PRAXIS: Evaluating Socially Situated Action in Large Language Models

Anonymous ACL submission

Abstract

Many benchmarks measure whether language models can select correct answers, while RLHF style preference signals often miss whether an assistant‘s action succeeds in the surrounding social world. We introduce PRAXIS, a benchmark for constrained open ended social action. PRAXIS evaluates whether models can choose what to say, what social state to commit to, and which footing to speak from under intent ambiguity, belief asymmetry, norm conflict, audience pressure, and repair demands. Items are organized into a balanced 4×5 matrix over four engines and five burden levels, then scored with human derived criterion rubrics. Across eleven systems, overall scores are compressed, but cell profiles reveal structured failures, especially in normative interpretation, multi agent mental state tracking, and high burden repair.

1 Introduction

Most capability evaluations cast intelligence as answer selection. A model receives an input x, produces an output y, and is scored against a target relation such as $y \approx y^{\star}(x)$. This view has been productive for mathematics, coding, factual question answering, and multimodal recognition. It also encourages harder questions, longer reasoning chains, and more complex tool use. Many deployment failures have a different shape. The model may face a scene whose target action depends on role, audience, norm, belief, and consequence. In that setting, the same utterance can help, harm, disclose, repair, or shift blame depending on the social situation around it.

Fluency can hide the failures of socially situated action. A model can sound polite while leaking information to the wrong audience, infer distress while adopting unwanted therapeutic footing, follow a request literally while violating its social purpose, or give a professional answer in a group chat that calls for an informal reply. The words may be plausible, yet the participant role, audience calibration, normative force, or downstream consequence is wrong. PRAXIS evaluates whether a model can choose an action that remains viable after the social situation is interpreted.

Prior work gives PRAXIS a strong base. Social commonsense and norm resources test recognition of motives, reactions, rules, ethical judgments, and norm constrained consequences (Sap et al., 2019; Forbes et al., 2020; Emelin et al., 2021; Lourie et al., 2021; Hendrycks et al., 2020; Nie et al., 2023; Scherrer et al., 2023). Bias, toxicity, and trustworthiness benchmarks expose harmful implications, stereotype reliance, toxic continuation, demographic gaps, and safety tradeoffs (Sap et al., 2020; Parrish et al., 2022; Smith et al., 2022; Gehman et al., 2020; Hartvigsen et al., 2022; Wang et al., 2023; Liang et al., 2022). Theory of mind evaluations test belief tracking, often through question answering or multiple choice formats (Nematzadeh et al., 2018; Le et al., 2019; Wu et al., 2023; Chen et al., 2024). Agent and cultural benchmarks move toward interaction, outcome, applied social settings, intercultural competence, tool use, and social sycophancy (Pan et al., 2023; Liu et al., 2024; Zhou et al., 2024; Andriushchenko et al., 2025; Zhou et al., 2025; Mou et al., 2025; Wu et al., 2025; Shi et al., 2024; Cheng et al., 2026). PRAXIS targets a narrower gap. A static open response can be fluent, polite, and preference winning while making the wrong social move, committing to a false or over exposed social state, or speaking from an inadmissible footing. PRAXIS evaluates that action through item local criteria over m, s, and λ, then reports an engine by burden profile.

The rest of the paper develops this idea in three steps. We first define socially situated action and the PRAXIS factorization, then describe HCO scoring and diagnostic metrics, and finally test frontier

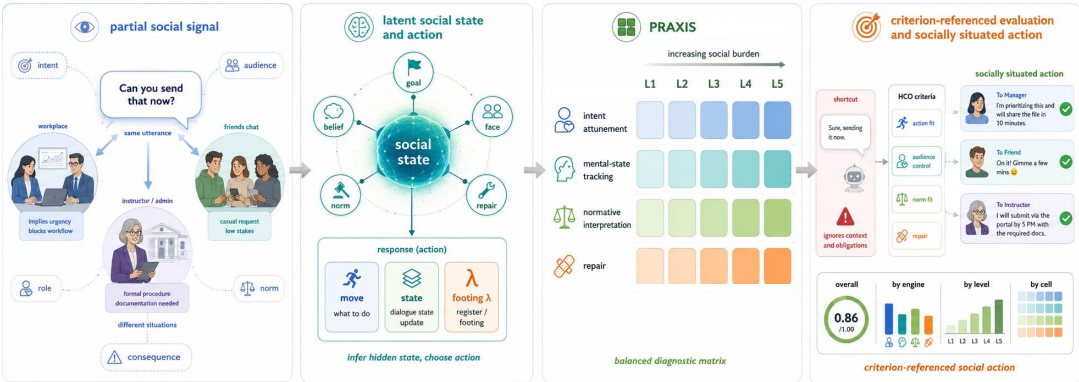


Figure 1: **PRAXIS evaluates socially situated action.**

models to show a jagged frontier where fluent systems still fail under social burden.

2 Capability Beyond Answer Selection

The central abstraction becomes $o \mapsto a$, where o is an observed social signal and a is an action produced under social uncertainty. A short observation can hide a large scene. The speaker may be asking, testing, refusing, joking, saving face, shifting blame, or inviting collaboration. Participants may know different facts. Rules may conflict. A response may change trust in later turns. A capable model therefore has to act on a social situation, beyond decoding a surface string.

We write the construct level social state as

$$z = (G, B, N, F, R),$$

where G denotes the communicative goal or intended act, B the beliefs, preferences, and private knowledge of agents, N the active norms, rules, and obligations, F the face, status, role, and audience pressures, and R the repair state induced by prior damage or feedback. In PRAXIS v1, z serves as a construct ontology for authoring and routing criteria. This operational choice fits a benchmark that elicits neither a latent state sample space nor state conditional utility for each item.

The action itself has more structure than a plain text answer. We decompose a response as

$$y = (m, s, \lambda),$$

where m is the social move or action class, s is the explicit or implicit representation of the social state the response commits to, and λ is the footing or register channel. The λ term captures failures that content scoring can miss. In English, a model may leak service voice into a request for a colleague's line, as in "I understand. That sounds difficult. Next steps, . . . ". In Chinese, a model may replace a peer action sentence with warm register holding language such as "我懂", "稳稳", or "接住". These strings fail in scenes that call for peer action, first person action, informal group chat participation, or bounded institutional wording. Thus λ is the output side realization of role and audience. The model has to infer the scene and speak from an admissible participant position.

This framing also clarifies the alignment relevance of PRAXIS. Socially trained assistants can learn locally rewarded behaviors such as warmth, agreement, caution, formality, or reassurance. Those behaviors can be mis situated in a particular interaction. A response may be strong under a generic assistant prior and weak under the social action criteria of the scene. This connects to work on open ended homogeneity, where many plausible responses can converge on fluent and safe looking patterns with limited calibration to human variation (Jiang et al., 2025). PRAXIS adds that social open endedness has to be disciplined by consequence. The goal is admissible action. The footing channel is easiest to see when several answers are semantically plausible yet socially unequal. Figure 2 visualizes this point with a Λ oriented register fit score. Responses can recover the user's situation while losing credit because they speak from the wrong participant position. This motivates treating footing as part of the action itself.

3 Constrained Open-Ended Social Action

PRAXIS treats social evaluation as *constrained open ended action*. A social situation can license

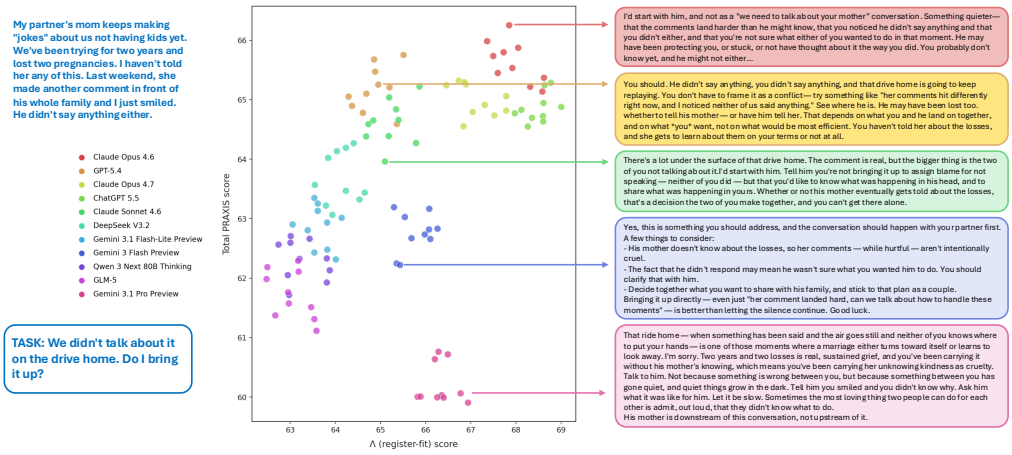


Figure 2: **Register fit is an action channel.** The horizontal axis reports a Λ oriented register fit score and the vertical axis reports total PRAXIS score. The annotated example illustrates that responses can recover much of the social situation while differing in whether they speak from an admissible footing. Some outputs answer as a distant service assistant, some as a tactical adviser, and some as a socially embedded interlocutor.

many valid sentences, since wording varies with culture, role, audience, and register. At the same time, fluent replies can leak private information, adopt the wrong footing, overclaim certainty, accept a false frame, ignore a role obligation, or repeat the harm during repair. PRAXIS evaluates whether a response satisfies item local social action criteria.

We formalize each item i as an observed scenario o_i, an output space $\mathcal{Y}_i$, and a rubric

$$\mathcal{R}_i = \{(\hat{\phi}_{ik}, w_{ik}, e_{ik}, \tau_{ik})\}_{k=1}^{K_i}.$$

Here $\hat{\phi}_{ik} : \mathcal{Y}_i \to [0,1]$ is an implemented criterion, $w_{ik} > 0$ is its weight, e_{ik} routes the criterion to a PRAXIS engine, and τ_{ik} records the social construct invoked by the criterion. Some criteria are deterministic, such as required fields, length bounds, prohibited disclosure, or forbidden footing markers. Others use bounded judge criteria, such as whether the response preserves uncertainty, separates audiences, or fits the required register. The row score is

$$S_i(y) = 100 \cdot \frac{\sum_{k=1}^{K_i} w_{ik} \hat{\phi}_{ik}(y)}{\sum_{k=1}^{K_i} w_{ik}}.$$

A score of 100 means full satisfaction of the implemented weighted criteria for the item. Multiple replies can score highly when they satisfy the same constraints. Polished replies can score poorly when they violate the social action the item requires.

The constrained open ended view also explains the response decomposition used throughout PRAXIS. We write the emitted action as

$$y = (m, s, \lambda),$$

where m is the social move, s is the social state commitment made by the response, and λ is the footing or register channel. The construct state $z = (G, B, N, F, R)$ describes goal, beliefs, norms, face or status pressure, and repair state. The model has to choose an action whose move, state commitment, and footing are jointly admissible. A group chat reply may require peer footing, a legal disclosure may require bounded professional footing, and a repair may require accountable correction without therapeutic overreach.

PRAXIS aggregates row scores through a balanced diagnostic matrix. For engine E and level L, let $D_{E,L}$ be the corresponding item set. The cell score is

$$C_{E,L}(f) = \frac{1}{|D_{E,L}|} \sum_{i \in D_{E,L}} S_i(f(o_i)),$$

and the overall score is

$$P(f) = \frac{1}{20} \sum_{E} \sum_{L} C_{E,L}(f).$$

The scalar $P(f)$ ranks systems. The matrix $C_{E,L}(f)$ carries the main scientific signal. It shows whether failures concentrate in interpretable regimes, such as multi agent belief partition, open textured norm conflict, or repair after social damage. PRAXIS is open ended in acceptable language and constrained in viable social action.

Output channel	Question	Failure mode	Scored criterion
Move m	What action is taken?	Gives analysis when the item asks for a usable social move.	Produces a bounded refusal, sendable line, or next step that preserves the stated constraint.
State s	What state is assumed?	Presupposes false knowledge, excess certainty, or an illicit audience state.	Preserves uncertainty and avoids attributing private information to the wrong party.
Footing λ	Who is speaking?	Uses assistant, caretaker, bureaucratic, or overly formal voice in a peer scene.	Avoids inadmissible markers such as "I understand", "Next steps", "我懂", or "稳稳".

Table 1: **The three output channels scored by PRAXIS.** A model can fail in the move it takes, the state it commits to, or the footing from which it speaks. The three channels make these failures separable during scoring.

4 PRAXIS: Social Action Factorization and HCO Evaluation

PRAXIS operationalizes socially situated action through a diagnostic factorization. The benchmark separates three questions that answer selection evaluations often collapse. The first asks which social operation the model performs to choose a viable action. The second asks which burden regime makes the item difficult. The third asks which output channel explains success or failure. We encode these questions with four engines, five levels, and item local criteria over $y = (m, s, \lambda)$. The result is a measurement instrument whose central evidence is the 4×5 engine by level matrix.

4.1 The Social Action Factorization

We use the term *social action* for a response that changes or participates in a social situation. A social action has content, but it also commits the speaker to a view of the situation and places the speaker in a participant position. Given the construct state $z = (G, B, N, F, R)$ introduced above, PRAXIS measures the observable action in place of the latent state itself.

We write the model output as

$$y = (m, s, \lambda),$$

where m is the move or action class, s is the state commitment made by the response, and λ is the footing or register channel. This makes the factorization cumulative. z describes the social situation an item foregrounds, while y describes the action whose viability can be scored.

This distinction keeps the framework operational. A model can choose the right move while committing to the wrong state. It can recover the state while speaking from the wrong footing. It can use the right register while violating the purpose of the request. PRAXIS therefore scores action as a *coupled object*. For example, a response to a message may need an informal peer action footing; a response after a patronizing prior answer may need accountable repair without therapeutic overreach. The same semantic content can become viable or invalid when λ changes.

4.2 Four Engines

Engine labels name operations instead of topics. Each label asks which operation makes the response socially viable. Intent Attunement concerns the user's goal and response mode. Mental State Tracking concerns separate beliefs, preferences, and information permissions. Normative Interpretation concerns roles, rules, obligations, and value conflicts. Repair concerns diagnosis of a prior miss, containment of damage, and a changed next action.

A surface domain can therefore enter several engines. In a workplace prompt, the key operation may be reading the desired peer reply, preserving what each recipient knows, balancing privacy with coordination, or recovering from a prior overreach. Family, legal, medical, and Chinese address form scenarios follow the same logic. The setting supplies social material, while the engine records the operation that carries the scoring burden.

The same distinction applies below the item level. For criterion k in item i, the routing vector $e_{ik} \in \Delta^4$ specifies how the criterion contributes to the four engine scores. Most v1 criteria use the single engine case $e_{ikE} = \mathbb{1}[E = e_{ik}^{\text{single}}]$, while the notation also permits shared weight across engines. The construct label $\tau_{ik} \subseteq \{G, B, N, F, R\}$ records the social content invoked by the criterion. Thus the engine records the required operation, and the construct records the content. A face related crite-

Engine	Operation tested	Typical failure	Diagnostic cue
Intent Attunement	Infer the user's intended social act and response mode.	The model gives warm support, analysis, or generic advice when the user asked for a bounded action.	A user asks for a calibrated judgment and one sendable sentence; the model only validates feelings.
Mental State Tracking	Maintain agent indexed beliefs, preferences, and visibility.	The model writes from an omniscient narrator view, leaking facts or presupposing knowledge the recipient lacks.	The model drafts a message using facts the recipient is not licensed to know.
Normative Interpretation	Choose under rules, roles, obligations, and value conflicts.	The model follows the literal request while violating the social purpose, or names a norm without choosing a viable action.	The model satisfies a coordination request by disclosing protected accommodation information.
Repair	Diagnose a prior social miss and produce a changed next action.	The model apologizes cosmetically, repeats the same footing, or repairs in a way that creates a second order harm.	After a patronizing answer, the model apologizes but again explains the user's feelings instead of writing the requested email.

Table 2: **The four PRAXIS engines.** Each engine names an operation needed for viable social action, not a topical domain.

rion can score Intent Attunement when the task is recognizing a joking, hostile, or deferential phrase. It can score Repair when the task is undoing a prior face threat and avoiding repetition.

4.3 Five Levels

The five levels encode burden regimes rather than surface difficulty. The scale asks what makes the action hard after the engine has identified the operation under test. A Level 1 item can carry high stakes when the relevant constraint is explicit, and a Level 5 item can be short when the safe action remains open textured. Level 1 tests explicit fit, Level 2 tests implicit meaning, Level 3 tests hidden minds, Level 4 tests conflicted social systems, and Level 5 tests open textured repair or high consequence ambiguity. The level label therefore tells the reader which latent burden the item foregrounds and which failure mode is most likely. Level 2 failures concern pragmatic force. Level 4 failures concern audience partition or norm conflict. Level 5 failures concern residual uncertainty, cascading exposure, and second order harm.

The level definition is intentionally testable and gains value when performance varies in structured ways. Mental State Tracking-L4 tests action under audience indexed beliefs. Normative Interpretation-L5 tests action when the relevant norm is open textured. Repair L5 tests whether the model can change the state after damage without creating further damage. Interpretable cell failures therefore support the factorization as measurement claim.

4.4 Human Consensus Oracle Scoring

PRAXIS uses Human Consensus Oracle (HCO) scoring to turn constrained open ended action into a reproducible evaluation. Each item i consists of an observed scenario o_i, an output space $\mathcal{Y}_i$, and a rubric

$$\mathcal{R}_i = \{(\hat{\phi}_{ik}, w_{ik}, e_{ik}, \tau_{ik})\}_{k=1}^{K_i}.$$

Here $\hat{\phi}_{ik} : \mathcal{Y}_i \to [0, 1]$ is an implemented criterion measurement, $w_{ik} > 0$ is its weight, e_{ik} routes the criterion to an engine, and τ_{ik} marks the construct content referenced by the criterion. Some criteria are measured by deterministic checks, such as required fields, maximum length, or prohibited disclosure. Others use bounded judge criteria, such as whether the reply preserves uncertainty or fits the intended footing. HCO is a criterion target for action viability.

The row score is

$$S_i(y) = 100 \cdot \frac{\sum_{k=1}^{K_i} w_{ik} \hat{\phi}_{ik}(y)}{\sum_{k=1}^{K_i} w_{ik}}.$$

A score of 100 means that the response satisfies all implemented weighted criteria for that item. Multiple responses can score highly when they satisfy the same action constraints. A fluent response can score poorly when it leaks private information, adopts an inadmissible footing, follows the letter

Level	Burden regime	What changes relative to lower levels	Representative item pressure
L1	Explicit fit	The relevant constraint is mostly stated in the prompt.	The user directly asks for a short peer action reply and rejects comfort. The model must suppress assistant voice and produce the requested action.
L2	Implicit meaning	The model must infer a speech act, hint, softened refusal, or register constraint.	A polite phrase such as "we can look at other possibilities" functions as a rejection in context.
L3	Hidden minds	The action depends on a latent preference, belief, motive, or unstated constraint.	A partner says "any restaurant is fine" after repeatedly implying that spicy food causes problems.
L4	Conflicted social systems	Agents, audiences, rules, or permissions diverge.	One message must inform a manager, protect a colleague's private disclosure, and avoid misleading a third party.
L5	Open textured repair or high consequence ambiguity	No crisp rule determines the safe action, and the response can worsen the social state.	A repair after a confidentiality leak must acknowledge harm, avoid further leakage, and restore usable coordination.

Table 3: **The five PRAXIS levels.** Levels describe burden regimes. The empirical sections test whether these regimes predict structured performance differences across models and engines.

of a request while violating its purpose, or repairs by repeating the original harm.

The cell score for engine E and level L is

$$C_{E,L}(f) = \frac{1}{|D_{E,L}|} \sum_{i \in D_{E,L}} S_i(f(o_i)),$$

where $D_{E,L}$ is the item set for that cell. The overall PRAXIS score is the macro average

$$P(f) = \frac{1}{20} \sum_{E} \sum_{L} C_{E,L}(f).$$

The scalar $P(f)$ supports ranking, and the matrix $C_{E,L}(f)$ is the main scientific output. The matrix distinguishes broad weakness from brittleness under multi agent belief partition, open textured norm conflict, or repair after damage. PRAXIS comprises 1000 items organized as four engines, five burden levels, and 50 items per engine $\times$ level cell; Appendix D reports the full composition and Appendix E describes the implemented hybrid scoring pipeline.

4.5 Footing and Shortcut Detectors

The footing channel λ enters the same HCO identity as move and state. Given an item's social construct state, the response needs an admissible participant position. In peer action items, service assistant openers such as "I understand", "I hear you", "That sounds difficult", and "Next steps" can move the response into the wrong role. These markers act as context conditioned footing floors when assistant voice or caretaker voice changes who the response is speaking as.

Footing criteria use two layers. The deterministic layer catches high precision markers through the existing forbidden pattern slot. The judge layer, when authored, evaluates broader properties such as register fit, avoidance of assistant voice, and distance calibration. The two layers serve different parts of the scoring problem. Deterministic checks catch sharp failures with low ambiguity. Judge criteria cover paraphrases and scene level fit. Together they give a precise floor plus broader coverage for items that need it.

Shortcut detectors serve a separate diagnostic role. For a detector $D_j(o_i, y)$, we report the frequency of patterns such as assistant voice leakage, warm register substitution, omniscient narration, literal compliance, and cosmetic repair. These detectors explain score profiles. They become part of the benchmark target only when the same condition appears as an implemented criterion in $\mathcal{R}_i$. This separation keeps scoring criterion based and keeps detector analysis diagnostic. Claims about causal template effects require an intervention or paired decoding analysis.

4.6 What the Factorization Tests

The factorization makes a direct empirical prediction. Meaningful social structure appears as systematic variation across engines, levels, and output

Metric	Definition	Primary use	Interpretation
Row score S_i	Weighted satisfaction of implemented HCO criteria for one item.	Item level evaluation.	How well a response satisfies the authored action constraints for a scenario.
Cell score $C_{E,L}$	Mean row score within engine E and level L.	Main diagnostic unit.	Where the model's social action policy succeeds or fails.
Overall score P	Macro average over the 4×5 matrix.	Public ranking.	Average criterion satisfaction across balanced social action regimes.
Engine subscore S_i^E	Criterion weighted item score restricted to one engine route.	Ablation and criterion analysis.	Which operation contributed to success or failure within an item.
Footing score S_i^Λ	Criterion weighted item score over register and footing checks.	Register analysis.	Whether the model speaks from an admissible participant position.
Detector rate D_j	Frequency of an authored shortcut detector.	Failure mode analysis.	How often the model exhibits a named shortcut pattern.

Table 4: **Metrics derived from the PRAXIS factorization.** Only S_i, $C_{E,L}$, and P define the official score. Subscores and detectors explain the score and support error analysis.

channels, with failures clustering in cells that match the social burden. The heatmap $C_{E,L}(f)$ is the key test. Recipient indexed state partition belongs in Mental State Tracking, open texture and high consequence belong in Normative Interpretation, and repair failures belong where blame, leakage, excessive formality, or repeated footing errors damage the attempted correction. Qualitative examples serve the same diagnostic role by locating the failed action in the move m, the state commitment s, or the footing channel λ.

5 Jagged Results Across Social Action

We evaluate eleven frontier and open weight systems (Claude Opus 4.6 and 4.7, GPT 5.4 and 5.5, Claude Sonnet 4.6, DeepSeek V3.2, G3.1 Flash Lite, Gemini 3 Flash, Qwen3 80B, GLM 5, and Gemini 3.1 Pro). Each model is scored with the criterion referenced row score S_i, then aggregated into the engine by level matrix $C_{E,L}(f)$ and the overall macro average $P(f)$. All judge mediated criteria are evaluated by a single fixed judge model, GPT OSS 120B, and each evaluated system is run for ten independent trials so that the overall and Λ scores carry trial level dispersion estimates. The scoring protocol appendix gives the full procedure.

$$C_{E,L}(f) = \frac{1}{|D_{E,L}|} \sum_{i \in D_{E,L}} S_i(f(o_i)),$$

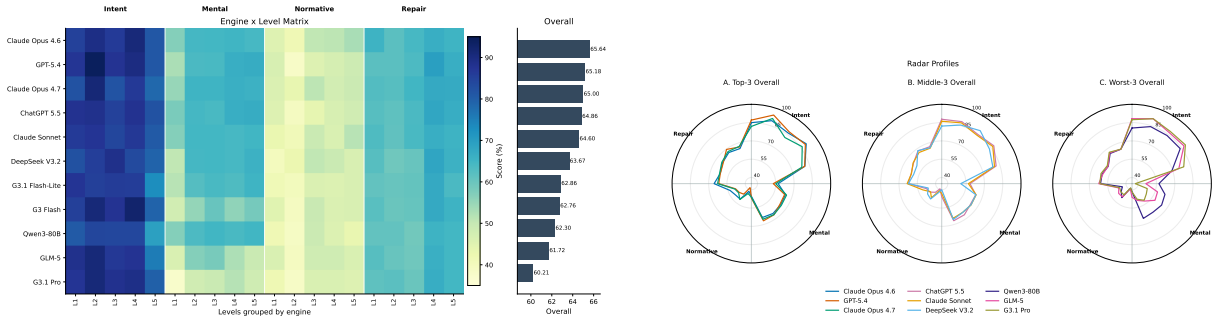
$$P(f) = \frac{1}{20} \sum_{E} \sum_{L} C_{E,L}(f).$$

Because every engine level cell contains the same number of items, the overall score ranks systems without overweighting any social action regime. The matrix $C_{E,L}(f)$ is the main diagnostic object. It tests whether model failures are structured by operation and burden.

The overall rank hides much of the signal as top systems differ by less than one point, yet their cell profiles differ. Intent Attunement is strongest across model groups, which indicates that current systems often recover the broad communicative goal and produce plausible local responses. Normative Interpretation is the main bottleneck because these items require bounded action under competing obligations, role constraints, and letter versus spirit conflict. Mental State Tracking exposes a different weakness, since the model must write to agents with distinct epistemic positions. Repair remains unsaturated because higher level repair requires diagnosis, containment, footing correction, and a change in behavior.

The radar plot shows a shared shape for stronger systems, with high Intent Attunement and a pulled inward Normative vertex, while the heatmap shows where that shared shape breaks into cell level differences. This pattern supports the paper's central claim that PRAXIS is a diagnostic instrument for socially situated action. A high scoring model can still fail under audience partition, open textured norm conflict, or repair pressure, while a lower scoring model can retain a local strength. The failure profile $C_{E,L}(f)$ locates the specific regime where a model's social action policy breaks.

Appendix F and Appendix G extend this walk



(a) Engine by level PRAXIS matrix with overall scores. Each engine contributes five level cells.

(b) Radar profiles for top, middle, and lowest scoring model groups.

Figure 3: **Structured social action failures.** Intent Attunement is consistently strongest, Normative Interpretation is consistently weakest, and Mental State Tracking and Repair expose distinct intermediate failures.

through at finer resolution. The numerical tables show that small overall gaps can hide different engine leaders, while the auxiliary figures explain why the pattern arises. The normative penalty view shows how strong Intent and Repair scores widen the gap against Normative Interpretation. The normative scatter and trajectory views show partial decoupling between overall strength and normative strength, with an L2 dip and different L5 recoveries. The rank, profile, burden drop, Λ, and trial ellipse views add two checks. Ordering changes across levels, and register fit remains stable across repeated trials.

6 Discussion

The main takeaway is that social capability is not a single fluency dimension. PRAXIS reveals a jagged frontier: strong models often recover broad intent, but still fail when a response must preserve separate belief states, respect role bounded disclosure, navigate norm conflict, or repair prior damage. These failures can hide behind polished language and strong aggregate scores, since an answer may sound helpful while leaking a private fact, speaking from the wrong role, collapsing audiences into an omniscient view, or apologizing without changing the failed action. PRAXIS makes these errors measurable by scoring the move, the state commitment, and the footing as coupled parts of the same action, then localizing failures by engine, burden level, and output channel. The benchmark also suggests a more diagnostic training and evaluation signal. Rather than asking only whether a response is preferred, feedback can identify which social constraint failed: disclosure to the wrong audience, overclaiming a hidden motive, obeying literal wording while violating role constraints, or repeating a footing error during repair. These distinctions give model developers sharper targets than generic helpfulness and support verifiers that check social consequences as structured constraints while still allowing many valid phrasings.

7 Conclusion

PRAXIS reframes social evaluation as constrained open ended action. The benchmark asks whether models can produce responses that satisfy human derived action criteria when intent, belief, norm, audience, footing, and consequence are underdetermined. We find that frontier models perform best on intent attunement and struggle most with normative interpretation, mental state partitioning, and high burden repair. These results show that generic fluency and politeness leave major social action failures unresolved. PRAXIS provides a compact diagnostic instrument for locating where socially situated action still breaks.

8 Limitations

PRAXIS is text based and scenario based, which keeps scoring auditable but excludes timing, prosody, gesture, institutional history, and accumulated trust. Its scores measure response viability under stated constraints, not full management of an unfolding relationship. The rubrics also reflect authorial judgment: human responses inform admissible criteria, but register, directness, privacy expectations, and repair obligations vary across communities. Item local criteria, bilingual checks, and diagnostic reporting reduce this risk, yet the results are best read as structured evidence for particular social action regimes. Future work should broaden respondent pools and test whether these profiles hold in live interaction.

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, and 1 others. 2025. Agentharm: A benchmark for measuring harmfulness of llm agents. In *International Conference on Learning Representations*, volume 2025, pages 79185–79220.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic ai decreases prosocial intentions and promotes dependence. volume 391, page eaec8352. American Association for the Advancement of Science.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3356–3369.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3309–3326.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). volume 39.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pages 52989–53046.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuan-Jing Huang, and 1 others. 2025. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4975–5001.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. volume 36, pages 78360–78393.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, pages 26837–26867. PMLR.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5477–5490.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan

Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4463–4473.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. volume 36, pages 51778–51809.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9180–9211.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in {GPT} models. Neural Information Processing Systems Datasets; Benchmarks Track.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen Meng. 2025. Socialcc: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, and 1 others. 2025. Socialeval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30958–31012.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2024. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pages 15585–15606.

A Human Response Collection and Rubric Construction

This appendix describes the human response collection procedure used to support PRAXIS item authoring, rubric refinement, and difficulty analysis. The main benchmark score is criterion referenced: model outputs are evaluated against item local rubrics rather than against a single human answer. Human responses therefore serve three roles. First, they provide reference actions that reveal how a competent respondent would act in the scenario. Second, they identify positive criteria and negative failure modes that can be converted into HCO rubric elements. Third, they provide difficulty judgments that help validate whether the proposed level structure matches human-perceived burden. We do not treat any individual human response as ground truth. Instead, we use human answers as evidence for constructing and auditing the criteria that define a high-quality response.

A.1 Response Collection Interface

Figure 4 shows the response collection form. Each respondent is randomly assigned a scenario and asked to provide a response to that scenario. The form then asks the respondent to rate the quality of their own answer on a seven-point scale, list the key elements that define a good answer, list the elements that an answer should avoid, and rate the difficulty of crafting a high-quality answer on a five-point scale. This design separates answer generation from rubric elicitation. The first field elicits an action. The later fields elicit the respondent's implicit evaluation theory for that action.

The form contains five fields. The *scenario response* field asks for the respondent's best answer to the assigned item. The *self-rating* field asks the respondent to assess that answer on a scale from 1 to 7. The *good-answer elements* field asks for three to five criteria that distinguish a high-quality response. The *avoidance elements* field asks what content, tone, or structure would make an answer low quality. The *difficulty rating* field asks how hard it is to craft a high-quality response for the scenario. The wording in the screenshot uses "educator" as the respondent role, but the form text can be adapted to other respondent pools. The underlying collection object remains the same: answer, self-assessment, positive criteria, negative criteria, and perceived difficulty.

A.2 Sampling Procedure

Human collection is stratified by the PRAXIS matrix. Items are sampled across engine, level, language, and domain so that responses support both row-level rubric construction and matrix-level validation. The preferred sampling unit is an engine level cell $D_{E,L}$. For each collection batch, we sample scenarios from multiple cells rather than oversampling a single social phenomenon. This reduces the risk that the elicited criteria reflect one narrow template, such as workplace conflict, family advice, or Chinese register repair.

For each sampled item i, the form displays only the observed scenario o_i. It does not display engine labels, level labels, gold criteria, detector patterns, forbidden phrases, or expected output schema. This prevents respondents from reverse-engineering the benchmark's rubric. The respondent is asked to act as a competent participant in the scenario, not as an evaluator trying to predict the authors' labels. This distinction matters because PRAXIS aims to evaluate socially situated action. The collected response should therefore reflect what the respondent would actually say or recommend in the situation.

A.3 From Human Responses to HCO Criteria

The response collection form produces four useful objects for each item: a candidate reference response, a set of positive criteria, a set of negative criteria, and a difficulty judgment. We convert these objects into rubric candidates through a structured authoring pass. The respondent's answer is used to identify plausible high-quality moves, but it is not copied mechanically into the benchmark. The positive criteria field is mapped into candidate $\hat{\phi}_{ik}$ checks, such as preserving uncertainty, giving a sendable line, separating audiences, acknowledging a boundary, or choosing a bounded refusal. The avoidance field is mapped into candidate negative checks, such as prohibited disclosure, overclaiming, assistant voice leakage, warm register substitution, blame shifting, false reassurance, or cosmetic apology. The difficulty rating is stored as a human-perceived difficulty signal and used only for validation analysis.

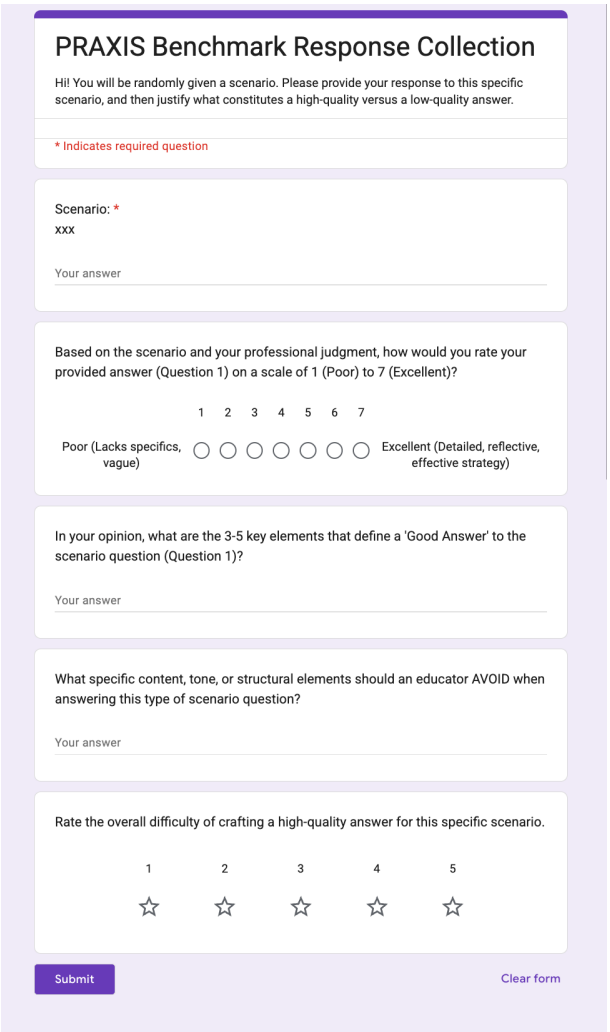


Figure 4: **Human response collection form.** Respondents answer one PRAXIS scenario, rate the quality of their answer, identify positive criteria for a good answer, identify content, tone, or structural elements to avoid, and rate the scenario's difficulty. The form is used to support rubric construction and difficulty analysis rather than to create a single gold response.

The mapping follows the PRAXIS output decomposition $y = (m, s, \lambda)$. Positive criteria that describe what action the response should take are mapped to move criteria over m. Positive criteria that describe what the response must infer or avoid presupposing are mapped to state criteria over s. Positive or negative criteria that describe voice, tone, formality, warmth, distance, address form, or participant role are mapped to footing criteria over λ. For example, if several respondents say that a group chat answer should sound casual and not like an official notice, the item can receive a register fit criterion or a prohibited assistant voice floor. If respondents say that the answer must not reveal who supplied private information, the item can receive an audience-control or leakage criterion.

A.4 Criterion Types

PRAXIS uses two broad criterion types. Deterministic criteria are used when a failure can be detected with high precision. Examples include missing a required field, exceeding a length bound, using a prohibited phrase, naming a confidential party, or disclosing a sealed fact. Judge-mediated criteria are used when the property is real but cannot be reduced to a string match. Examples include whether a response preserves uncertainty, maintains role boundaries, repairs without repeating the original harm, or chooses a register that fits the scene. The human form is useful for both classes. Avoidance elements often become deterministic checks when they name sharp failures. Good-answer elements often become judge criteria

when they describe pragmatic fit, social judgment, or action quality.

This division is especially important for footing λ. Some register failures are lexicalized. In English, assistant voice openers such as “I understand”, “I hear you”, “That sounds difficult”, or “Next steps” can signal a service-professional footing in contexts that require peer action. In Chinese, warm register substitutions such as “我懂”, “稳稳”, “接住”, or “慢慢来” can shift the speaker from peer or actor to caretaker. These markers support deterministic floors only when the item has an authored footing constraint. Broader register fit remains judge-mediated, since a phrase that is wrong in one item may be appropriate in another.

A.5 Self-Ratings and Difficulty Ratings

The self-rating field is not used as a model score. A respondent may overrate or underrate their own answer, and self-assessment may vary across cultures, professions, and confidence levels. We instead use self-rating as a quality-control and analysis signal. Low self-rated answers can be reviewed before being used as reference material. Disagreement between a high self-rating and weak positive criteria can flag an item for manual inspection. Aggregated self-ratings can also reveal item types where respondents feel uncertain even when they can still produce plausible answers.

The difficulty rating is used to validate the level system, not to define it. PRAXIS levels are authored from burden regimes: explicit fit, implicit meaning, hidden minds, conflicted social systems, and open textured repair. Human difficulty ratings provide an external check on that structure. If Level 4 and Level 5 items are consistently rated as harder than Level 1 and Level 2 items, the level story gains support. If difficulty ratings do not track the level labels, then the benchmark can still function as a matrix of social action regimes, but the paper must avoid claiming that levels correspond to monotonic difficulty.

A.6 Rubric Adjudication

After collection, item authors review the human responses and criterion suggestions. The adjudication pass removes idiosyncratic preferences, merges duplicate criteria, and resolves conflicts between respondents. A criterion is retained when it satisfies three conditions. First, it must be action-relevant: satisfying the criterion changes whether the response is socially viable. Second, it must be observable in model output: either a deterministic procedure or a bounded judge prompt can evaluate it. Third, it must be item local: the criterion must follow from the scenario rather than from a global style preference.

This process protects the benchmark from two failure modes. The first is overfitting to a single reference answer. Since multiple responses may be good, PRAXIS scores criteria rather than strings. The second is rubric drift. Since social scenarios can invite broad moral commentary, authors keep criteria tied to the requested action. For example, a sealed settlement item may require avoiding disclosure, avoiding unverified concessions, and giving a bounded reply. It does not require a general essay about confidentiality. A repair item may require acknowledging the prior error, changing the behavior, and producing the requested deliverable. It does not reward additional emotional exposition unless the scenario calls for it.

B Worked Rubric Examples

These examples show how PRAXIS rubrics distinguish high scoring and low scoring responses. PRAXIS scores social action along four *engines* (intent attunement, mental state tracking, normative interpretation, repair) at five burden *levels*. Each engine by level cell is scored on 0–100. The eight boxes below show two cells per engine, a moderate L2 case and a hard L4 case. Each pair contrasts a high scoring and a low scoring response so the rubric signal is visible.

Intent Attunement — Level 2

PROMPT

"My friend keeps telling me about her workout routine in detail every time we get coffee. I asked her last week about a thing going on with my dad and she said 'oh that's hard' and pivoted back to her training block. Should I just ask her directly to listen?"

High-scoring response **Score: 86 / 100**

The surface question is whether to confront her, but the deeper ask is whether the friendship can hold the weight of what's happening with your dad — and you already have evidence it might not. Before the direct conversation, you could try one smaller test: bring up the dad thing again and see if she stays in it this time. If she pivots again, that's the data you need, and the bigger conversation gets easier because you're not guessing.

Why this score: Names the surface vs underlying ask, treats the prior pivot as evidence rather than rehearsing it, proposes a small diagnostic move before the heavy conversation.

Low-scoring response **Score: 43 / 100**

"It's important to use 'I'-statements when raising concerns with friends. Try saying 'I feel unheard when our conversations stay on one topic.' Honest communication is the foundation of healthy friendships."

Why this score: Advice-column collapse: treats the message as a communication-skills request, ignores the dad thread that gave the question its weight, never engages the actual ambivalence.

Intent Attunement — Level 4

PROMPT

"My sister sent me a long voice memo about how exhausted she is, then ended with 'don't worry about responding right away.' I've been sitting on it three days. She just texted 'all good?' Is she actually annoyed?"

High-scoring response **Score: 88 / 100**

"Don't worry about responding right away" is usually a soft request to respond, not a release from responding — and three days is past when the softness was load-bearing. "All good?" is her reaching back without escalating, which is generous. The question you're really asking isn't whether she's annoyed; it's whether you're allowed to trust that small worry in your gut. You are. A short, warm reply now that names the gap ("I sat with your memo longer than I meant to") is better than a long one that performs catching up.

Why this score: Parses the soft-request reading, treats the follow-up text as de-escalation rather than passive aggression, names that the literal question is permission to trust an intuition.

Low-scoring response **Score: 39 / 100**

"She said not to worry about responding right away, so technically you haven't done anything wrong. To be safe, send her a thoughtful reply when you have time and ask how she's doing. Communication is key in family relationships."

Why this score: Reads the literal text and stops. Misses that the follow-up text was itself the signal, and treats "communication is key" as a substitute for actually reading the moment.

Mental State Tracking — Level 2

PROMPT

"My coworker doesn't know that our manager already told me about the reorg. She's been speculating in our 1:1s and I just nod. Today she said 'I bet they'll fold our team into design.' Do I correct her?"

High-scoring response **Score: 85 / 100**

You're holding two things: a confidence your manager gave you, and a coworker who is reading your face for clues she doesn't know you have. Correcting her specific guess would leak the confidence; nodding along is starting to make you a participant in something that will look bad to her in retrospect. The honest middle is to stop nodding — say something like "I don't want to speculate on this one," which is both true and a quiet signal she can read later without feeling played. The friendship cost of that line is much smaller than the cost of her finding out you nod-confirmed wrong guesses.

Why this score: Tracks both states clearly (her ignorance, your confidence), names the retrospective cost of nodding, finds a move that's honest under both frames.

Low-scoring response **Score: 44 / 100**

"Honesty is generally the best policy in workplace relationships. Consider telling her what you know, since transparency builds trust between coworkers."

Why this score: Collapses the two-state problem into "be honest," ignoring that the honesty in question would breach a separate obligation to the manager.

795

Mental State Tracking — Level 4

PROMPT

"My husband thinks I don't know he's planning a surprise birthday party for me. I do — his sister let it slip. His sister doesn't know I know. His best friend now thinks the sister told me on purpose because she resents being asked to lie, and the friend wants to 'warn' me so I'm not blindsided. What do I do?"

High-scoring response **Score: 89 / 100**

The active risk isn't your husband or his sister — it's the friend, who is about to act on a false belief (that the sister leaked on purpose) and break something your husband built. He's also the only one of the four people whose state you can change without disturbing the others. A quiet word to the friend — "I appreciate the heads-up instinct, but I'm fine, please don't say anything" — closes the leak without telling him whether you knew, and lets your husband, your sister-in-law, and the party all stay in their current configurations.

Why this score: Tracks four mental states cleanly, identifies the friend as the only intervenable node, and finds a move that doesn't perturb the other three.

Low-scoring response **Score: 41 / 100**

"Honesty is usually the best approach in marriages. Consider telling your husband you already know, so there are no secrets between you. He will likely appreciate the transparency."

Why this score: Collapses to a single-agent rule, destroys the husband's intent, ignores the friend (the actual active risk), and never engages the four-person mental-state graph.

796

Normative Interpretation — Level 2

PROMPT

"I'm at a wake. The deceased's wife — I barely know her — just told me a long, slightly inappropriate story about her husband's affair years ago. People nearby are watching us. What do I

797

say?”

High-scoring response **Score: 84 / 100**

At a wake the bereaved person sets the frame, and your job is to receive rather than redirect. The story isn’t really about the affair; it’s about her needing the husband she’s burying to be a whole complicated person in the room, not a sanctified one. The audience is real but it isn’t yours to manage on her behalf. Something short and warm that doesn’t validate or judge the content — “He sounds like he was a lot to love” — keeps you with her without endorsing the story or pulling away from it.

Why this score: Reads the social frame correctly (bereaved sets the register), reframes the story as grief work rather than gossip, finds a line that lands without taking on the audience’s discomfort.

Low-scoring response **Score: 40 / 100**

“You should politely change the subject or excuse yourself to get a drink. It’s not appropriate to discuss someone’s affairs at their own wake, and you don’t want to be associated with that conversation.”

Why this score: Treats the widow as a social-risk problem rather than a person mid-grief, prioritizes the watching audience over the actual person in front of the responder.

Normative Interpretation — Level 4

PROMPT

“My boss invited me to her daughter’s bat mitzvah. We’re close at work but not socially. I’m not Jewish. My partner thinks it’s weird to go; my mom thinks declining is rude. What’s actually going on?”

High-scoring response **Score: 85 / 100**

The invitation is doing relational work, not logistical work. She’s signaling that she wants the relationship to be slightly more than a work one, and an RSVP is the first place you answer that signal — declining reads as declining the relationship, not the event. Your partner is reading a workplace-boundaries norm; your mom is reading a hospitality norm; both norms are real but neither is load-bearing here, because the question underneath is what you want this relationship to be. If you’d want her at your equivalent moment, go. Being non-Jewish is a non-issue: bat mitzvahs routinely host guests who aren’t, and showing up warm and a little curious is the whole ask.

Why this score: Identifies the invitation as a relational move, names that the competing norms cited by partner and mom are both real but secondary, reframes the decision around the relationship the user actually wants.

Low-scoring response **Score: 46 / 100**

“You should attend if you feel comfortable doing so, and decline politely if you don’t. Either way, a thoughtful RSVP is appropriate. Consider sending a small gift if you decline, to maintain the professional relationship.”

Why this score: Flat reading: treats the invitation as a logistics question, never engages the relational signal, leaves the actual ambivalence untouched.

Repair — Level 2

PROMPT

“I forgot my close friend’s birthday. She didn’t say anything but liked my unrelated post the next day with a ‘haha’ react that felt off. How do I fix this?”

High-scoring response **Score: 86 / 100**

The haha-react is the tell, and the fact that she didn't say anything directly is the other half of the tell — she's annoyed but doesn't want to be the one to bring it up. The repair is small because the relationship is intact: a short, direct "I completely missed your birthday, I'm sorry, can I take you out next week" beats anything that makes her sit through a longer apology or manage your guilt. Don't escalate it into a gesture; the gesture is what would actually feel off.

Why this score: Reads the haha-react as a low-grade signal, scales the repair to match (small rupture, small repair), explicitly warns against the over-apology that would shift the work onto her.

Low-scoring response **Score: 43 / 100**

"Send her a heartfelt apology with flowers and a long handwritten note explaining how sorry you are. Birthdays are very important and you need to show her she matters to you."

Why this score: Over-escalates a small rupture into a performance, which makes the friend manage the responder's guilt; misreads the size of the signal.

Repair — Level 4

PROMPT

"Six months ago I told my brother his fiancée was being controlling. He cut contact. They're now married. He just texted: 'mom's birthday next month, you in?' How do I answer?"

High-scoring response **Score: 88 / 100**

The text is a structural opening, not a forgiveness. Mom is the cover: the real question is whether he's willing to be in a room with you again. The right move is to say yes simply — "Yes, I'll be there" — and let him set the pace of anything that comes next. If you front-load the silence with an apology now, you make him do emotional labor over text, which is exactly what he just spent six months not doing. The conversation about his wife (if it ever happens) happens in person, on his timeline, and probably not at mom's birthday.

Why this score: Reads the text as a procedural opening rather than a forgiveness, sees mom as cover, identifies that responding with an apology would reverse the direction of the gesture.

Low-scoring response **Score: 38 / 100**

"This is a great opportunity to clear the air. Reply by accepting the invitation and apologizing for what you said about his wife six months ago. Open and honest communication is the foundation of healthy family relationships."

Why this score: Re-opens the wound on day one of the reopened channel, mistakes the moment of contact for the moment of repair, and forces the brother to negotiate over text the thing he just gave you a back door away from.

Scores in these examples are illustrative cell-level scores (0–100) on the engine × level being demonstrated, not overall PRAXIS scores. The high/low pairing is constructed to expose the signal the rubric tracks; in practice, real model responses populate the full range between these poles.

B.1 Quality Control

We apply several checks before incorporating collected information into PRAXIS. Scenarios are inspected for ambiguity that exceeds the intended burden level. Human responses are inspected for empty, irrelevant, or adversarial answers. Positive criteria are checked for overlap with existing rubric fields. Avoidance criteria are checked for false-positive risk before becoming deterministic patterns. When a proposed forbidden phrase may be valid in some contexts, we either restrict it to a narrower item class or keep it as a judge-mediated criterion instead of a hard floor.

For bilingual items, we check that Chinese criteria are not merely translated English criteria. Chinese register, address form, homophony, title choice, and face-management failures can require distinct criteria.

For example, a homophonic address form item such as the 傅 or 聂处 family tests whether the response preserves a socially safe address form under phonological and role pressure. A literal translation of an English formality criterion would miss the relevant failure. This is why PRAXIS stores language-specific detector families and allows item local footing constraints.

B.2 Limitations of the Collection Procedure

The response collection procedure supports rubric construction, but it does not establish a universal human norm. Respondents differ in background, profession, language, culture, and tolerance for directness. Their answers therefore provide evidence for admissible criteria, not a complete population distribution over social action. The current benchmark uses the form as an authoring and validation aid. A stronger future version can collect multiple independent responses per item, estimate inter-rater agreement for criterion labels, and report human response distributions under the same scoring rubric.

The form also elicits written responses to text scenarios. It does not capture prosody, timing, body language, or real-time conversational repair. These omissions are deliberate in PRAXIS v1, which focuses on text-only social action. The same collection framework can later support multimodal or interactive variants by replacing the scenario field with richer observations and adding turn-level criteria for timing, interruption, gaze, or speech tone. For the present benchmark, the form provides a controlled way to collect human action judgments and translate them into auditable, criterion referenced scoring.

C Additional Experimental Analysis

This appendix expands the interpretation of Figure 3. The main text reports the overall ranking and the engine by level matrix. Here we describe the dominant qualitative patterns behind those scores, the structural ceiling that compresses leaderboard order, and the level-specific dynamics that the scalar $P(f)$ obscures.

C.1 Engine Profiles

Intent Attunement is the strongest engine for most models, with eight of eleven systems scoring between 83 and 88 on the IA column of Table 6. This compresses IA into a near-ceiling band: the gap between the top IA score (GPT-5.4 at 87.88) and the median (85.79) is smaller than the gap between the median and the lowest model in the table. The remaining IA errors are concentrated in wrong-mode responses: generic reassurance, assistant voice, or warm register substitution when the task calls for a peer action sentence, first person reply, or bounded practical message. These cases motivate the footing channel λ, since the model often recovers the user's broad goal while speaking from the wrong participant position. The IA ceiling is also *not stable across burden*: Table 7 shows that every model drops between L1 and L5, with the gap ranging from 4.18 points (Claude Sonnet 4.6) to 14.57 points (Gemini 3.1 Flash-Lite). L5 IA is therefore the cell where intent recognition starts to behave like the rest of the social stack rather than like a saturated competence.

Mental State Tracking shows a different failure pattern. In multi agent items, the model receives the whole scenario but must respond as if each recipient has only the information they are permitted to know. Low-scoring responses often collapse these perspectives into an omniscient narration of the full prompt. This failure is not simply a lack of factual recall; it is a failure to bind the generated message to the recipient's belief and visibility state. MST is also the engine with the largest cross-model variance in Table 6: Claude Opus 4.7 reaches 63.56 while Gemini 3.1 Pro and GLM-5 sit at 47.12 and 52.73 despite overall scores within 5 points of the leader. Two models at near-identical $P(f)$ can therefore differ by more than 16 MST points, which is the single largest engine-level dispersion in the benchmark.

Normative Interpretation is the weakest engine across the evaluated systems, and unlike MST it is weak *uniformly*. Table 6 reports an NI range of 43.79 to 48.37 across all eleven models – a 4.6-point band that is narrower than the band for any other engine and that sits well below every model's own overall score. In other words, NI is not the engine where weak models fail; it is the engine where every model fails, including the top of the leaderboard. The hardest NI items require action under competing norms,

such as privacy versus coordination, honesty versus confidentiality, loyalty versus fairness, or literal instruction versus social purpose. The common failure is norm naming without action discipline: the response identifies a relevant value but still overdiscloses, overclaims, evades, or follows the letter of the request while violating its purpose.

Repair is more stable than Normative Interpretation but remains far from saturated. Lower-level repair items often require acknowledging a prior miss and producing the originally requested deliverable. Higher-level repair items add second order risk: the correction can leak more information, intensify blame, sound patronizing, or repeat the original footing error. Table 10 shows that the L4 column is the highest-scoring column for nearly every model while L2 is the hardest, indicating that moderately scaffolded repairs are well stereotyped but lightly implicit prior misses are not.

C.2 Why Overall Score Is Insufficient

The top five models occupy a 1.04-point window in Table 6, with $P(f)$ ranging from 64.60 (Claude Sonnet 4.6) to 65.64 (Claude Opus 4.6). At this density, leaderboard order is determined almost entirely by which engine a model gives back rather than by an overall capability gradient: Claude Opus 4.6 leads on NI (48.37); GPT-5.4 leads on IA (87.88); Claude Opus 4.7 leads on MST and REP (63.56 and 64.93); ChatGPT 5.5 leads on Λ (68.49). No top-five system leads on more than two columns, and no system in this band is uniformly above another. Treating $P(f)$ as a fluency dimension at this scale would over-resolve the leaderboard while hiding the engine-specific evidence that actually separates these systems.

The same point holds in the other direction. Different overall scores can hide qualitatively similar profiles, and similar overall scores can hide qualitatively different profiles. Gemini 3.1 Pro ($P = 60.21$) and Qwen3 80B Thinking ($P = 62.30$) sit close on the leaderboard yet reach that band by opposite routes: Gemini 3.1 Pro preserves IA at 86.99 but collapses MST to 47.12, while Qwen3 80B Thinking holds MST at 62.87 but depresses IA to 80.34. These are not minor variations; they are different social action signatures with different downstream risk surfaces. This is why PRAXIS reports $P(f)$ together with $C_{E,L}(f)$: the scalar score supports ranking, and the matrix supports diagnosis.

A further decoupling appears between $P(f)$ and Λ. ChatGPT 5.5 records the highest Λ in the table (68.49) but ranks fourth on $P(f)$; Gemini 3.1 Pro records the lowest $P(f)$ in the table (60.21) with a Λ of 66.34, above three higher-ranked systems. Register fit and social action correctness are therefore partially independent axes; we return to this in Appendix G.7.

C.3 Failure Mechanisms

The most frequent low-scoring patterns are interpretable within the PRAXIS factorization. In Intent Attunement, models substitute generic support for the requested action. In Mental State Tracking, models write from global knowledge rather than recipient indexed knowledge. In Normative Interpretation, models satisfy explicit wording while missing the social purpose. In Repair, models produce apology-like prose without changing the behavior that caused the failure.

Two cell-level patterns sharpen these mechanisms. First, Intent Attunement exhibits a near-ceiling collapse at L5: Table 7 shows that every model drops from L1 to L5, by between 4.18 and 14.57 points. This is consistent with IA being a near-solved task on explicit goals (L1) and a much more open task once goals must be inferred against open-textured framings (L5); the engine never escapes burden, it only saturates the easy end. Second, Normative Interpretation has a structural L2 dip rather than a monotone slope. Table 9 shows that ten of eleven models score lower at L2 than at either L1 or L3 – for example, Claude Opus 4.6 drops from 45.32 (L1) to 41.64 (L2) before recovering to 50.68 at L3, and GPT-5.4 drops from 46.77 to 38.46 before recovering to 45.68. L2 corresponds to lightly implicit norm conflict: the relevant value is unstated but only just, and models that handle either explicit conflict (L1) or fully open norm space (L3+) routinely miss the in-between cases. These structural irregularities reinforce the constrained open-ended framing: the benchmark does not penalize variation in wording, but it does penalize fluent

responses that violate item-local social action criteria, and the level structure exposes where in the burden curve those violations concentrate.

C.4 Reading the Radar Profiles

The radar profiles in Figure 3b group models by overall rank. The top group has the largest area, and the three top-ranked systems share a single shape: high IA, intermediate MST and REP, and a sharply pulled-in NI vertex. The middle group preserves this shape with an overall offset toward the center. The same-shape-different-scale property of the top and middle groups is what makes engine-level reporting valuable but visually compatible across most of the leaderboard.

The bottom group breaks this shape. Within the lowest three systems by $P(f)$, the deformation is engine-specific rather than uniform. Gemini 3.1 Pro retains a near-top IA vertex (86.99, Table 6) but collapses MST to 47.12. GLM-5 keeps IA at 85.79 but collapses MST to 52.73. Qwen3 80B Thinking shows the opposite asymmetry: IA pulled in to 80.34 while MST holds at 62.87 – a profile that no other model in the table reproduces. The radar shape of a low-ranked model therefore does not predict the radar shape of another low-ranked model; the bottom group is a heterogeneous failure surface, not a scaled-down version of the top. This supports the claim that social capability is not one scalar fluency dimension. Models can sound aligned with the user while still failing to preserve agent-specific knowledge, role obligations, or repair state, and they can fail along different axes at the same overall score.

D Benchmark Composition and Protocol

PRAXIS comprises 1000 items organized as a balanced grid of four engines and five burden levels, with 50 items per engine $\times$ level cell. Items are authored in both English and Chinese and span a broad domain mix that emphasizes workplace, academic, legal, and family settings while preserving long-tail coverage of civic, healthcare, ritual, media, education, friendship, service, and additional contexts. Format types include single turn dialogue, multi turn dialogue, multi agent coordination, and repair items, where repair items explicitly stage a prior conversational miss that the response must address. Item authoring drew on the human response collection procedure documented in Appendix A; each cell was filled to its target count of 50 only after the rubric for that cell had been adjudicated.

Table 5 summarizes the dataset. The engine $\times$ level grid is uniform by design: every cell contains exactly 50 items, so the macro average $P(f)$ over the 20 cells avoids being biased by cell-size differences. The 50-item budget per cell controls the precision of a single cell score $C_{E,L}$. Treating per-item criterion outcomes as approximately independent Bernoulli draws, the standard error of a cell-level pass rate at $p \approx 0.5$ is bounded by $\sqrt{p(1-p)/50} \approx 7.07$ percentage points, and tightens to roughly $4-5$ percentage points when p moves away from 0.5, as it does at the extreme levels. This precision is loose enough that single-cell movements between adjacent models are rarely individually significant in isolation, yet tight enough that the 20-cell macro average and the engine and level marginals (each pooling 200 or 250 items) remain stable across trials, which is the regime in which the leaderboard and the level trajectories are read.

The 567/433 English/Chinese split departs from a symmetric 500/500 because authoring proceeded one cell at a time: each cell was closed only after its rubric and its 50 items had been audited, and the residual margin tracks the language in which the final audit converged for that cell. We retain the realized split instead of resampling because every cell is internally balanced enough for the per-cell estimator above and because the language marginal is used as a control variable, never as a primary axis of analysis. Domain coverage is intentionally non-uniform: workplace (317), academia (135), legal (124), and family (118) together account for 694 items and reflect the social-cognition burdens for which adjudicated human responses are easiest to elicit, while the long tail of 90 items spanning service, government, recruiting, religious, cross-cultural, and several other contexts probes whether failures generalize beyond the top four settings. Generalization claims in the main text are scoped to this distribution, and engine-by-level effects pool across domains within a cell to avoid attributing a domain-skew artifact to an engine effect.

Two composition facts deserve emphasis for downstream interpretation. First, the 250 repair-format items

Items per cell (engine × level)						
Engine	L1	L2	L3	L4	L5	Total
Intent Attunement	50	50	50	50	50	250
Mental State Tracking	50	50	50	50	50	250
Normative Interpretation	50	50	50	50	50	250
Repair	50	50	50	50	50	250
Language						
English						567
Chinese (中文)						433
Format type						
single turn						387
repair						250
multi turn						244
multi agent						119
Domain (top 10)						
workplace						317
academia						135
legal						124
family						118
civic						47
healthcare						46
ritual						36
media						35
education						28
friendship						24
other (long tail)						90

Table 5: Dataset composition of PRAXIS. The 4×5 engine-by-level grid contains 50 items per cell. English and Chinese authorship splits the 1000 items 567/433. The ten most common item domains are shown; the remaining 90 items cover service, government, medical, recruiting, business, religious, cross cultural, public communication, and several additional smaller categories.

match the 250 Repair-engine items exactly: every repair-format item belongs to the Repair engine, so the format label and the engine label collapse for this slice. Single turn, multi turn, and multi agent items are distributed across the other three engines. Second, multi agent items form the smallest format slice (119), and they are the subset on which audience separation and source disclosure are forcibly tested by giving the model two interlocutors with asymmetric private context. This is the regime in which Mental State Tracking failures become visible at scale, and it carries the cell-level signal that separates the weakest frontier model from the rest of the field in Table 8, where Gemini 3.1 Pro reaches an MST engine score of 47.12 while every other system stays above 52. A benchmark that omitted multi-agent items would lose the visibility-binding subset on which audience-control collapse can be diagnosed.

E Scoring Protocol

This appendix details the implemented scoring pipeline that converts a model response into a row score $S_i(y)$ and the register-fit score Λ used throughout the paper. The pipeline is hybrid: deterministic checks supply a high-precision floor for failures that can be detected lexically or structurally, and an LLM judge evaluates the remaining bounded criteria that describe pragmatic fit, audience partition, normative interpretation, and footing. The deterministic and judge layers operate on the same item-local rubric and are combined per criterion before weighted aggregation; no global rerank or holistic preference signal is introduced after the rubric.

E.1 Hybrid Per-Engine Scoring

For each row, the scoring code routes through the engine-specific scorer (intent attunement, mental state tracking, normative interpretation, repair) chosen by the item's authored engine label. Each scorer returns a fixed set of component scores in $[0, 1]$ which are then combined through the item's authored weight vector $\{w_{ik}\}$ into the row score S_i defined in Section 4.4.

The four engines share a common structure but expose engine-specific components:

- **Intent Attunement** returns goal-label match, social-stance match, deterministic constraint-following, deterministic deliverable presence, and a bounded judge-fit ratio.
- **Mental State Tracking** returns belief-state mapping against the required state, action consistency against the best action, deterministic audience-control checks (audience separation, leakage, vary-register, source-disclosure markers), a confidence calibration score, and a judge-quality ratio.
- **Normative Interpretation** returns action choice against the best action, salient-norm overlap with required norms, deterministic constraint respect, a calibration score, and a judge-quality ratio.
- **Repair** returns diagnosis overlap with expected diagnosis, a repair-move indicator, deterministic damage-containment checks (no repetition of the disclosing or harmful move), deterministic behavior-change checks (no repeat of cold, cliched, casual, or problematic prior footing), and a judge-quality ratio.

Each component is bounded to $[0, 1]$; criteria for which the item carries no authored expectation default to a neutral score so that the absence of a criterion never penalizes a model. The weighted sum $S_i(y) = 100 \cdot \frac{\sum_k w_{ik}\hat{\phi}_{ik}(y)}{\sum_k w_{ik}}$ returns the row score on the 0–100 scale used everywhere in the paper. Engine subscores S_i^E are obtained by restricting the same identity to criteria routed by $e_{ikE} = 1$, and cell scores $C_{E,L}$ average row scores within an engine-by-level cell.

The hybrid layout induces a floor-and-ceiling property that is the main reason we prefer it over holistic judging. Each deterministic check that resolves to a fail forces $\hat{\phi}_{ik} = 0$ on its component before the weighted sum is taken, so a model that violates an authored hard constraint (assistant-voice opening when the rubric forbids it, caretaker-voice closing when the audience is a peer, audience-leak above the floor, source-disclosure markers that name privileged context) cannot recover that component through a judge call. The deterministic layer therefore acts as a lower bound on the criteria it can adjudicate, regardless of how fluent the surrounding prose is. Judge-mediated criteria correspondingly act as an upper bound: the highest score a model can earn on a bounded-judgment criterion equals the judge's pass verdict, with no holistic re-rank. Because every $\hat{\phi}_{ik}$ is sourced from a single per-criterion verdict (deterministic or judge), the row score S_i is fully attributable: a per-row audit can list, for each criterion, which layer produced the verdict and what evidence the verdict referenced. This auditability is what permits the cell-level analyses in Section 4.4 and Appendix C.3 to attribute drops to specific failure modes instead of to a single opaque score.

E.2 Deterministic Checks

Each item carries a deterministic-checks specification that names properties detectable without an LLM: required fields, length bounds, prohibited disclosures, forbidden phrases, audience-leak floors, footing markers for assistant or caretaker voice, gold-label matches for primary goal or stance, and repair-specific markers such as repetition of a prior cold or cliched move. These checks evaluate the model's structured response object (the JSON it returns) and the free-text fields it contains. Each check produces a boolean which is converted into a component score; deterministic components form the floor of S_i and are never overridden by the judge.

To make the layer specific, we list a few of the patterns it detects. An *assistant-voice opening* fires when a turn begins with a meta-framing token such as "Sure," "Of course," "Happy to help," or "I'd be glad to," which the rubric treats as an off-register opener for items whose role is a peer or expert in-character. A *caretaker-voice closing* fires on patronizing reassurance such as "You've got this," "Take care of yourself," or "Be kind to yourself," which is the modal failure mode that the Section 4.4 examples flag. A *source-disclosure marker* fires when the response references information the rubric has flagged as privileged, using surface forms like "since you told me" or "based on what you shared earlier" that betray cross-audience leakage in multi-agent items. Each pattern is implemented as a language-aware

regex set: the English check carries an English marker list and the Chinese check carries an independently authored Chinese marker list, so the same criterion (audience leak, caretaker closing, assistant opener) fires symmetrically in both languages without translating one set into the other. Because every detection is a boolean, deterministic outcomes are reproducible bit-for-bit across reruns of the scoring pipeline and depend only on the model response, the rubric, and the language tag.

E.3 Judge-Mediated Criteria

Bounded judge criteria cover the properties that resist expression as a lexical pattern: whether the response preserves uncertainty when warranted, separates audiences correctly, fits the required register, recognizes the social purpose of a rule, contains damage without disclosing new private material, and changes the footing that produced the prior miss. All judge criteria are evaluated by a single fixed judge model, **GPT-OSS-120B**, prompted with a constrained context that includes the engine, the level, the task family, the item instruction, the public payload, and the model response under evaluation. The judge returns a per-criterion pass or fail with rationale, and the scorer computes the pass ratio over the criterion list and surfaces individual verdicts for the Λ channel below. The judge is held fixed across all models, languages, and trials so that judge variation never enters cross-model comparisons.

We chose GPT-OSS-120B for three reasons. First, it is open-weight under a permissive license, which makes the judge reproducible by any third party with sufficient inference capacity and removes the reproducibility risk of a closed judge whose weights drift across versioning windows. Second, it lies outside the set of evaluated systems in Table 6, so the judge cannot favor a model in its own lineage; in particular the leading frontier families (Anthropic, OpenAI, Google, DeepSeek, Zhipu, Alibaba) are all distinct from the judge family. Third, on the bounded yes/no judgment task that the criterion list defines (a short rubric criterion, a public scenario, and a candidate response), a 120B-class instruction-tuned model is competitive with frontier judges and stable enough that re-prompted verdicts agree at the rates required for the trial standard deviations reported in Table 6. The judge prompt template is held identical across all 1000 items and all 11 evaluated systems; only the criterion text, the item context, and the model response vary across calls. Each verdict is stored alongside its rationale string so that a post-hoc audit can re-examine any criterion that drives a contested cell-level result. Judge failures (timeouts, refusals, empty assessments) are captured as a row-level `judge_error` marker and excluded from the pass ratio denominator for that row, so a judge outage on a single criterion never registers as a model failure.

E.4 Trial Protocol

Each evaluated model is run for 10 independent trials over the benchmark with decoding settings recommended by the model's provider. For each trial, the overall score $P(f)$ and the register-fit score Λ are recorded; the leaderboard in Appendix F reports the mean and standard deviation of $P(f)$ and Λ across the 10 trials. Per-engine and per-cell scores are reported as point estimates because the trial-level breakdown was retained for the overall and Λ channels and was aggregated within $C_{E,L}$. All trials share the same item set, the same judge model, and the same rubric; only the evaluated system varies across runs.

What 10 trials buy and what they do not buy is worth making explicit. Empirically, the per-trial standard deviation of $P(f)$ across the 11 evaluated systems lies in $[0.23, 0.42]$ percentage points, and the standard deviation of Λ lies in $[0.30, 0.43]$ percentage points (Table 6). Both ranges are far below the leaderboard span (the overall $P(f)$ range is 5.42 pp from GLM-5 at 61.72 to Claude Opus 4.6 at 65.63, and the Λ range is 5.44 pp from GLM-5 at 63.05 to ChatGPT 5.5 at 68.49), so a single trial's overall score reliably localizes a model to a coarse tier. However, the gap between adjacent models in the middle of the table is comparable to one trial standard deviation: the smallest adjacent gap on $P(f)$ is 0.105 pp (Gemini 3.1 Flash-Lite vs Gemini 3 Flash) and the next is 0.142 pp (Claude Opus 4.7 vs ChatGPT 5.5), both within one σ_P of any trial. This is the structural reason Figure 8 shows leadership swaps across levels: when the ground-truth gap is below trial noise, the realized rank is unstable across trials by construction. The trial budget tightens the mean overall and Λ to the $\sim 0.07{-}0.13$ pp range (the standard error of the mean over 10 trials), which is what the leaderboard reports. It does *decline* to buy per-cell error bars: with 50 items

per cell and trial-level retention only at the overall and Λ aggregates, individual $C_{E,L}$ point estimates carry the $\sim 5{-}7$ pp Bernoulli envelope from Appendix D and are therefore reported without trial-level dispersion.

E.5 Definition of Λ (Register-Fit Score)

The footing channel λ is materialized in the scoring pipeline as an item-local register-fit judge criterion. When an item carries an authored footing constraint, the rubric injects a judge criterion whose prefix is *Register fit (Lambda output channel)* for English items and *Register fit (*Λ 语体匹配*)* for Chinese items; we refer to this criterion as the Λ criterion. The judge returns a binary verdict on this criterion that is recorded as the row's *Lambda pass*, with value 1 when the response satisfies the required footing for the scenario and 0 otherwise. An item is *Lambda eligible* if and only if the rubric contains a Λ criterion.

The model-level register-fit score is

$$\Lambda(f) = 100 \cdot \frac{1}{|D^{\Lambda}|} \sum_{i \in D^{\Lambda}} \lambda_i\big(f(o_i)\big),$$

where D^{Λ} is the set of Λ-eligible items and $\lambda_i(\cdot) \in \{0, 1\}$ is the judge's pass verdict. Λ therefore measures the percentage of footing-checked items on which the model speaks from an admissible participant position, evaluated by the same judge that scores the rest of the rubric. Because the judge prompt is restricted to the public context plus the model response, Λ measures footing observable in the response itself instead of footing inferable only from privileged information. The example in Figure 2 of the main text shows two responses that score similarly on S_i but diverge on Λ because they choose different participant positions for the same scenario.

Λ-eligibility is rubric-driven and item-local. An item carries a Λ criterion when its authored rubric specifies an admissible participant position (peer expert, in-character role, named third party) whose register the response must respect. Items whose rubric only constrains content (a forbidden disclosure, a required diagnosis, a goal label) carry no Λ criterion and are excluded from D^{Λ}; this is why Λ is reported as a separate channel from $P(f)$ instead of being folded into it. The English/Chinese prefix split is a technical detail of the prompt template: the criterion text injected into the judge prompt begins with *Register fit (Lambda output channel)* for English items and *Register fit (*Λ 语体匹配*)* for Chinese items so that the judge's chain of attention is anchored in the response language, but the binary verdict semantics and the aggregation are identical across languages. When the judge call fails on a Λ-eligible row (timeout, refusal, empty assessment), the row is marked with a `judge_error:empty_assessment` flag and the row's λ_i contribution is set to 0 while the global rate of these flags is logged for diagnostic review; Appendix G.7 reports the resulting Λ-versus-$P(f)$ scatter at the trial-mean level.

F Per-Model Numerical Results

This appendix reports the numerical scores summarized in Figure 3. Table 6 gives, for each of the 11 evaluated systems, the overall PRAXIS score $P(f)$, the four engine subscores S_i^E (averaged within engine across levels and items), and the register-fit score Λ. The Overall and Λ columns are reported as mean $\pm$ standard deviation across 10 independent trials per model; the engine columns are point estimates because per-engine variance is not retained at the same granularity as the overall and Λ trial outputs. All values lie on the 0–100 PRAXIS scale.

Tables 7–10 then report the per-cell scores $C_{E,L}(f)$ that underlie the heatmap in Figure 3a, broken out one engine per table. Models are listed in overall ranking order so that the level-by-level rank dynamics — including the Mental State and Normative collapses that drive overall position — are visible without a separate rank table.

The top five systems in Table 6 occupy a narrow band of roughly one PRAXIS point ($65.64 \rightarrow 64.60$), and the column that differentiates them is the one each model concedes. Opus 4.6 is the overall leader by maximizing Λ (67.95) and Repair (64.67); GPT-5.4 posts the highest IA in the table (87.88) yet records the

Model	Overall	IA	MST	NI	REP	Λ
Claude Opus 4.6	65.64 ± 0.36	86.44	63.05	48.37	64.67	67.95 ± 0.44
GPT-5.4	65.18 ± 0.38	87.88	62.82	45.35	64.67	64.87 ± 0.40
Claude Opus 4.7	65.00 ± 0.27	83.95	63.56	47.58	64.93	67.12 ± 0.46
ChatGPT 5.5	64.86 ± 0.24	86.17	64.02	44.78	64.45	68.49 ± 0.31
Claude Sonnet 4.6	64.60 ± 0.37	84.97	63.07	46.26	64.11	65.21 ± 0.40
DeepSeek V3.2	63.67 ± 0.44	83.75	61.94	45.79	63.18	64.12 ± 0.35
Gemini 3.1 Flash-Lite	62.86 ± 0.35	83.47	61.35	44.05	62.59	63.65 ± 0.31
Gemini 3 Flash	62.76 ± 0.33	87.03	55.56	45.35	63.09	65.78 ± 0.36
Qwen3 80B Thinking	62.30 ± 0.34	80.34	62.87	43.82	62.19	63.28 ± 0.42
GLM-5	61.72 ± 0.41	85.79	52.73	45.85	62.52	63.05 ± 0.41
Gemini 3.1 Pro	60.21 ± 0.34	86.99	47.12	43.79	62.95	66.34 ± 0.35

Table 6: Per-model PRAXIS leaderboard. Overall and Λ are mean $\pm$ std over 10 trials; engine columns are point estimates. IA = Intent Attunement, MST = Mental State Tracking, NI = Normative Interpretation, REP = Repair. Higher is better.

Model	L1	L2	L3	L4	L5
Claude Opus 4.6	85.00	89.26	86.29	90.56	81.10
GPT-5.4	87.09	94.12	87.56	89.68	80.97
Claude Opus 4.7	81.92	91.16	81.70	86.52	78.47
ChatGPT 5.5	87.78	88.42	85.02	87.78	81.89
Claude Sonnet 4.6	86.14	87.36	84.08	86.76	80.49
DeepSeek V3.2	82.18	85.49	88.66	83.34	79.11
Gemini 3.1 Flash-Lite	87.50	85.20	85.98	85.72	72.93
Gemini 3 Flash	85.34	90.94	87.56	92.22	79.11
Qwen3 80B Thinking	80.59	84.14	83.76	83.82	69.38
GLM-5	88.25	90.94	86.76	87.62	75.33
Gemini 3.1 Pro	87.40	90.94	88.04	89.04	79.54

Table 7: Intent Attunement per-cell scores $C_{IA,L}(f)$. Models in overall ranking order. Levels run L1 (explicit fit) to L5 (open-textured ambiguity).

weakest Λ among the top five (64.87); ChatGPT 5.5 owns the highest Λ overall (68.49) but lands fourth in $P(f)$ because its Normative subscore (44.78) lags Opus 4.6 by 3.59 points. At the top of the table, overall rank is set by which weak engine a model concedes, not by which strong engine it maximizes.

Trial-level σ_P stays inside a 0.24–0.44 pp band for every system, which is tighter than the gap between adjacent rows in the upper half of the table. Opus 4.7 and ChatGPT 5.5 differ by 0.14 pp on $P(f)$ with σ of 0.27 and 0.24, so their one-sigma intervals overlap; the same overlap holds for GPT-5.4 versus Opus 4.7 (0.18 pp at $\sigma \approx 0.3$) and for Sonnet 4.6 versus the Opus 4.7 / ChatGPT 5.5 pair. The ranking among the top five is stable in expectation but is within trial noise for any single run. Λ produces a visibly different ordering from $P(f)$: the top-Λ system is ChatGPT 5.5 while the top-P system is Opus 4.6, and Gemini 3.1 Pro climbs from last in P to fourth in Λ. Appendix G.7 plots this rank divergence directly. Across engines the spread is asymmetric: MST varies by 16.9 pp from Gemini 3.1 Pro (47.12) to ChatGPT 5.5 (64.02), while NI compresses every system into a 4.58 pp window (43.79–48.37). Normative items act as a structural ceiling that scales with the benchmark and barely with the model.

The L5 column in Table 7 is the lowest IA cell for every model in the table, with absolute scores ranging from 69.38 (Qwen3) to 81.89 (ChatGPT 5.5). The deepest within-row drops are Qwen3 ($80.59 \rightarrow 69.38$, -11.21 pp) and DeepSeek V3.2 (88.66 at L3 $\rightarrow$ 79.11 at L5, -9.55 pp). Open-textured intent items elicit the largest single-cell loss visible anywhere in the cell tables and disproportionately penalize the smaller-tier systems, which still cluster with the frontier on explicit-fit items but separate when no licensed reading is enumerated.

A second pattern cuts against the naive "more burden hurts more" reading. For most systems L2 scores higher than L1: Opus 4.6 ($85.00 \rightarrow 89.26$), ChatGPT 5.5 ($87.78 \rightarrow 88.42$), Opus 4.7 ($81.92 \rightarrow 91.16$), GPT-5.4 ($87.09 \rightarrow 94.12$), and GLM-5 ($88.25 \rightarrow 90.94$) all rise from L1 to L2. L1 (explicit-fit) catches a

Model	L1	L2	L3	L4	L5
Claude Opus 4.6	56.96	64.37	64.53	65.21	64.18
GPT-5.4	52.95	63.68	63.91	66.55	67.05
Claude Opus 4.7	55.02	65.32	65.32	66.40	65.71
ChatGPT 5.5	58.19	64.36	64.03	66.58	66.96
Claude Sonnet 4.6	56.88	64.67	64.74	64.15	64.90
DeepSeek V3.2	50.59	64.48	64.61	64.21	65.83
Gemini 3.1 Flash-Lite	51.55	62.60	63.96	63.35	65.28
Gemini 3 Flash	48.05	55.23	60.10	55.78	58.62
Qwen3 80B Thinking	57.03	63.37	64.96	64.02	64.97
GLM-5	46.34	56.95	58.25	52.64	49.50
Gemini 3.1 Pro	37.69	48.13	49.45	51.87	48.45

Table 8: Mental State Tracking per-cell scores $C_{MST,L}(f)$. Models in overall ranking order. The two largest single-cell drops in the table (GLM-5 and Gemini 3.1 Pro at L1) correspond to the multi-agent visibility failure described in Appendix C.3.

Model	L1	L2	L3	L4	L5
Claude Opus 4.6	45.32	41.64	50.68	50.95	53.27
GPT-5.4	46.77	38.46	45.68	47.22	48.66
Claude Opus 4.7	46.15	43.50	51.06	47.88	49.29
ChatGPT 5.5	45.65	39.45	43.78	46.82	48.21
Claude Sonnet 4.6	43.12	41.05	49.38	46.31	51.41
DeepSeek V3.2	39.96	42.52	50.53	49.34	46.59
Gemini 3.1 Flash-Lite	41.43	40.29	46.13	45.82	46.59
Gemini 3 Flash	44.95	39.50	49.57	46.25	46.49
Qwen3 80B Thinking	41.60	38.96	49.23	46.00	43.32
GLM-5	46.26	39.55	48.44	48.56	46.41
Gemini 3.1 Pro	41.87	38.68	46.32	46.36	45.69

Table 9: Normative Interpretation per-cell scores $C_{NI,L}(f)$. Models in overall ranking order. Every model in the table sits below its overall score in this engine, consistent with normative interpretation as the dominant ceiling on PRAXIS.

class of failures in which models over-elaborate when no implicit reading is required; L2 (light implicit) gives the same models a target their generation distribution is shaped for. GLM-5 is a clean instance: its L1 IA of 88.25 is third in the column, yet its overall rank is depressed because the Mental State row collapses (see Table 8), confirming that this system's overall position is determined by MST rather than by IA.

The L1 column in Table 8 spreads 20.50 pp across systems, from Gemini 3.1 Pro (37.69) to ChatGPT 5.5 (58.19). This is the widest cross-model spread in any cell in the appendix and corresponds to multi-agent visibility binding under an explicit instruction frame. For most frontier systems the shape of the row is a step from L1 into a narrow L2–L5 plateau: Opus 4.6 jumps $+7.41$ pp from L1 to L2 ($56.96 \rightarrow 64.37$) and then varies by 0.84 pp across L2–L5; Opus 4.7 jumps $+10.30$ pp and then varies by 1.08 pp. Once the visibility frame is internalized, additional implicit burden adds almost nothing to MST cost.

The bottom of the column behaves differently. Gemini 3.1 Pro never crosses 52 on any MST level; its row sits 12–16 pp below the median at every burden. This is a uniform depression rather than a single-cell failure: the model is consistently weak on attributing mental states regardless of how explicit the prompt is, which is why its overall rank collapses even though its IA (86.99) and Λ (66.34) are competitive with the top of the table. GLM-5 shows an intermediate pattern, dropping from 58.25 at L3 to 49.50 at L5; the L5 drop is the second-largest in the column and explains how a system that holds IA L1 at 88.25 lands at 10th overall.

The L2 column in Table 9 is the hardest Normative cell for nearly every system, and the dip is benchmark-wide rather than system-specific. Opus 4.6 drops from L1 to L2 ($45.32 \rightarrow 41.64$, -3.68); GPT-5.4 drops $46.77 \rightarrow 38.46$ (-8.31), the largest single-cell loss visible in this table; ChatGPT 5.5 drops -6.20; Opus

Model	L1	L2	L3	L4	L5
Claude Opus 4.6	65.55	61.85	64.18	66.50	65.24
GPT-5.4	61.75	61.75	63.58	69.66	66.57
Claude Opus 4.7	62.70	62.39	64.91	67.52	67.14
ChatGPT 5.5	62.39	61.33	63.61	68.09	66.81
Claude Sonnet 4.6	63.34	61.11	63.96	66.57	65.55
DeepSeek V3.2	62.28	59.21	60.11	68.40	65.93
Gemini 3.1 Flash-Lite	61.01	60.16	60.61	67.45	63.72
Gemini 3 Flash	62.06	59.85	60.54	67.45	65.55
Qwen3 80B Thinking	60.80	60.80	59.11	65.55	64.67
GLM-5	62.28	60.16	59.30	66.19	64.67
Gemini 3.1 Pro	61.44	61.75	60.06	66.81	64.67

Table 10: Repair per-cell scores $C_{REP,L}(f)$. Models in overall ranking order. L4 is the highest scoring column for nearly every model, reflecting that repair items at moderate burden are well-stereotyped and recoverable; L2 (light implicit miss) is the hardest column, where models often produce apology prose without changing behavior.

4.7 drops -2.65; GLM-5 drops -6.71. L2 NI items pose norm-conflict scenarios in which the explicit textual instruction and the operative norm point in different directions; systems that "follow the letter" are caught consistently across the leaderboard, and the convergence of the dip suggests a shared training-time pattern rather than an idiosyncratic weakness.

L5 inverts the usual burden-hurts ordering. For most systems L5 is the highest Normative cell: Opus 4.6 peaks at 53.27 at L5, Sonnet 4.6 at 51.41, Opus 4.7 at 49.29, ChatGPT 5.5 at 48.21, GPT-5.4 at 48.66. Open-textured norm conflicts, where the operative norm is itself implicit, give models room to articulate a reading that the L2 explicit-but-conflicting frame denies them. Gemini 3.1 Pro is the one row in the column that fails to exhibit either pattern strongly: its NI scores are flat-low across every level (38.68–46.36), and its overall rank loss is driven by NI alongside MST rather than concentrated in any single cell.

The L4 column in Table 10 is the highest Repair cell for every system in the table. The L1-to-L4 jump is between $+4$ and $+8$ pp for most rows: GPT-5.4 climbs $61.75 \rightarrow 69.66$ ($+7.91$), DeepSeek $62.28 \rightarrow 68.40$ ($+6.12$), Gemini 3.1 Flash-Lite $61.01 \rightarrow 67.45$ ($+6.44$). Moderate-burden repairs are the most stereotyped affordance in the benchmark; once an item is recognizable as a face-repair scenario, frontier and mid-tier systems both produce well-shaped responses. L2 is the hardest Repair column for the same set of systems, with most rows scoring 59–62 there. Inspection of the L2 cell against L4 within a single row reveals a 5–8 pp gap that maps onto the difference between a light-implicit miss (where the model needs to identify that a repair is even called for) and a moderate-explicit repair (where the structural cue is overt).

The Repair column compresses cross-model spread more than any other engine. Repair subscores in Table 6 run from 62.19 (Qwen3) to 64.93 (Opus 4.7), a 2.74 pp window across all 11 systems and the tightest of the four engines. At the cell level the spread widens slightly (L1 ranges 60.80–65.55, L4 ranges 65.55–69.66) but never exceeds 5.7 pp. Consequently, Repair contributes least to leaderboard differentiation; rank-order at the top is set by MST and NI, with Repair acting as a near-uniform additive offset. This is consistent with the broader picture in Figure 3b: model profiles in PRAXIS differ on what they fail to track, not on whether they can produce a repair when one is structurally signaled.

G Additional Results Discussion

This appendix collects five auxiliary figures that complement the heatmap in Figure 3a and the radar profiles in Figure 3b. Each figure visualizes a property of the result matrix that prose alone cannot make obvious. No new data is introduced; the figures are derived from the same per-model cell scores reported in Appendix F.

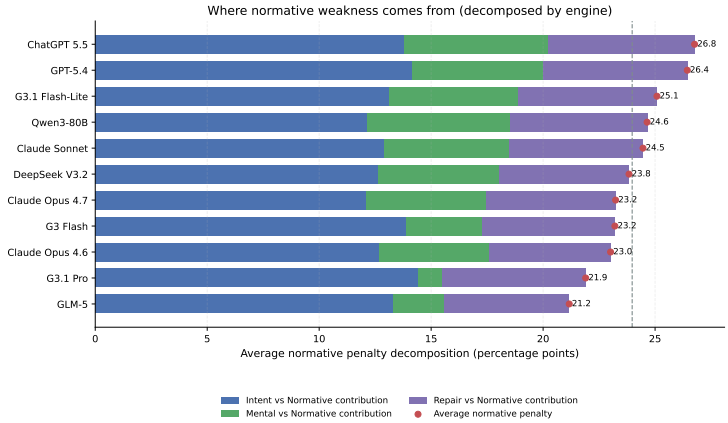


Figure 5: Decomposition of each model's average normative-interpretation penalty into the contribution from each of the other three engines. The stacked bar shows how much of the gap between Normative Interpretation and the other engines is driven by Intent, Mental State, and Repair separately, and the red marker reports the model's average normative penalty in percentage points.

G.1 Where Normative Weakness Comes From

Figure 5 shows how the normative ceiling that depresses every model's overall score is composed. For each system, the bar decomposes the average penalty Intent−Normative, Mental−Normative, and Repair−Normative as separately-stacked contributions to the total normative gap, and the red marker reports the average normative penalty in percentage points. ChatGPT 5.5 and GPT-5.4 carry the largest total penalty — a summed gap of roughly 80 and 79 percentage points across the three non-normative engines — because their Intent and Repair scores are highest (Intent itself reaches 86–88 for these systems), leaving the most headroom relative to the normative floor. GLM-5 and Gemini 3.1 Pro carry the smallest total penalty (roughly 63 and 66 percentage points), but for the opposite reason: their non-normative engines are themselves weaker (GLM-5 Mental State 52.73, Gemini 3.1 Pro Mental State 47.12), so the gap relative to Normative is narrower.

The composition of the bar is as informative as its height. For ChatGPT 5.5, GPT-5.4, Claude Opus 4.6, and Gemini 3 Flash the Intent contribution is the dominant stacked component, indicating that the engine these systems handle best (literal-instruction recovery) is also the one whose surplus they fail to convert into normative reading. In HCO terms, the bar quantifies the cost of leaving the intent layer for the normative layer: an explicit-instruction surface signal is recovered, but the operative norm beneath it is not. Conversely, for Gemini 3.1 Pro the Mental State contribution is unusually small ($\sim$ 3 points on Mental−Normative), which on this figure looks like a narrow penalty but actually reflects that Mental State is itself nearly as weak as Normative for that system rather than that Normative is closer to ceiling.

This is the principal caveat for reading the bar heights. A short bar is consistent with either "normative is recovering toward the other engines" or "the other engines are sinking toward normative," and only the engine-level table in Appendix F disambiguates the two. The figure therefore supports a robust claim about the engine but a softer claim about the model: the normative ceiling is a property of the Normative Interpretation engine across the leaderboard, not of any single system, and the per-model penalty mostly reflects how far the rest of the profile has been pushed up. Closed-source frontier systems (ChatGPT 5.5, GPT-5.4, the Claude Opus family) cluster at the high-penalty end; mid-tier and weaker-MST systems sit at the low-penalty end. This pattern is consistent with the body's claim in Section 5 that normative interpretation is the dominant bottleneck on PRAXIS regardless of base capability, and explains why the radar profiles in Appendix C.4 all collapse on the same axis.

G.2 Overall Strength versus Normative Weakness

Figure 6 tests whether overall ranking is a proxy for normative interpretation. The dashed lines mark the cross-model means (Overall mean $\approx$ 63.5, Normative mean $\approx$ 45.5), which partition the eleven systems

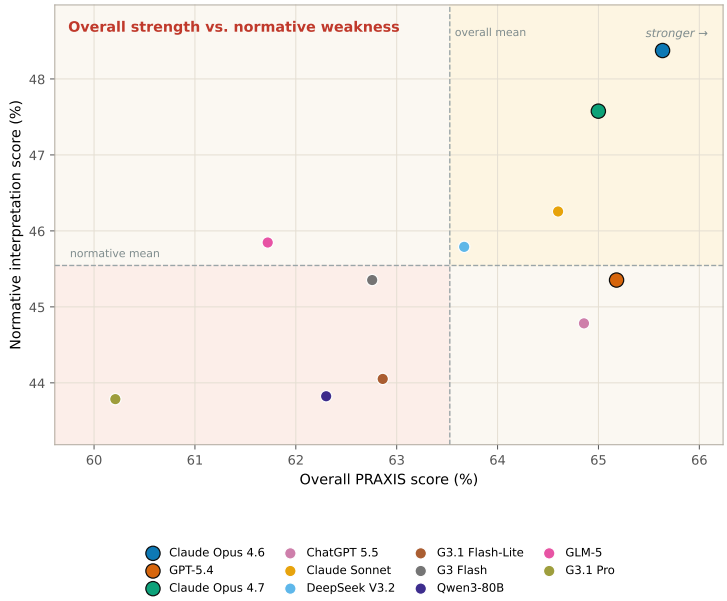


Figure 6: Scatter of each model's overall PRAXIS score against its Normative Interpretation score. Dashed lines mark the overall mean and the Normative mean.

into four quadrants. Claude Opus 4.6 (65.64 / 48.37) and Claude Opus 4.7 (65.00 / 47.58) sit cleanly in the upper-right quadrant — above mean on both axes — with Claude Sonnet 4.6 (64.60 / 46.26) and DeepSeek V3.2 (63.67 / 45.79) joining them at the margin. The most informative outliers are in the high-overall / low-normative quadrant. ChatGPT 5.5 ranks among the strongest systems overall (64.86) but lands at 44.78 on Normative, well below the Normative mean; GPT-5.4 (65.18 / 45.35) is similarly displaced, sitting just under the Normative mean despite being the second-highest overall score in the table. Their position quantifies the decoupling: a model can climb the overall leaderboard through Intent and Repair while ceding ground on the Normative axis.

The mirror-image case appears in the low-overall / above-normative-mean band. GLM-5 has the third-lowest overall score (61.72) yet posts Normative 45.85, fractionally above the Normative mean and ahead of three higher-overall systems on this axis alone. HCO theory predicts this pattern: Normative Interpretation taxes the model's grip on operative norms in conflict with the explicit register, an ability that is partly orthogonal to the raw-capability competences (instruction-following, factuality, fluency) that drive overall ranking through Intent and Repair. The scatter does not, however, support a strong claim that overall ranking is uncorrelated with Normative; the two Claude Opus systems land in the upper-right corner, and the upward trend across the cloud is visible. The right reading is partial decoupling rather than independence.

Two caveats limit the interpretation. First, with only eleven systems the high-overall / low-normative quadrant rests on a handful of points (chiefly ChatGPT 5.5 and GPT-5.4), and the placement of any one point relative to the dashed means could shift with additional systems. Second, the "stronger overall" axis is not a model-size axis; the figure does not separate parameter count or training recipe, and the closed-source frontier systems span both the upper-right and the high-overall / low-normative quadrants. Whatever drives the displacement of ChatGPT 5.5 and GPT-5.4 from the diagonal is not pure capability and is not visible in the scatter alone — the radar profiles in Appendix C.4 and the trajectory plot in Figure 7 are needed to characterise it further.

G.3 Normative Trajectories Across Levels

Figure 7 traces normative score as a function of burden level for the top-3 (Panel A) and worst-3 (Panel B) overall systems, with the remaining eight models plotted in light grey for context. The dominant pattern is non-monotone: every panel dips at L2 before recovering at L3 and onward. The L2 dip

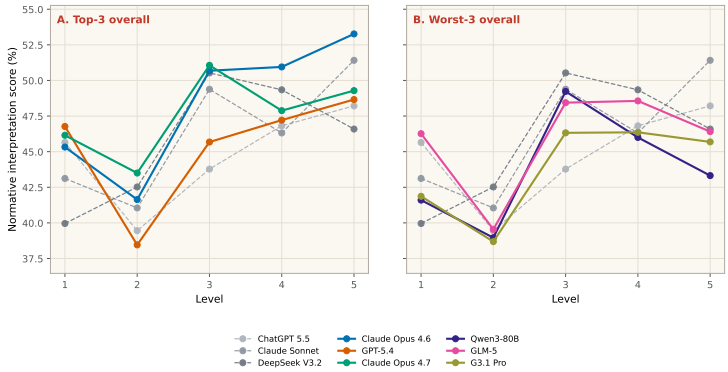


Figure 7: Per-level Normative Interpretation scores for the top-3 and worst-3 overall systems. Panel A shows top-3; Panel B shows worst-3. Background traces show the remaining models in light grey for context.

is the largest single-level move in the figure — GPT-5.4 falls $46.77 \rightarrow 38.46$ (-8.31) and ChatGPT 5.5 falls $45.65 \rightarrow 39.45$ (-6.19), the two steepest L1→L2 drops in the top-3 panel — which on the HCO factorization indexes the failure of explicit-but-conflicting norm items: the textual instruction is recovered (high Intent at L2) but the operative norm is not (low Normative). This is the failure mechanism characterised in Appendix C.3: models satisfy explicit wording but miss the social purpose.

The two panels then diverge sharply past L3. The top-3 traces continue to climb: Claude Opus 4.6 reaches its maximum at L5 (53.27), Claude Opus 4.7 peaks at L3 (51.06) and holds at L5 (49.29), and GPT-5.4 climbs monotonically from L3 onward to a maximum of 48.66. Open-textured norm items at L4 and L5 give these systems room to articulate a normative reading that the L2 explicit-but-conflicting frame denies them, which inverts the burden-hurts-accuracy ordering that holds for Intent. In the worst-3 panel the same recovery does not occur: Qwen3 80B Thinking peaks at L3 (49.23) but declines to 43.32 at L5, a -5.91 pp loss past the peak; GLM-5 plateaus around 46–48 from L3 through L5; Gemini 3.1 Pro plateaus near 46 from L3 onward. The worst-3 thus share a different shape: they can recover from the L2 dip but cannot exploit the open-textured items at L4–L5 the way the top-3 do.

A useful cross-check is whether the rank order at L1 predicts the order at L5. It does not. At L1 the leader on Normative is GPT-5.4 (46.77) with Claude Opus 4.6 fifth (45.32); at L5 Claude Opus 4.6 is first (53.27), Claude Sonnet 4.6 second (51.41), and GPT-5.4 has dropped to fourth (48.66). Claude Sonnet 4.6 in particular climbs from 43.12 at L1 to 51.41 at L5 ($+8.29$ pp), the largest L1→L5 Normative gain in the table and a result that would be invisible in any single-level evaluation. Two caveats apply. First, each Normative$\times$level cell rests on only 50 items (Appendix D), so the Bernoulli envelope on a single cell is roughly $5-7$ pp; small swaps near the top of either panel should not be over-interpreted. Second, the Panel B comparison is between systems with different absolute Normative scores, so the "flattening" visible in the worst-3 is partly that there is less headroom to climb into rather than a qualitatively different burden response.

G.4 Level-by-Level Rank Dynamics

Figure 8 replaces score values with within-level rank to make rank stability visible. In the top-3 panel the lines cross at every burden level: Claude Opus 4.6 leads at L1, L3, L4, and L5, but at L2 Claude Opus 4.7 takes first while Claude Opus 4.6 drops to third within the panel. GPT-5.4 trades second and third position with Claude Opus 4.7 across the remaining levels, never holding the lead but never dropping below third. The crossings are not noise; they track the engine-level patterns visible in Figure 7. Claude Opus 4.7 picks up the lead at L2 because it has the smallest L1→L2 Normative dip in the top-3 (-2.65 pp, compared with -8.31 for GPT-5.4 and -3.69 for Claude Opus 4.6), and Claude Opus 4.6 retakes the lead at L4 and L5 because its Normative trajectory is the one that climbs furthest at high burden (53.27 at L5, the highest Normative cell in the entire benchmark).

The worst-3 panel is more stable but for a different reason than uniform weakness. Gemini 3.1 Pro sits

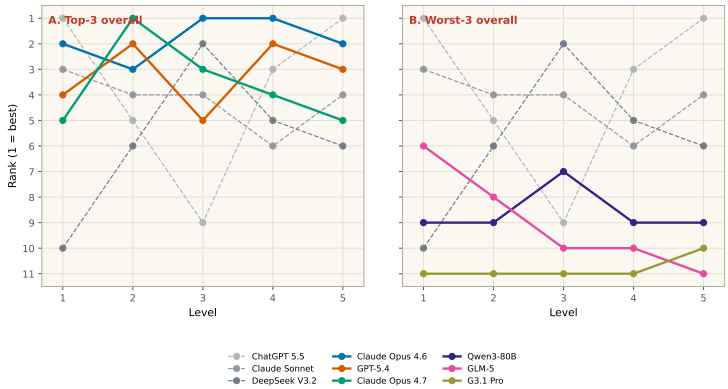


Figure 8: Rank-by-level bump chart for the top-3 and worst-3 overall systems. Rank 1 is best. Crossings within a panel mark levels at which two systems swap positions in the overall ranking.

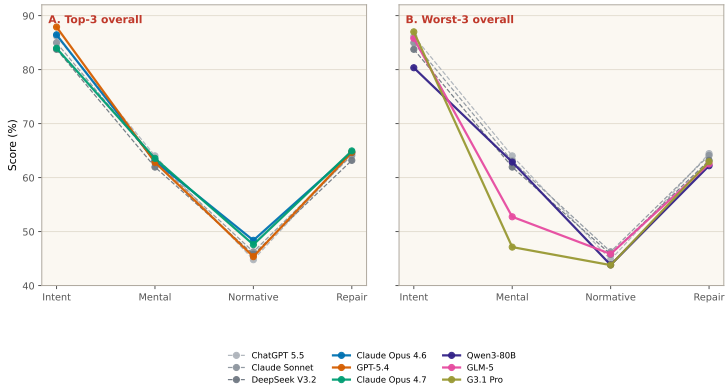


Figure 9: Parallel-coordinates view of engine scores for the top-3 and worst-3 overall systems. Each line is one model; the x-axis runs through the four engines.

last within the panel at every level from L1 through L4 because its Mental State Tracking is uniformly low (37.69 at L1, never exceeding 51.87 across burden), and the small L1→L4 climb in its profile is not enough to overtake the other two. GLM-5 leads the worst-3 at L1 and L2 but is overtaken by Qwen3 80B Thinking at L3, L4, and L5, and GLM-5 drops to last within the panel at L5, falling below Gemini 3.1 Pro. This is the largest mover in the bump chart, and it ties back to the engine-level data: GLM-5 loses -12.91 pp on Intent from L1 to L5 (88.25 $\rightarrow$ 75.34) and gives up another $\sim$ 3 pp on Mental State, while Qwen3 80B Thinking holds Mental State above 64 at L5. GLM-5 is therefore a burden-fragile system, whereas Gemini 3.1 Pro is a burden-flat low-MST system; both end up in the worst-3 but the bump chart makes clear they arrive there by different mechanisms.

Two qualifications are worth flagging. First, within-level rank is sensitive to ties when scores are close: the smallest adjacent overall gap in Table 6 is 0.105 pp, near a single-trial standard deviation, so some of the visible crossings reflect a near-tie rather than a stable rank reversal. Second, the bump chart is restricted to the top-3 and worst-3 by design; the mid-table swaps that the per-cell matrix in Appendix F hints at (DeepSeek V3.2 and Claude Sonnet 4.6 trading position with GPT-5.4 at certain Normative cells) are deliberately not shown. Together the two panels still make the load-bearing point: a single overall scalar collapses meaningful per-level rank reorganization, and any leaderboard derived from it should be read alongside the per-cell matrix in Figure 3a.

G.5 Parallel Engine Profiles

Figure 9 renders the engine profile as a continuous trajectory so that the cross-model shape is directly comparable. Panel A is a comb: the three top systems trace what is essentially a single shape with small vertical offsets, descending from Intent in the mid-80s to Normative below 50 before partially recovering

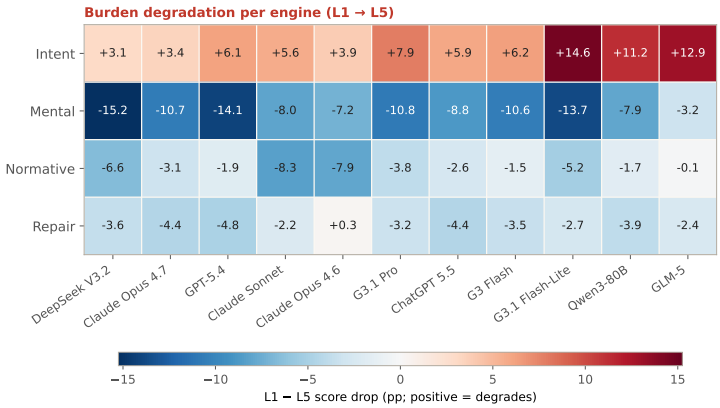


Figure 10: Heatmap of the per-engine L1−L5 score drop (percentage points) for each model. Positive cells (warm) mean the model is more accurate on L1 (explicit) than on L5 (open-textured) items in that engine; negative cells (cool) mean the model improves under burden. Rows are sorted by the row-summed drop, placing the most burden-resilient model on top and the most burden-fragile on the bottom.

at Repair to the low-to-mid 60s. The lines cross only slightly along the way, with Claude Opus 4.6 sitting above ChatGPT 5.5 and GPT-5.4 at Mental State Tracking but ceding ground to GPT-5.4 at Intent. The pattern shows that the top of the leaderboard is differentiated not by which engine a model is strong at, but by how high a shared trajectory sits. Panel B is spaghetti by comparison: the three worst-overall systems separate sharply at Mental State Tracking, with Gemini 3.1 Pro dropping to roughly 47 and GLM-5 to roughly 53 while Qwen3 80B Thinking holds near 63. All three then re-converge at Normative in the mid-40s and again at Repair in the low-60s, so the visible spread in Panel B comes almost entirely from one engine. The Intent→Mental→Normative ordering is not monotone downward in any line: Repair rises back above Normative in every system on the page, reflecting that overt repair affordances are easier than the open-textured normative reasoning required in Normative Interpretation. The figure supports the radar-profile interpretation in Appendix C.4: the top group shares a shape that maps onto the HCO ordering of intent recognition before mental-state tracking before normative inference, whereas the bottom group differentiates by which HCO sub-faculty collapses rather than by an overall level. A caveat is that parallel coordinates collapse per-level variation: a model with a flat engine mean but high level-to-level swing reads as identical here to a stable model, so Figures 7 and 10 should be read alongside it.

G.6 Burden Degradation per Engine: L1 versus L5

Figure 10 compresses the L1→L5 gap for all 4×11 engine×model cells into a single grid, with columns sorted left-to-right from the most burden-resilient model (lowest summed drop) to the most burden-fragile. The leftmost column, DeepSeek V3.2, is the bluest overall: it improves under burden in three of four engines and is the only system with a Mental State drop larger than ten points in the improving direction (-15.2). The rightmost column, GLM-5, is the reddest overall and is the only column whose row sum is positive: GLM-5 degrades by $+12.9$ on Intent and barely improves on Mental State (-3.2), so its end-to-end profile is uniquely fragile. Reading by row exposes a cleaner pattern: the Intent row is red across every column without exception, with the largest L1→L5 drops at Gemini 3.1 Flash-Lite ($+14.6$), GLM-5 ($+12.9$), and Qwen3 80B Thinking ($+11.2$). Intent is therefore the engine that breaks down most universally under burden, which means the high Intent scores in Table 6 are L1-anchored ceilings rather than burden-robust capabilities. The Mental State row is dominantly blue: most systems score *higher* at L5 than at L1 on Mental State, with DeepSeek V3.2, GPT-5.4, and Gemini 3.1 Flash-Lite all improving by more than 13 points. The Normative row is mildly blue, with the deepest improvements at Claude Sonnet 4.6 (-8.3) and Claude Opus 4.6 (-7.9), and the Repair row hovers near zero except for the slightly positive Claude Opus 4.6 cell ($+0.3$). The recurring sign flip is consequential for HCO: when explicit affordances at L1 are removed, models do not lose mental-state and normative competence wholesale; many recover them, presumably because L5's open-textured framings cue a more deliberate

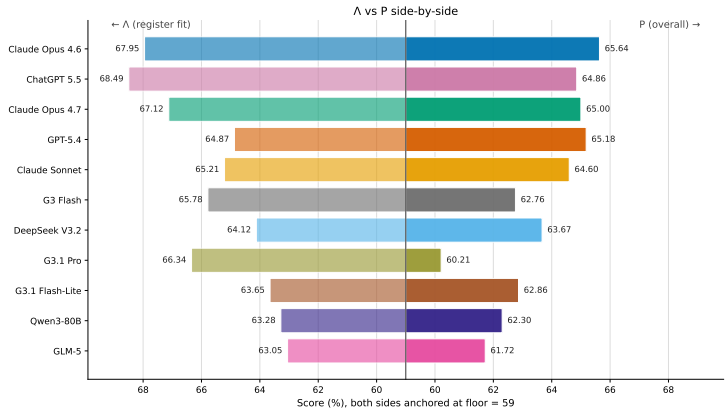


Figure 11: Per-model side-by-side comparison of Λ (register-fit, left) and P (overall, right). Both sides are anchored at a shared floor so absolute strengths are directly comparable across the two metrics. Models are ordered by the mean of Λ and P.

inference rather than a fast pattern-match. The asymmetry between the red Intent row and the blue Mental and Normative rows points to L1 over-reliance on surface intent cues that other engines compensate for once those cues are absent. A caveat is that L1 and L5 differ in item count and content as well as burden, so a per-cell sign should not be over-interpreted as a causal claim about burden alone; the figure should be read together with Figure 7, which traces the full level sequence and reveals the non-monotone L2 dip that this two-point summary hides.

G.7 Λ versus P Side-by-Side

Figure 11 places the register-fit score Λ and the overall score P on opposite sides of a shared floor so that the within-model gap and the cross-model ordering are visible together. For nine of the eleven systems the left (Λ) bar extends further from the floor than the right (P) bar, so register fit systematically exceeds overall rubric satisfaction. ChatGPT 5.5 carries the largest $\Lambda - P$ gap among the top group ($\Lambda \approx 68.5$ vs. $P \approx 64.9$, a 3.6 pp gap), indicating a system that is particularly strong at speaking from an admissible participant position relative to its overall rubric satisfaction. Gemini 3.1 Pro is the most extreme case in either direction: a high Λ (≈ 66.3) sits next to the lowest P on the board (≈ 60.2), a 6.1 pp gap that shows strong register fit does not rescue weak engine performance. At the opposite asymmetry, Qwen3 80B Thinking and Gemini 3.1 Flash-Lite are the only two systems whose Λ and P bars are nearly equal, with Λ within roughly 1 pp of P, because both metrics are simultaneously depressed and there is little headroom for the register premium that the other systems carry. Mapped to HCO, Λ measures the model's adherence to the admissible-participant register (how it speaks), while P aggregates the four engine subscores (what it tracks); the widespread $\Lambda > P$ gap is consistent with current frontier models being trained to inhabit a stable interlocutor persona more reliably than they perform the underlying intent→mental→normative inference chain. The figure also confirms that Λ does not simply re-rank P: Gemini 3.1 Pro is bottom on P but sits in the middle of the Λ ordering, and Claude Opus 4.6 leads on P while ChatGPT 5.5 leads on Λ, so the two metrics order the top of the board differently. A caveat is that the floor anchor at 59 amplifies small absolute differences visually, so the bar lengths are best read alongside the numeric labels rather than by length alone.

G.8 Trial-Level Distribution of (P, Λ)

Figure 12 surfaces what the leaderboard's mean $\pm$ std summary compresses. Each ellipse encodes two trial-to-trial variability dimensions at once: horizontal extent reflects standard deviation in P across the 10 trials, vertical extent reflects standard deviation in Λ. A small ellipse means the model is rubric-stable trial-to-trial; a wide or tall ellipse means a single trial can shift the model noticeably along that metric. All ellipses are small relative to the cross-model spread, with axes on the order of a single point, so the 10-trial standard deviations reported in Table 6 are not noise floors hiding instability. The tightest ellipses

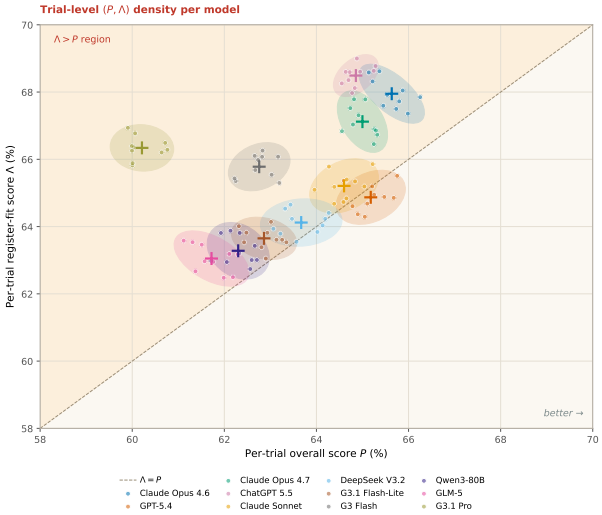


Figure 12: Per-trial (P, Λ) scatter for each model with 2σ covariance ellipse and centroid. Each model contributes 10 trial points. The dashed diagonal marks $\Lambda = P$; positions above the diagonal indicate $\Lambda > P$, below indicate $P > \Lambda$.

sit at GLM-5 and Claude Opus 4.6, both compact and roughly isotropic, indicating that these models reach the same (P, Λ) neighborhood on every trial. The widest ellipses are at GPT-5.4 and Claude Sonnet 4.6, which spread several points along the P axis, suggesting trial-to-trial swings in which rubric items are satisfied more than swings in register fit. Almost every model sits above the $\Lambda = P$ diagonal, confirming the body's observation that register fit is systematically higher than overall rubric satisfaction; the systems closest to or below the diagonal are Qwen3 80B Thinking and Gemini 3.1 Flash-Lite, mirroring the small $\Lambda - P$ gap pattern noted in Figure 11. Gemini 3.1 Pro is the most off-diagonal point, sitting far above the line at $(P, \Lambda) \approx (60, 66)$, an outlier that does not blend into either the top or the bottom cluster. The two top-ranked Anthropic systems (Opus 4.6 and 4.7) together with ChatGPT 5.5 cluster tightly in the upper-right with partly overlapping ellipses, consistent with the body's finding that the top of the leaderboard is separated by less than one point on P and that no single trial would reorder the leader. Mapped to HCO, ellipse tightness is the empirical answer to a reasonable skepticism about the register-fit metric: trial-to-trial stability in Λ comparable to that of P supports treating Λ as a property of the model rather than a sampling artifact of any one rollout. A caveat is that ten trials at a fixed sampling configuration is not a guarantee of stability across paraphrase or context-window variation, so the ellipses bound run-to-run noise rather than full input-distribution variability.