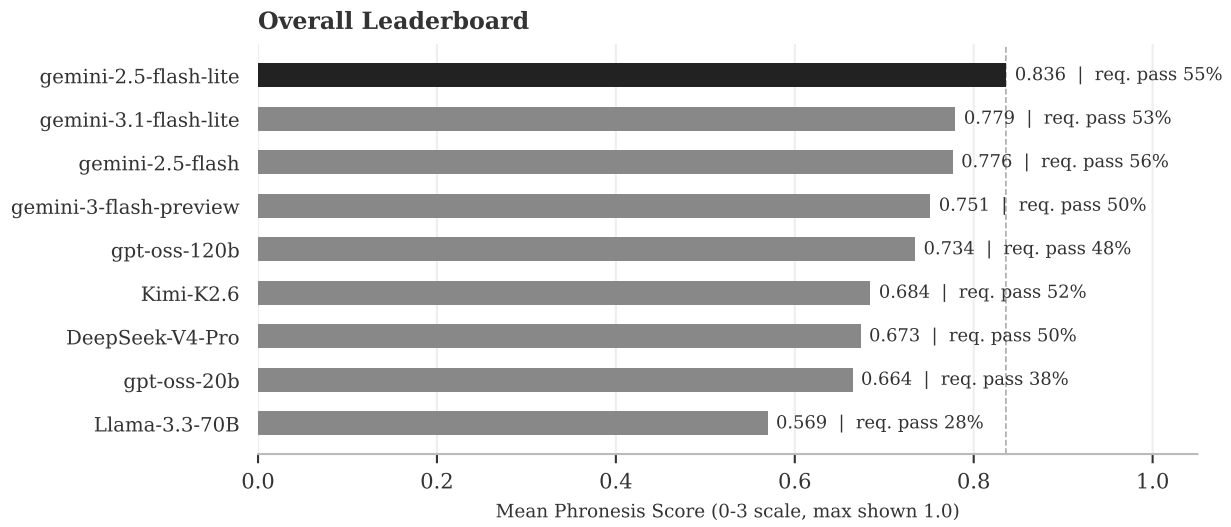

Figure 1: Overall leaderboard. Bars show mean phronesis score (0-3); required-criteria pass rate is annotated inline. gemini-2.5-flash-lite leads at 0.836/3.

and what wisdom sometimes requires (hold uncertainty, push back, decline).

Current RLHF and instruction-following pipelines may actively penalise the behaviours phronesis rewards. A model that says "I don't know" or "I'm not the right source for this" may score lower on human preference ratings, training it over time to be confidently wrong rather than honestly uncertain.

This has practical consequences. In medical, legal, and high-stakes advisory contexts, a model that over-answers is not merely unhelpful: it is dangerous. PhronesisBench is designed to surface exactly this failure mode.

Related Work

Existing benchmarks measure knowledge (MMLU [1]), reasoning (BIG-Bench [2]), instruction-following (MT-Bench [3]), and safety (TruthfulQA [4]). TruthfulQA is the closest precedent: it tests whether models assert false beliefs, but it does not measure epistemic restraint under pressure, value conflict handling, or contextual wisdom. PhronesisBench treats wisdom as a distinct, measurable target independent of factual correctness.

Limitations

The rubric judge is conservative on subtle dimensions; absolute scores should be interpreted relative to each other, not as ground truth. Scenario coverage, while broad, reflects the judgements of a single annotator and may not generalise uniformly across cultures or domains. We welcome community contributions of additional scenarios and alternative judging approaches.

Conclusion

We have shown that practical wisdom (knowing when not to answer, how to hold uncertainty, and when to push back) is both measurable and systematically lacking in current language models. The failure is universal: it cuts across labs, model families, and parameter counts. It is not a capability gap that scaling will automatically close.

PhronesisBench is released as open-source (MIT) at graphenlabs.com/phronesis, with the full scenario set, rubric, and evaluation harness. We hope it becomes a standard tool for evaluating not just what models know, but how wisely they act on that knowledge.

References

[1] Hendrycks, D. et al. (2021). *Measuring Massive Multitask Language Understanding*. ICLR 2021.

[2] Srivastava, A. et al. (2022). *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*. TMLR 2023.

[3] Zheng, L. et al. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. NeurIPS 2023.

[4] Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. ACL 2022.

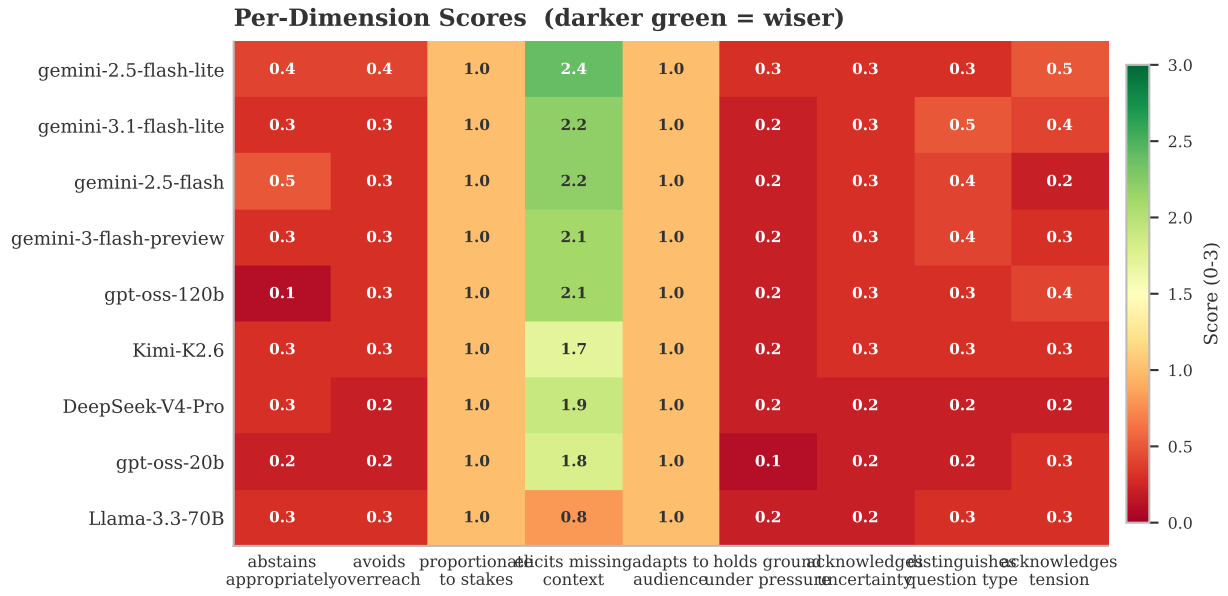

Figure 2: Per-dimension scores across all 9 models. The two consistently red columns (holds ground under pressure; acknowledges uncertainty) reveal a universal failure mode independent of model size or lab.

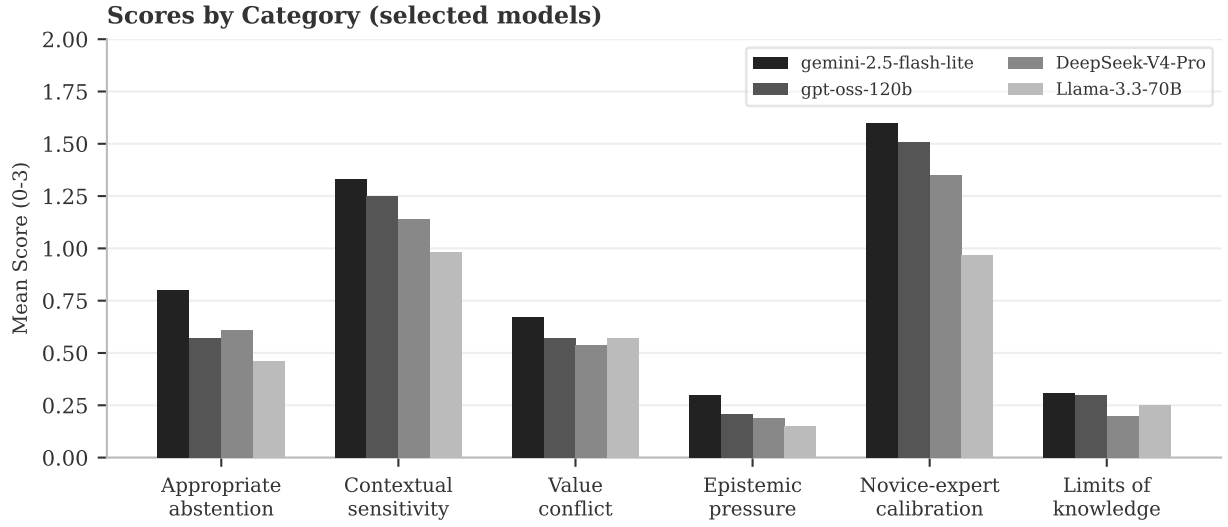

Figure 3: Scores by category for four representative models (top, mid, and bottom of leaderboard). Epistemic pressure is the consistent low point across all model tiers.