

PCP: A Controlled Benchmark for Pragmatic Intention Recognition in Clinical Communication

Kun Zhang

School of Computer Science, University of Leeds

pjdf0740@leeds.ac.uk

Abstract

Medical LLMs are commonly evaluated on factual question answering, summarization, and explicitly stated clinical information, but less on pragmatic understanding: recovering intended meaning from indirectness, hedging, urgency cues, politeness, and implicature. We present PCP, a 3,533-item benchmark for pragmatic intention recognition in clinical communication, with a controlled synthetic Core split, a close-distractor Hard split, and a small externally grounded Real-Clean split derived from public doctor–patient dialogue datasets. PCP combines answer-choice artifact audits with blind human validation of all Hard and Real-Clean items plus a 300-item Core sample, and targeted expert review of 280 items. Across eleven systems (six API-accessed, five open-weight), our main finding is a recognition-vs-naming dissociation: ten of the eleven select the intended clinical action at 95.5–98.4% aggregate accuracy but name the pragmatic phenomenon licensing it at only 56.6–64.5%, a 32–40 pp gap on the same items. Recognition is also split-dependent: API-accessed models are near ceiling on PCP-Core (98.7–99.8%) but drop on PCP-Hard-469 (83.4–90.6%) and PCP-Real-Clean-64 (79.7–90.6%), so broad controlled probes can mask difficulty exposed by close distractors and externally grounded cases. Because gold actions are often the most medically defensible choice, accuracy is a confound-limited estimate; PCP is an evaluation resource for controlled clinical-pragmatic inference, not for deployment readiness.

1 Introduction

Large language models have made substantial progress on USMLE-style and medical QA (Nori et al., 2023; Singhal et al., 2023, 2025), clinical summarization, and diagnostic interaction (Johri et al., 2025; Tu et al., 2025; Van Veen et al., 2024), but this progress does not establish that models reliably handle pragmatic understanding in clinical

communication. In medical encounters, risk can arise not only from factual error but also from misreading a speaker’s intent, stance, urgency, or face-management strategy. Following work that distinguishes surface-form success from meaning and world-grounded understanding (Bender and Koller, 2020), we focus on a narrower construct: whether a model can recover what the speaker wants the clinical listener to understand and do when that intent is expressed indirectly or socially cautiously.

We use *pragmatic intention recognition* to mean recovery of the clinically relevant intent that a listener should infer from an utterance in context, not its literal propositional content alone. This includes indirect escalation requests, minimized serious concerns, polite urgency signals, and implied dissatisfaction, consistent with clinical communication work on delayed presentation, accounting, and disclosure (Heritage and Robinson, 2006b,a; Hallford et al., 2023). PCP therefore differs from clinical factual QA, summarization, and NLI: those tasks primarily reward recovery of explicitly stated information, whereas PCP probes intended meaning under pragmatic variation.

Misreading such indirect expression carries clinical weight: communication failures are implicated in patient-safety incidents, and patients do not always disclose high-risk concerns in direct terms (Keshtkar et al., 2025; Hallford et al., 2023).

Despite the clinical importance of pragmatic competence, existing medical NLP benchmarks mostly target explicit clinical knowledge, medical QA/MCQA, summarization quality, NLI, dialogue generation, or broad patient interaction (Jin et al., 2021, 2019; Pal et al., 2022; Romanov and Shivade, 2018; Ben Abacha et al., 2023a; Zeng et al., 2020; Saley et al., 2024; Johri et al., 2025). Recent work increasingly evaluates multi-turn consultations, simulated clinical interactions, physician-authored rubrics, open-ended health conversations, and holistic medical task suites (Liu et al., 2025;

Fan et al., 2025; Arora et al., 2025; Bedi et al., 2026). PCP complements this line by isolating a narrower step: pragmatic intention recovery from indirect, hedged, polite, or implicature-laden speaker utterances under controlled variants and close distractors.

We make four contributions. First, we define and operationalize clinical pragmatic intention recognition as a benchmarkable sub-capability distinct from medical QA, summarization, and NLI. Second, we introduce PCP, a three-split benchmark that combines controlled pragmatic variation, close same-domain distractors, and a small public-source-grounded transfer split. Third, across eleven systems (six API-accessed, five open-weight) we report a recognition-vs-naming dissociation: ten of the eleven select the intended clinical action at 95.5–98.4% aggregate accuracy but name the pragmatic phenomenon licensing it at only 56.6–64.5%, a 32–40 pp gap on the same items, so acting on a pragmatic cue does not require naming it. Fourth, we treat answer-choice artifact auditing, blind validation, and expert sign-off as part of benchmark construction rather than post-hoc analysis.

2 Related Work

PCP sits at the intersection of three lines of work that existing benchmarks rarely combine: classical and LLM-era pragmatics, medical interaction evaluation, and benchmark validity / artifact control.

2.1 Pragmatic Competence in LLMs

Pragmatic intention recognition in our setting is grounded in classical work on speech acts, implicature, indirectness, intention, politeness, and common ground. Work on speech acts, conversational implicature, indirect speech acts, and linguistic communication explains how listeners recover intended meaning that is not explicitly stated (Austin, 1962; Grice, 1975; Searle, 1975; Bach and Har-nish, 1979). Politeness, hedging, face management, and common ground are especially relevant for clinical encounters, where deference, mitigation, stigma, or shared context can reshape how urgency or dissatisfaction is expressed (Brown and Levinson, 1987; Lakoff, 1973; Levinson, 1983; Stalnaker, 2002; Clark, 1996). We draw on these traditions to define PCP’s construct boundary, not to build a theory paper.

Recent LLM-focused pragmatics benchmarks provide the closest NLP-side precedent. Hu et

al. compare humans and language models across multiple pragmatic phenomena, PUB benchmarks general-domain implicature, presupposition, reference, and deixis, and grammar-based and indirect-speech-act work confirms that direct counterparts can be much easier than pragmatic variants (Hu et al., 2023; Sravanthi et al., 2024; Zheng et al., 2021; Ruis et al., 2023; Koo et al., 2025). Work on conversational intention recognition and recent pragmatics surveys further motivate evaluating intended meaning rather than surface form alone (Holtgraves, 2008; Ma et al., 2025). PCP differs in its clinical focus, close distractors, answer-choice artifact control, and separation of intention recognition from phenomenon naming.

Clinical interaction research shows that presenting concerns are organized through hedging, delayed presentation, bad-news sequences, other-attribution, and face management (Heritage and Robinson, 2006b; Maynard, 2003; Heritage and Maynard, 2006; Heritage and Robinson, 2006a).

PCP is not intended as another general medical LLM leaderboard. Compared with general pragmatics benchmarks, PCP focuses on clinical contexts where communication failures and indirect disclosure can matter for safety (Keshtkar et al., 2025; Hallford et al., 2023) and asks whether the same underlying intention is recovered across direct and pragmatic realizations. Compared with medical QA, NLI, and summarization benchmarks, PCP does not test whether a model can answer medical questions or compress text; it tests whether the model can infer the intended clinical action from indirect speaker expression. This framing keeps claims bounded: surface-form success alone does not establish robust meaning or clinical dialogue competence (Bender and Koller, 2020).

2.2 Clinical Benchmarks and Interaction Evaluation

Clinical LLM work shows strong performance on knowledge-centered tasks while also emphasizing safety, subjectivity, and deployment-risk gaps (Singhal et al., 2023, 2025; Tu et al., 2025; Agrawal et al., 2025; Bedi et al., 2025). Interaction-oriented evaluations now cover task-oriented medical dialogue (Liu et al., 2025), multi-agent clinical simulators (Fan et al., 2025), patient-interaction rubrics (Johri et al., 2025), open-ended health conversations (Arora et al., 2025), and clinician-validated holistic task taxonomies (Bedi et al., 2026). PCP is complementary to these broader

evaluations: it isolates single-turn pragmatic intention recognition under controlled variants and close distractors rather than full multi-turn diagnostic interaction, general patient-interaction competence, or broad clinical workflow performance. It also differs from physician-intent trajectory annotation in doctor–patient dialogues (Röhr et al., 2025): PCP focuses on speaker-side pragmatic intention recovery, not clinician-side SOAP-stage intent classification.

2.3 Behavioral Testing and Artifact Control

Our benchmark construction primarily follows the behavioral testing logic of **CheckList** (Ribeiro et al., 2020): controlled variants foreground specific linguistic phenomena while holding much of the scenario content fixed. Perturbation sensitivity analysis provides secondary motivation for testing whether local wording changes expose unintended model behavior (Prabhakaran et al., 2019). This allows us to report not just overall accuracy but phenomenon- and variant-associated performance differences.

Because multiple-choice benchmarks are vulnerable to annotation artifacts and answer-choice shortcuts, our methodology is also informed by work on benchmark biases, shortcut learning, and adversarial or dynamic evaluation. Prior research has shown that models can obtain misleadingly strong results by exploiting superficial cues rather than the intended task signal, including syntactic heuristics in NLI and artifacts in clinically specialized settings (Gururangan et al., 2018; McCoy et al., 2019; Herlihy and Rudinger, 2021; Geirhos et al., 2020; Bras et al., 2020; Kiela et al., 2021). Recent medical-evaluation work further cautions that benchmark success can create an illusion of deployment readiness (Agrawal et al., 2025; Bedi et al., 2025). PCP therefore treats artifact audits as a first-class part of benchmark construction rather than a post-hoc diagnostic.

3 Benchmark Design and Construction

PCP is constructed as a three-split benchmark resource spanning controlled synthetic probes, close-distractor challenge items, and a small dialogue-grounded external split. Figure 1 summarizes the full construction, validation, adjudication, and evaluation pipeline.

3.1 Design Principles

The PCP dataset is constructed around three design principles:

1. **Clinical grounding:** Scenarios are motivated by clinical communication patterns documented in medical discourse analysis (Heritage and Maynard, 2006; Heritage and Robinson, 2006a,b) and conversation analysis research (Maynard, 2003).
2. **Controlled perturbation:** Each clinical scenario serves as its own control, with utterance variants ranging from direct/explicit to pragmatically transformed, enabling within-scenario comparison.
3. **Phenomenon isolation:** While real utterances may involve multiple pragmatic phenomena simultaneously, each probe is designed to foreground a primary phenomenon, with secondary phenomena tagged for analysis.

3.2 Pragmatic Phenomena

We target five pragmatic phenomena particularly relevant to clinical communication:

- **Indirect Speech Acts (INDIRECT):** Utterances where the speaker’s communicative goal differs from the literal meaning (e.g., “I was wondering if there are other options” → “I want a different treatment”).
- **Urgency Signaling (URGENCY):** Linguistic devices that convey time-sensitivity or severity, ranging from explicit (“I need help now”) to implicit (“I almost didn’t come in”).
- **Hedging (HEDGING):** Tentative or minimizing language that obscures the true severity of concern (e.g., “It’s probably nothing, but...” for a potentially serious symptom).
- **Politeness Strategies (POLITENESS):** Deferential language that manages social relationships but may mask clinical urgency, especially relevant in cross-cultural encounters.
- **Conversational Implicature (IMPLICATURE):** Meaning conveyed through implication rather than assertion, requiring the listener to bridge inferential gaps.

3.3 Clinical Domains

Probes cover six clinical domains with distinct pragmatic pressures: primary care, emergency (600

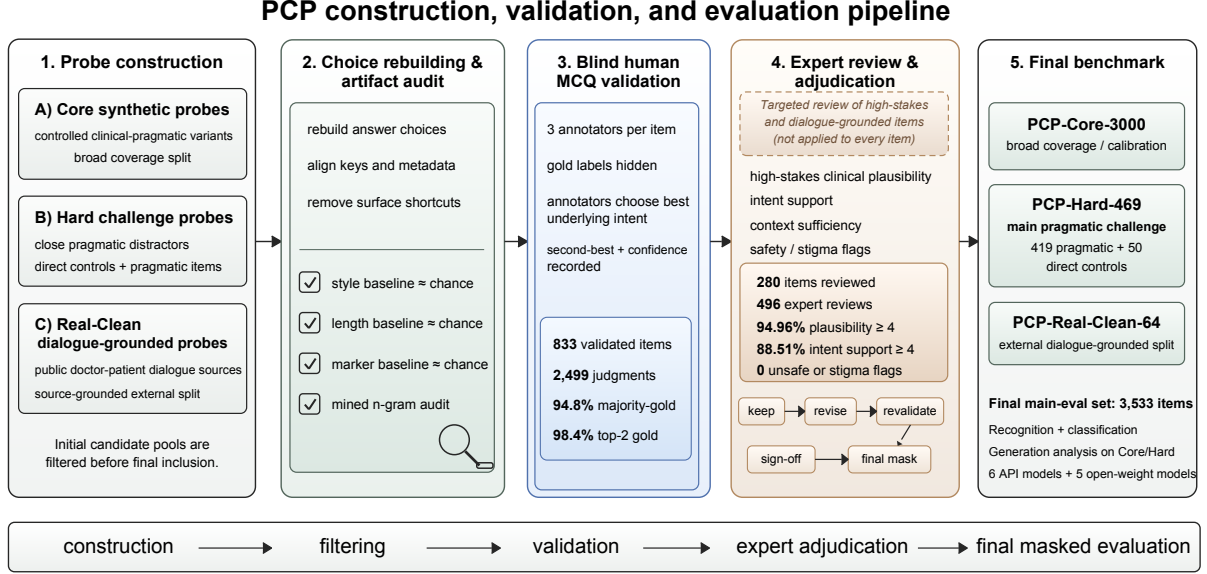


Figure 1: PCP construction, validation, and evaluation pipeline. Core and Hard are controlled clinical-pragmatic splits, while Real-Clean provides a small dialogue-grounded external split. Items pass answer-choice artifact audits, blind human MCQ validation, targeted expert review, and final inclusion masking before model evaluation.

each), mental health, specialist (540 each), pediatric, and geriatric care (360 each), exposing routine negotiation, time-critical escalation, stigma, privacy, referral uncertainty, caregiver mediation, and autonomy concerns.

3.4 Utterance Variants

Each scenario is realized through variants with different pragmatic profiles: direct controls, hedged, indirect, urgent, polite/deferential, casual/offhand, normalized, third-party-attributed, euphemistic, and deflected forms.

Operationalization. We distinguish four levels: *literal content* (the surface proposition), *pragmatic cue* (a hedging / deference / urgency / implicature signal), *underlying communicative intention* (what the speaker wants the clinical listener to recognize), and *clinical action abstraction* (the answer-choice-level action, e.g. evaluation, escalation, treatment adjustment). PCP scores recognition of the intended action, not diagnosis, generation, or deployment readiness.

3.5 Task Format

Each probe is a multiple-choice intention recognition task. The example below is a PCP-Hard item; all four options stay within the same chest-symptom sub-area, so lexical or topic matching is insufficient.

Clinical Context: A 28-year-old patient during a first visit. Patient with chest pressure that started earlier in the day. The speaker worries about being dismissed.

Speaker says: “Maybe I’m overthinking this: there is a little pressure in my chest, probably indigestion.”

What is the speaker’s most likely underlying communicative intention?

- A. The speaker wants the clinician to assess anxiety-related physical symptoms and reassurance needs given minimized wording.
- B. The speaker wants the clinician to evaluate chest pressure as possible cardiac symptoms urgently today given minimized wording.
- C. The speaker wants the clinician to provide anxiety-reduction guidance for chest discomfort and monitor symptom patterns given minimized wording.
- D. The speaker wants the clinician to evaluate respiratory causes of chest discomfort and exertional symptoms given minimized wording.

Option B is labeled correct: minimizing cues (“Maybe I’m overthinking,” “a little,” “probably indigestion”) frame a possible cardiac concern deflationarily, so the supported action is urgent cardiac evaluation rather than anxiety reassurance (A, B) or respiratory workup (D). A, B, D are the item’s tagged same-target-wrong-action, adjacent-target-same-action, and wrong-pragmatic-focus close distractors (Hard item `pcp_hard_000042`, hedged).

Table 1: PCP benchmark overview across final evaluation splits.

Split	Source / realism	Validation / primary use
Core-3000	Controlled probes across six domains and five pragmatic phenomena	Artifact-audited; 300-item blind validation sample; broad calibration
Hard-469	Close same-domain distractors; 419 pragmatic + 50 direct controls	Full blind validation, expert review/sign-off; main challenge split
Real-Clean-64	Public-dialogue-grounded MTS-Dialog/ACI-Bench paraphrases	Blind validation and expert sign-off; external transfer check

3.6 Benchmark Overview

Table 1 summarizes the final benchmark splits.

3.7 PCP-Core-3000 and PCP-Hard-469

The 3,000 probes above form **PCP-Core-3000**, a broad-coverage benchmark constructed as 50 scenario-intent types $\times$ six demographic/context variants $\times$ ten utterance variants. We derive the close-distractor Hard split by selecting scenario/context blocks where the intended action is plausibly confounded with adjacent same-domain actions, then applying expert adjudication, revised-item post-validation, and retained-revision expert sign-off. The final evaluation release is **PCP-Hard-469**, with 419 pragmatic hard items and 50 direct controls. Hardness is based on non-leaky lexical-baseline errors on PCP-Core-3000 and construction-time difficulty diagnostics (lexical baseline failure, scenario/context-block grouping, and human/expert revision rounds); no model from the final eleven-system evaluation table was used at Hard-selection time. Hard-set recognition distractors are rewritten as close same-domain confounds rather than generic alternatives.

3.8 PCP-Real-Clean-64

PCP-Real-Clean-64 is the final externally grounded split: 64 recognition-only items derived from speaker-side turns of public MTS-Dialog (Ben Abacha et al., 2023b,c) and ACI-Bench (Yim et al., 2023a,b) resources licensed under CC BY 4.0 after pragmatic-marker screening, scenario mapping, de-identifying paraphrase, blind validation, expert adjudication, and retained-revision sign-off. Raw source dialogues are not redistributed; released metadata records source/license provenance, keyed HMAC identifiers instead of raw source record IDs, and DeepSeek-V4-Pro as the construction-assistance model, which is not among the eleven evaluated systems. Final labels were reassigned by blind

Table 2: Human validation summary for final included items. *Indiv.* is single-annotator accuracy and *Top-2* the judgment-level rate at which gold is among an annotator’s top two labels, both over all judgments; *Maj.* and *Zero* are item-level percentages; *Conf.* is mean annotator confidence on a five-point scale.

Split	Items	Judg.	Indiv.	Maj.	Top-2	Zero	Conf.
Core	300	900	98.7	99.0	100.0	0.7	4.91
Hard	469	1407	90.6	93.2	97.7	3.2	4.57
Real-Clean	64	192	86.5	87.5	96.4	1.6	4.53
Aggregate	833	2499	93.2	94.8	98.4	2.2	4.69

annotators and expert adjudicators rather than accepted from construction suggestions.

3.9 Human Validation and Expert Review

We validate the final recognition labels with blind human annotation on 833 included items, using three independent judgments per item. Blind validation covers all 469 Hard items, all 64 Real-Clean items, and a 300-item Core sample. Annotators saw only the clinical context, speaker utterance, and answer options, not the gold label or source metadata. Non-majority and zero-gold cases were adjudicated before finalization: Core and Hard items were retained only when adjudication found no clear key, semantic, or medical-support error, while Real-Clean items could be kept for close ambiguity or removed when the grounded utterance did not support the gold label. Relabel recommendations were removed rather than silently changing the public gold label.

Table 2 reports single-annotator and majority agreement with the final gold label, top-2 and zero-gold rates, and annotator confidence. Because validation asks whether three annotators select a hidden final gold label from shuffled answer options rather than inducing an open label set, we use majority-gold, top-2-gold, zero-gold, and confidence as the primary quality statistics rather than treating κ or α as primary agreement measures. A targeted expert review by three reviewers spanning clinical communication, clinical NLP/LLM evaluation, and health communication covered 280 items and 496 reviews against a plausibility, intent-support, actionability, and safety rubric: plausibility was rated ≥ 4 in 94.96% of reviews, intent support in 88.51%, and unsafe/stigma flags were 0%. Remove-level items were excluded; revise-level items were edited, revalidated blind, and retained only after sign-off, which kept seven and excluded 18 (Appendix B).

3.10 Answer-Choice Artifact Control

Multiple-choice pragmatic evaluation is sensitive to answer-choice artifacts: a model can appear to infer intention while exploiting style, length, punctuation, or repeated opening phrases. We therefore rebuild PCP-Core-3000 and the Hard split recognition answer choices with unified speaker-centered frames. Core options use the form “The speaker is expressing the intention that the speaker wants the clinician to ...”, while Hard options use the shorter close-distractor frame “The speaker wants the clinician to ...”. Gold answers and distractors are length-matched, share the same abstraction level, and are balanced across answer positions. We then add non-semantic artifact baselines: style-only, length-only, punctuation-only, starts-with, shortest/longest-word, distractor-marker, and mined-ngram shortcuts. A preliminary Core construction exhibited severe artifacts and is excluded from all main results; Appendix C records why reconstruction is necessary.

Appendix audit material. Appendix A–H reports count/static/artifact audits, validation and expert-review summaries, inclusion-mask definitions, result summaries, model metadata, and prompt templates. Full model-row recomputation requires excluded raw per-call outputs or rerunning the documented endpoints. Raw source dialogue repositories, raw model responses, private annotator assignment keys, participant-facing packets, local runtime files, and API credentials are excluded.

4 Experimental Setup

4.1 Models

We benchmark eleven systems: six API-accessed (GPT-5.5, Claude Opus 4.7, Gemini 3.1 Pro, MiMo-V2.5-Pro, MiniMax-M2.7, GLM-5.1) and five open-weight, run through local OpenAI-compatible vLLM endpoints (MedGemma-27B-it, BioMistral-7B, Gemma-3-27B-it, Mistral-Small-3.2-24B-Instruct, Qwen3-30B-A3B-Instruct). Each reported row was frozen after applying the final inclusion mask to item-level run outputs; summary-only and audit-only records from earlier runs are not carried forward, and raw response logs are excluded from the public release. Following reporting recommendations for experimental NLP results (Dodge et al., 2019), the reproducibility checklist records final data hashes, model identifiers or Hugging Face revisions,

prompt hashes, shuffle seeds, endpoint/API family, decoding settings, parse-failure counts, final-mask version, and aggregation script. Appendix G provides API-access metadata summaries, open-weight reproducibility sidecars, and frozen result summaries.

4.2 Evaluation Protocol

For each probe, the model receives the clinical context, speaker utterance, and four answer choices (one gold, three distractors). To control for answer-position artifacts, we evaluate five deterministic shuffles of the answer choices under two prompt formats, yielding ten recognition runs per probe. We canonicalize raw rows by evaluation id, aggregate repeated runs by majority vote over the selected answer text at the probe level, and count probes with no parseable valid run as incorrect. We parse only standalone answer labels or explicit answer fields, avoiding substring-based extraction errors such as interpreting “The answer is C” as option A. Performance is measured as multiple-choice accuracy.

Algorithm 1: Final-mask-filtered recognition scoring

Input: item set I , five answer-order shuffles, two output formats, final inclusion mask M .

For each item $i \in I$: (1) run all 10 prompt variants; (2) parse only standalone labels or explicit JSON answer fields; (3) map valid parses to canonical answer text, ignoring position; (4) majority-vote valid canonical predictions; (5) mark no-parse items incorrect.

After aggregation: (6) apply M ; (7) compute accuracy on the masked item set.

Output: final-mask-filtered item-level accuracy.

4.3 Baselines

We establish lexical, position, and artifact baselines. **Random** selects uniformly among the four choices (25.0% accuracy), and **Always A** is a position-only baseline that diagnoses whether a model or parser exploits a fixed option position. **Lexical overlap** selects the choice with highest word overlap against either the speaker utterance alone or the context plus utterance, excluding the answer choices themselves to avoid leakage. **Artifact-only** baselines select by answer-choice style, length, terminal punctuation, opening phrase, word-length extremes, marker phrases, or mined n-grams, auditing whether the choices themselves leak the gold label.

4.4 Metrics

Our primary metric is **intention recognition accuracy**. From the same item-level outputs we also compute a **pragmatic gap**, $\text{Accuracy}(\text{direct}) - \text{Accuracy}(\text{pragmatic variants})$; **domain-** and **phenomenon-specific** accuracy slices for exploratory error localization; and a separate aux-

iliary classification task asking whether models can *name* the pragmatic phenomenon. Except where tabled, we interpret these slices conservatively.

5 Results

5.1 Recognition Accuracy Across Controlled and Grounded Splits

Table 3 presents final sign-off-filtered recognition results for all eleven systems on PCP-Core-3000, PCP-Hard-469, the Hard pragmatic/direct slices, PCP-Real-Clean-64, and the 3,533-item aggregate. Non-leaky lexical baselines remain far below a solved benchmark on Hard and Real-Clean: context-plus-utterance overlap reaches 49.0% and 37.5%, respectively. Aggregate artifact-only baselines are near chance; residual split-specific signals are documented in Appendix C Table 12.

API-accessed models are near ceiling on PCP-Core (98.7–99.8%) but drop on PCP-Hard-469 (83.4–90.6%), with loss concentrated in the 419 pragmatic hard items (81.9–90.2%) rather than the 50 direct controls (92.0–96.0%). Open-weight models show the same pattern with wider spread. On PCP-Real-Clean-64, systems range from 79.7–90.6%; with only 64 items and overlapping CIs (Table 3), we treat this split as a range-level transfer check rather than a ranking, with compact statistical notes in Appendix D.

5.2 The Pragmatic Gap

Across all eleven evaluated systems, accuracy on the Hard direct controls exceeds accuracy on the Hard pragmatic items (pooled +7.1 pp; per-system range +2.2 to +14.1 pp, median +5.5 pp; 11/11 positive, a consistent direction rather than a tested effect), though this gap conflates pragmatic difficulty with close-distractor pressure since the two slices are different probes. A within-block contrast on PCP-Core (each scenario–context block’s one direct versus nine pragmatic variants of the *same* content) isolates the pragmatic effect, but Core’s near-ceiling compresses it to a sub-2 pp signal (median +0.74, pooled mean +1.72 pp, block-level bootstrap 95% CI [1.3, 2.2]) that we read as a ceiling-limited lower bound rather than an informative estimate. The more discriminative construct evidence is on Hard pragmatic items: a non-semantic escalation-keyword baseline that always selects the most-escalation option reaches 52.8% there, 30.5% on Core, and 36.3% on Real-Clean pragmatic items (Appendix C), yet every

system exceeds this medical-conservatism shortcut on all three splits (by 24–37 pp on Hard), so accuracy is not merely safe-action selection. A stratified control sharpens this: partitioning the Hard pragmatic items by whether the gold action is the argmax of the same escalation score, pooled accuracy is 89.8% on the 82 items where gold is *not* a most-escalation option, above the 83.7% on the 161 items where it is the unique most-escalation option (10/11 systems reach at least 82% on the former). Where the conservatism shortcut actively selects a wrong option, systems do not follow it. This still bounds only the escalation operationalization of the confound, not medical defensibility in general; separating the two fully requires the minimal-pair control we leave to future work.

5.3 Medical Fine-Tuning Effect (Controlled Pair)

PCP permits one controlled comparison: MedGemma-27B-it and Gemma-3-27B-it share base architecture, parameter count, chat templates, and shuffle seeds. MedGemma is at or above its base on every split (Core +0.9, Hard +1.7, Hard-direct +8.0, Real-Clean +3.1, Aggregate +1.0 pp), but paired McNemar gives $p = 0.374$ on Hard and $p = 0.727$ on Real-Clean, and phenomenon classification is essentially unchanged (56.6% vs. 57.6%). These nulls reflect limited statistical power rather than a demonstrated zero effect: the minimum detectable effect at the observed discordant counts is ≈ 3.8 pp on Hard and ≈ 12.5 pp on the $n=64$ Real-Clean split (Appendix D). Consistent with Jeong et al. (Jeong et al., 2024), PCP is evidence about clinical pragmatic intention recognition rather than medical adaptation overall.

5.4 Domain Analysis and the Recognition-vs-Naming Dissociation

Domain-level analysis is exploratory because some slices are small (Appendix E). Across the full release, the eleven-model mean is highest on Emergency (97.9%, 94.0–99.9) and lowest on Specialist (92.0%, 65.0–98.1); Mental Health spans 73.1–97.6. On the Hard split alone the spread widens further (BioMistral-7B reaches only 45.6% on Hard Pediatric), so per-cell spread is wider than the aggregate suggests; Geriatric and Pediatric require under-claiming because they contribute only 6 and 0 Real-Clean items. Auxiliary phenomenon classification (Appendix F) is substantially harder than

Table 3: Final sign-off-filtered recognition results. Hard and Real-Clean model cells include 95% bootstrap CIs; baseline/artifact rows are point estimates. Best artifact baseline is selected per split (Appendix C).

Model	Type	Core	Hard [95% CI]	Hard Prag.	Hard Direct	Real-Clean [95% CI]	Agg.
Random	Baseline	25.0	25.0	25.0	25.0	25.0	25.0
Always A	Baseline	25.0	25.4	25.8	22.0	26.6	25.1
Lexical overlap (ctx+utt.)	Baseline	71.0	49.0	47.3	64.0	37.5	67.5
Best artifact baseline	Artifact	26.2	34.3	—	—	34.4	26.1
GPT-5.5	Model	99.7	90.6 [88.1, 93.2]	90.2	94.0	90.6 [82.8, 96.9]	98.4
Claude Opus 4.7	Model	99.7	89.8 [87.0, 92.3]	89.0	96.0	87.5 [79.7, 95.3]	98.2
Gemini 3.1 Pro	Model	99.8	87.2 [84.0, 90.0]	86.6	92.0	79.7 [70.3, 89.1]	97.8
MiMo-V2.5-Pro	Model	99.3	87.0 [84.0, 89.8]	85.9	96.0	85.9 [76.6, 93.8]	97.5
MiniMax-M2.7	Model	98.7	83.4 [79.7, 86.6]	81.9	96.0	89.1 [81.2, 96.9]	96.5
GLM-5.1	Model	99.5	89.3 [86.4, 92.1]	88.8	94.0	84.4 [75.0, 92.2]	97.9
MedGemma-27B-it	Model	98.3	85.7 [82.5, 88.7]	84.7	94.0	90.6 [82.8, 96.9]	96.5
BioMistral-7B	Model	84.4	77.2 [73.3, 81.2]	76.6	82.0	82.8 [73.4, 92.2]	83.4
Gemma-3-27B-it	Model	97.4	84.0 [80.8, 87.2]	83.8	86.0	87.5 [78.1, 95.3]	95.5
Mistral-Small-3.2-24B-Instruct	Model	99.1	87.4 [84.4, 90.4]	86.4	96.0	82.8 [71.9, 92.2]	97.3
Qwen3-30B-A3B-Instruct	Model	98.1	89.1 [86.4, 91.7]	88.5	94.0	84.4 [75.0, 92.2]	96.7

recognition: aggregate primary-phenomenon accuracy is 56.6–64.5% for ten systems and 21.4% for BioMistral-7B, while the same ten reach 95.5–98.4% aggregate recognition. Because both tasks are run over the same item set, this 32–40 pp gap is a within-item dissociation rather than a cross-slice contrast: models select the intended clinical action without reliably naming the mechanism that licenses it.

5.5 Generation Analysis

Generation is an auxiliary diagnostic. Under a three-judge ensemble, aggregate overall-success rates span $77.6 \pm 9.3\%$ (MiniMax-M2.7) to $94.5 \pm 5.3\%$ (Claude Opus 4.7); pairwise Cohen’s κ is only 0.20–0.53, and the largest drop-self shift is -7.3 pp on Gemini 3.1 Pro (Appendix F). We rely on recognition for primary claims and treat generation rankings as suggestive.

5.6 Error Analysis

PCP-Hard pragmatic items are built with three explicit close-distractor types per item: `same_target_wrong_action`, `adjacent_target_same_action`, and `wrong_pragmatic_focus` (one of each per item, giving $419 \times 3 = 1,257$ close distractors; type definitions and per-model error summaries in Appendix F). Two released Hard items illustrate the two most consequential failure modes.

Missed urgency cue (`pcp_hard_000222`, possible-sepsis, hedged “it’s probably nothing” after post-procedure fever): gold action is urgent sepsis escalation, but the `wrong_pragmatic_focus` distractor accepts the hedge and treats the fever as routine recovery. **Missed privacy/stigma cue** (`pcp_hard_000212`, postpartum distress,

hedged “probably just hormones”): gold action is postpartum mood screening, but the `same_target_wrong_action` distractor mirrors the speaker’s minimization and converts a screening signal into routine reassurance.

Beyond these examples, per-model error distributions diverge by distractor type rather than collapsing to one generic failure pattern (Appendix F).

6 Discussion

PCP operationalizes one slice of language understanding: recovering a speaker’s intended clinical action when literal form under-specifies it. The clearest signal is the within-item dissociation: selecting the intended action is far more reliable than naming the mechanism licensing it, consistent with cue-response association rather than explicit pragmatic reasoning. PCP provides benchmark-level construct evidence for first-pass intent recovery, not for multi-turn repair, longitudinal care, or deployment readiness; because pragmatic cues vary by language, community, and institution, broader claims require interaction-based evaluation.

7 Conclusion

PCP is a controlled benchmark for clinical-pragmatic inference. Its main result is a within-item dissociation: systems select the intended action far more reliably than they name the mechanism licensing it. Broad probes can also overstate robustness when Hard and Real-Clean stay below solved performance, so PCP targets a high-stakes sub-capability, not deployment readiness.

Limitations

Several limitations warrant discussion. The benchmark is predominantly synthetic: 3,469 of the 3,533 released items (98.2%) are constructed

probes and only the 64 Real-Clean items are externally grounded. PCP-Core-3000 and PCP-Hard-469 are therefore controlled probes that may not capture the full complexity of real clinical pragmatics; PCP-Real-Clean-64 is public-dialogue-grounded and paraphrased but small, recognition-only, and potentially exposed through public-source pretraining. The multiple-choice format also constrains the response space and may overestimate model understanding compared to free-response or interactive formats. Accordingly, benchmark success here is evidence about a specific construct, pragmatic intention recognition under restricted multiple-choice conditions, rather than clinical dialogue competence or deployment readiness (Agrawal et al., 2025; Bedi et al., 2025). Answer options and distractor quality can shape difficulty, so option-order controls, close-distractor hard sets, parser audits, and artifact baselines are essential; residual shortcut signals remain in the final audit. Results reflect specific endpoints, snapshots, chat templates, and local vLLM runs at evaluation time. Three further limitations deserve emphasis. First, because gold actions tend to be the clinically cautious option, accuracy is partly confounded with selecting the safest action: a non-semantic escalation-keyword baseline already reaches 52.8% on Hard pragmatic items, and the gold option is the unique most-escalation choice in 38.4% of them. The stratified control in Section 5.2 argues against a pure conservatism heuristic, since pooled accuracy is highest (89.8%) precisely on the items where the shortcut selects a wrong option, but it bounds only this lexical operationalization. We did not run a minimal-pair control in which the same utterance content maps to different gold actions purely via the pragmatic cue with medical defensibility held equal; such minimal pairs are the cleanest way to isolate pragmatic inference from medical conservatism and are important future work. Second, PCP-Real-Clean’s source corpora (MTS-Dialog, ACI-Bench) are predominantly outpatient/primary-care, so its grounded items skew toward Primary care (38/64) and under-represent high-acuity domains (Emergency and Mental Health at 5 each, Pediatric at 0); the highest-stakes scenarios are therefore evidenced through the controlled Core and Hard splits rather than external grounding, and our high-stakes framing should be read as controlled-probe evidence, not externally validated coverage. Relatedly, of the five target phenomena, IMPLICATURE is

realized as a *primary* phenomenon only in Real-Clean (33/64); the controlled Core and Hard splits foreground Hedging, Indirect, Urgency, and Politeness (with implicature appearing as a secondary tag), so claims about implicature recognition rest mainly on the smaller grounded split. Third, headroom on these splits is narrow: against blind human validation (Table 2), the best model score matches single-annotator accuracy on Hard (both 90.6%) and exceeds it on Real-Clean (90.6% vs. 86.5%), while remaining below the stricter 3-vote majority-gold rate on Hard (93.2%) and above it on Real-Clean (87.5%), so top rankings should be read cautiously. Finally, annotation remains subjective, expert review is targeted rather than exhaustive, and English-only clinical communication leaves cross-linguistic, cross-cultural, and community-specific pragmatic variation for future work.

Ethical Considerations

PCP-Core-3000 and PCP-Hard-469 use synthetically constructed clinical scenarios inspired by real communication patterns. PCP-Real-Clean-64 is derived from public-source MTS-Dialog and ACI-Bench dialogues; the released split uses DeepSeek-V4-Pro-assisted paraphrasing with de-identification instructions, human-validated labels, and expert-adjudicated inclusion decisions, and does not redistribute raw source dialogues. Consistent with responsible data-use checklist practices (Rogers et al., 2021), we did not collect private patient records or use identifiable patient data. The PCP benchmark is designed to evaluate model capabilities, not to make clinical recommendations, and performance on this benchmark should not be interpreted as evidence of clinical readiness. Blind validation and expert review were conducted on benchmark items and labels rather than on model outputs used for leaderboard ranking. Annotators and expert reviewers saw de-identified materials, hidden gold labels, and no source-side identifying metadata. Assignment identifiers were used only for batching, quality control, and response merging; reviewer-facing packets, participant-facing packets, identifier mappings, recruitment records, compensation records, and institutional review documentation are not part of the public release. Expert reviewers included backgrounds in clinical communication, clinical NLP/LLM evaluation, and health communication research; remove-level expert findings were excluded from the final mask. We ac-

knowledge that clinical pragmatic evaluation in English may not generalize to other languages and cultural contexts, and we encourage multilingual extensions of this work.

Data and code availability. The PCP benchmark (all 3,533 items across Core, Hard, and Real-Clean), the derived result summaries, and the evaluation and aggregation code are released under CC BY 4.0 at <https://github.com/Beicicccc/pcp-benchmark>, with per-row source attribution to MTS-Dialog and ACI-Bench. Raw source dialogues are not redistributed; Real-Clean provenance is recorded by HMAC so that source records are verifiable without being exposed. Raw per-call model outputs and annotation administration records are likewise excluded, so full model-row recomputation requires rerunning the documented endpoints.

Ethical review and compensation. The annotated materials are synthetically constructed clinical scenarios and de-identified, paraphrased excerpts from publicly released dialogue corpora. No private patient records, identifiable patient data, or clinical care decisions were involved, and annotation targeted benchmark items and labels rather than participants’ own health information. Annotators and expert reviewers were recruited and compensated in line with institutional practice; the corresponding review, recruitment, and compensation documentation is retained by the authors and is available to the workshop organizers on request.

References

- Monica Agrawal, Irene Y. Chen, Freya Gulamali, and Shalmali Joshi. 2025. [The evaluation illusion of large language models in medicine](#). *npj Digital Medicine*, 8(1):600.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [HealthBench: Evaluating large language models towards improved human health](#). *Preprint*, arXiv:2505.08775.
- J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford.
- Kent Bach and Robert M. Harnish. 1979. *Linguistic Communication and Speech Acts*. MIT Press, Cambridge, MA.
- Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi Haredasht, Ivan Lopez, and 64 others. 2026. [Holistic evaluation of large language models for medical tasks with MedHELM](#). *Nature Medicine*, 32(3):943–951.
- Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soilemani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. [Testing and evaluation of health care applications of large language models: A systematic review](#). *JAMA*, 333(4):319–328.
- Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. [Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations](#). In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. [An empirical study of clinical note generation from doctor-patient encounters](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023c. [MTS-Dialog: A dataset of simulated patient-doctor conversations \(source repository\)](#). GitHub repository. Released under CC BY 4.0; companion repository to (Ben Abacha et al., 2023b).
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. [Adversarial filters of dataset biases](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.
- Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. [Show your](#)

- work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. 2025. [AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2:665–673.
- H. P. Grice. 1975. [Logic and conversation](#). In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 41–58. Academic Press, New York.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- David J. Hallford, Danielle J. Rusanov, Ben Winestone, Ryan A. Kaplan, Mark Fuller-Tyszkiewicz, and Glenn A. Melvin. 2023. [Disclosure of suicidal ideation and behaviours: A systematic review and meta-analysis of prevalence](#). *Clinical Psychology Review*, 101:102272.
- John Heritage and Douglas W. Maynard, editors. 2006. *Communication in Medical Care: Interaction Between Primary Care Physicians and Patients*. Cambridge University Press, Cambridge.
- John Heritage and Jeffrey D. Robinson. 2006a. [Accounting for the visit: Giving reasons for seeking medical care](#). In John Heritage and Douglas W. Maynard, editors, *Communication in Medical Care: Interaction Between Primary Care Physicians and Patients*, pages 48–85. Cambridge University Press, Cambridge.
- John Heritage and Jeffrey D. Robinson. 2006b. [The structure of patients’ presenting concerns: Physicians’ opening questions](#). *Health Communication*, 19(2):89–102.
- Christine Herlihy and Rachel Rudinger. 2021. [MedNLI is not immune: Natural language inference artifacts in the clinical domain](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1020–1027, Online. Association for Computational Linguistics.
- Thomas Holtgraves. 2008. [Automatic intention recognition in conversation processing](#). *Journal of Memory and Language*, 58(3):627–645.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. [A fine-grained comparison of pragmatic language understanding in humans and language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. [Medical adaptation of large language and vision-language models: Are we making progress?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. 2025. [An evaluation framework for clinical use of large language models in patient interaction tasks](#). *Nature Medicine*, 31(1):77–86.
- Leila Keshtkar, Amber Bennett-Weston, Ahmad S. Khan, Shaan Mohan, Max Jones, Keith Nockels, Sarah Gunn, Natalie Armstrong, Jennifer Bostock, and Jeremy Howick. 2025. [Impacts of communication type and quality on patient safety incidents: A systematic review](#). *Annals of Internal Medicine*, 178(5):687–700.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021.

- Dynabench: Rethinking benchmarking in NLP.** In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Youngeun Koo, Jiwoo Lee, Dojun Park, Seohyun Park, and Sungeun Lee. 2025. **Evaluating large language models on understanding korean indirect speech acts.** *Preprint*, arXiv:2502.10995.
- George Lakoff. 1973. **Hedges: A study in meaning criteria and the logic of fuzzy concepts.** *Journal of Philosophical Logic*, 2(4):458–508.
- Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge.
- Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025. **Interactive evaluation for medical LLMs via task-oriented dialogue system.** In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE. Association for Computational Linguistics.
- Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. **Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges.** In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.
- Douglas W. Maynard. 2003. *Bad News, Good News: Conversational Order in Everyday Talk and Clinical Settings*. University of Chicago Press, Chicago.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. **Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. **Capabilities of GPT-4 on medical challenge problems.** *Preprint*, arXiv:2303.13375.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. **MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering.** In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. **Perturbation sensitivity analysis to detect unintended model biases.** 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 5740–5745, Hong Kong, China. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. **Beyond accuracy: Behavioral testing of NLP models with CheckList.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. **“just what do you think you’re doing, dave?” a checklist for responsible data use in NLP.** In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tom Röhr, Soumyadeep Roy, Fares Al Mohamad, Jens-Michalis Papaioannou, Wolfgang Nejdl, Felix Gers, and Alexander Löser. 2025. **“where does it hurt?” dataset and study on physician intent trajectories in doctor patient dialogues.** In *ECAI 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 4569–4577, Bologna, Italy. IOS Press.
- Alexey Romanov and Chaitanya Shivade. 2018. **Lessons from natural language inference in the clinical domain.** In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.
- Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. **The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs.** In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.
- Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam. 2024. **MediToD: An English dialogue dataset for medical history taking with comprehensive annotations.** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.
- John R. Searle. 1975. **Indirect speech acts.** In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 59–82. Academic Press, New York.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. **Large language models encode clinical knowledge.** *Nature*, 620(7972):172–180.

- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31(3):943–950.
- Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. [PUB: A pragmatics understanding benchmark for assessing LLMs’ pragmatics capabilities](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.
- Robert Stalnaker. 2002. [Common ground](#). *Linguistics and Philosophy*, 25:701–721.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomašev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*, 30(4):1134–1142.
- Wen-wai Yim, Yajuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023a. [ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation](#). *Scientific Data*, 10(1):586.
- Wen-wai Yim, Yajuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023b. [ACI-Bench corpus](#). figshare dataset. Companion dataset to (Yim et al., 2023a); released under CC BY 4.0.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. [GRICE: A grammar-based dataset for recovering implicature and conversational reasoning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

A Audit Manifest

This appendix records audit details, statistical uncertainty, and per-cell breakdowns that support replication of the main paper. The construct definition, split definitions, primary results, limitations, and ethical considerations are all stated in the main paper. All numbers below come from the frozen artifacts used to generate the reported results or the retained reproducibility workspace.

A.1 Audit Map

The appendix covers final split composition, count and mask checks, source provenance, artifact and parse audits, validation and expert-review summaries, statistical uncertainty, auxiliary diagnostics, model metadata, and compact prompt-template descriptions. It deliberately avoids placing full raw model-response logs or private annotation administration materials in the review PDF.

A.2 Source-Grounded Provenance

PCP-Real-Clean uses public-source-grounded paraphrases from MTS-Dialog (Ben Abacha et al., 2023b,c) and ACI-Bench (Yim et al., 2023a,b). Released rows record an opaque HMAC of the source-record identifier; the private HMAC key and source-record mapping are excluded from the public release and retained locally for audit only. Raw source dialogues are not redistributed. Table 4 summarizes the released provenance record.

Table 4: Source-grounded provenance for PCP-Real-Clean-64.

Source	Items	License	Use in PCP
MTS-Dialog	48	CC BY 4.0	Speaker-side paraphrases only; raw dialogues excluded.
ACI-Bench	16	CC BY 4.0	Speaker-side paraphrases only; raw dialogues excluded.

The construction-assistance model recorded for Real-Clean drafting is DeepSeek-V4-Pro. Final released labels were assigned through blind human validation, expert adjudication, revised-item revalidation, and expert sign-off rather than accepted from construction suggestions (see Appendix B). DeepSeek-V4-Pro is not in the eleven-system evaluation table.

A.3 Items Excluded from the Public Release

Raw source-dialogue repositories, private annotator assignment keys, participant-facing packets, local

runtime files, API credentials, raw closed/API response logs, and the private Real-Clean HMAC mapping are excluded from the public release. Retained workspace paths recorded in metadata are audit references only.

B Construction, Validation, and Expert Sign-Off

B.1 Finalization Counts

Table 5 gives final release counts and task coverage. Core and Hard support both recognition and generation tasks; Real-Clean is recognition-only. Table 6 records the finalization flow preserved in the released audit logs. Hard starts from a 500-item close-distractor parent (450 pragmatic candidates plus 50 direct controls); Real-Clean starts from an 80-item pre-final grounded parent. Table 7 gives the split-level domain breakdown.

Table 5: Final split counts and task coverage.

Split	Items	Recognition	Generation
PCP-Core-3000	3000	3000	3000
PCP-Hard-469	469	469	469
PCP-Real-Clean-64	64	64	0
Total	3533	3533	3469

Table 6: Finalization flow preserved in the current audit artifacts. “Post-val.” denotes revised-item post-validation removal before final sign-off.

Split	Parent	Exp. rm.	Post-val.	Blocked	Final
Core	3000	0	0	0	3000
Hard	500	7	10	14	469
Real-Clean	80	8	4	4	64

Table 7: Split-level domain counts.

Domain	Core	Hard	Real-Clean
Emergency	600	95	5
Geriatric	360	59	6
Mental health	540	88	5
Pediatric	360	57	0
Primary care	600	85	38
Specialist	540	85	10

B.2 Blind Validation

Table 8 expands the blind-validation summary referenced in the main paper. For final included rows, blind validation covers all 469 Hard items, all 64 Real-Clean items, and a 300-item Core sample, with three independent judgments per item. The validation task asks whether annotators select a hidden final gold option from shuffled answer choices,

so we treat majority-gold, top-2-gold, zero-gold, and confidence as primary quality statistics. Table 9 reports secondary choice-level Fleiss’ κ and mean pairwise agreement; these are diagnostic only, because answer letters are item-local labels and the validation target is agreement with a hidden gold option.

Table 8: Final blind-validation summary after expert sign-off. Indiv. is single-annotator accuracy and Top-2 the judgment-level rate at which gold is among an annotator’s top two labels, both over all judgments; Maj. and Zero are item-level percentages; confidence is on a five-point scale.

Split	Items	Judg.	Indiv.	Maj.	Top-2	Zero	Conf.
Core	300	900	98.7	99.0	100.0	0.7	4.91
Hard	469	1407	90.6	93.2	97.7	3.2	4.57
Real-Clean	64	192	86.5	87.5	96.4	1.6	4.53
Aggregate	833	2499	93.2	94.8	98.4	2.2	4.69

Table 9: Secondary choice-level agreement diagnostics after final-mask filtering. Pairwise agreement is the mean agreeing-pair share per item.

Split	Items	Pairwise	Fleiss’ κ
Core	300	98.9	0.985
Hard	469	89.7	0.863
Real-Clean	64	80.2	0.733
Aggregate	833	92.3	0.897

B.3 Expert Review and Revised-Item Sign-Off

Targeted expert review covered 280 items and 496 reviews. At review level, clinical plausibility was rated at least 4/5 in 94.96% of reviews, intent support at least 4/5 in 88.51%, clinical actionability at least 4/5 in 95.16%, and context sufficiency at 98.19%. Unsafe or stigmatizing flags were 0%. At item level, expert aggregation produced 226 keep, 39 revise, and 15 remove recommendations. Of the 39 revise recommendations, 14 were dropped at revised-item post-validation (10 Hard, 4 Real-Clean; cf. Table 6) and the remaining 25 proceeded to expert sign-off. Table 10 gives the review buckets preserved in the assignment summary; Table 11 reports the final sign-off outcomes for those 25 revised items: 7 were retained and 18 were blocked.

Table 10: Expert-review buckets before final removals.

Bucket	Items	Reviews
Dialogue-grounded source items	80	96
Hard high-stakes pragmatic items	160	320
Core high-stakes sample	40	80
Total	280	496

Table 11: Final expert sign-off outcomes for revised items.

Split	Reviewed	Retained	Blocked
Hard	20	6	14
Real-Clean	5	1	4
Total	25	7	18

C Artifact, Parse, and Sensitivity Audits

C.1 Answer-Choice Integrity and Shortcut Baselines

Answer positions are balanced by construction in Core (750/750/750/750 across A/B/C/D) and near-balanced in the smaller final splits (Hard 119/117/116/117; Real-Clean 17/19/11/17). No final included row has a choice-metadata mismatch or correct-index mismatch. Table 12 reports deterministic shortcut baselines stored in the final artifact audit. The Hard split’s largest residual artifact-only signal is longest-word at 34.3%, still below every evaluated model row on Hard. Real-Clean’s length-only value is 34.4%, but the split has only 64 items and is reported as a small source-grounded transfer check rather than a ranking-stable leaderboard.

We additionally report a clinical **escalation-keyword baseline** that scores each answer option by escalation cues (e.g. *urgent*, *evaluate*, *assess*, *screen*, *risk*, *safety*) minus de-escalation cues (e.g. *monitor*, *routine*, *reassurance*, *follow-up*) and selects a most-escalation option, breaking ties at random. On the 419 Hard pragmatic items it reaches 52.8% (vs. 25% chance), with the gold option being the unique most-escalation choice in 38.4% of items and tied for most-escalation in a further 42%. This baseline bounds how much recognition accuracy could be carried by a medical-conservatism shortcut rather than pragmatic inference: every evaluated system exceeds it by 24–37 pp on Hard pragmatic items, but the gap is not unbounded, motivating the minimal-pair control discussed in the Limitations. The same baseline reaches only 30.5% on Core and 36.3% on Real-Clean pragmatic items, so the escalation shortcut is bounded

well below model accuracy on all three splits. Stratifying the Hard pragmatic items by the same score gives a sharper control than the baseline ceiling alone. The gold option is the unique argmax for 161 items, tied at the argmax for 176, and not at the argmax for 82; on that last stratum a conservatism-only policy is necessarily wrong. Pooled accuracy across the eleven systems is 83.7%, 85.5%, and 89.8% on the three strata respectively, so accuracy is highest exactly where the shortcut fails, and 10/11 systems exceed 82% there (BioMistral-7B, the weakest system overall, is the exception at 68.3%). Per-model values are recomputed by `compute_escalation_baseline.py`.

Table 12: Final artifact and shortcut baselines. All numbers are accuracies in percent.

Baseline	Core	Hard	Real-Clean	Aggregate
Style-only	25.0	25.4	26.6	25.1
Punctuation-only	25.0	25.4	26.6	25.1
Starts-with	25.0	25.4	26.6	25.1
Length-only	20.2	25.2	34.4	22.1
Shortest-word	25.4	6.8	12.5	22.7
Longest-word	23.9	34.3	23.4	25.2
Distractor-marker	24.6	25.4	26.6	24.7
Mined n-gram	26.2	29.9	26.6	26.1

C.2 Parse-Failure Audit

Each recognition probe receives ten model runs (five deterministic answer-order shuffles $\times$ two prompt formats). Table 13 reports the nonzero parse-failure cases and the zero-failure group. BioMistral-7B is the only system with non-trivial per-call failure (about 46% of its Core runs fail to follow the JSON answer format under the JSON style), but probe-level majority recovery from at least one parseable shuffle keeps its probe-level no-parse count at 55 across the 3,533-probe main evaluation.

Table 13: Recognition parse audit summary. Probe-level no-parse counts are after majority aggregation.

Model group	Run fail	Valid	No-parse
Nine systems	0	35,330 each	0
MedGemma-27B-it	1,152	34,178	0
BioMistral-7B	15,728	19,602	55

C.3 Zero-Gold Sensitivity

Table 14 reports sensitivity to excluding the 18 final-included validation items where no annotator selected the hidden gold option. Mean and max accuracy deltas across the eleven evaluated systems are all under 2.5 pp on Hard and under 1.5 pp

on Real-Clean, so the main paper’s recognition rankings are robust to dropping zero-gold items.

Table 14: Sensitivity to excluding final included validation items with zero blind gold votes. Accuracy deltas are percentage points across the eleven evaluated model rows; aggregate excludes all 18 zero-gold items from the 3,533 main-evaluation probes.

Scope	Zero items	Val. share	Mean Δ	Max $ \Delta $
Core	2	0.7	+0.04	0.07
Hard	15	3.2	+1.43	2.31
Real-Clean	1	1.6	+0.78	1.44
Aggregate	18	2.2	+0.27	0.41

D Statistical Uncertainty

The retained bootstrap audit reports, for each model/split cell, a percentile bootstrap 95% CI from 1,000 resamples over that cell’s per-probe correctness vector; the paired between-model deltas use the analogous paired percentile bootstrap over shared probes. Table 15 lists the per-cell CIs and the McNemar contrasts. On PCP-Hard-469, representative 95% CIs include GPT-5.5 90.6% [88.1, 93.2], Claude Opus 4.7 89.8% [87.0, 92.3], MedGemma-27B-it 85.7% [82.5, 88.7], Gemma-3-27B-it 84.0% [80.8, 87.2], and BioMistral-7B 77.2% [73.3, 81.2]. On PCP-Real-Clean-64, intervals are necessarily wide: GPT-5.5 and MedGemma-27B-it are both 90.6% [82.8, 96.9], while Gemini 3.1 Pro is 79.7% [70.3, 89.1]. Paired McNemar tests use the exact two-sided binomial test when the discordant count $b + c \leq 25$ and a continuity-corrected χ^2 approximation otherwise; every PCP-Real-Clean-64 pair falls in the exact regime and every PCP-Hard-469 pair uses the χ^2 form. We report the single controlled base/instruct pair (MedGemma-27B-it vs. Gemma-3-27B-it) together with pairs spanning the Hard accuracy-gap range: the controlled pair is not significant on Hard ($p = 0.374$) or Real-Clean ($p = 0.727$), while large Hard contrasts such as GPT-5.5 vs. BioMistral-7B reach $p = 7 \times 10^{-9}$. These p-values are unadjusted; under Bonferroni correction across the 55 pairwise contrasts per split (threshold $\approx 0.05/55 \approx 9 \times 10^{-4}$), 11 of the 26 nominally significant Hard contrasts survive, whereas no PCP-Real-Clean-64 contrast reaches even nominal significance (smallest $p = 0.065$), consistent with treating Real-Clean as a range-level check rather than a ranking. The MedGemma-vs-Gemma-3 nulls are power-limited: with 62 discordant pairs

on Hard, a difference must reach $|b - c| \geq 18$ (a ≈ 3.8 pp accuracy gap) to attain $p < 0.05$, so the observed $+1.7$ pp lies below the minimum detectable effect; on Real-Clean only 8 discordant pairs remain, giving a minimum detectable effect of ≈ 12.5 pp. The paired-difference bootstrap CIs are correspondingly wide: $[-1.7, +5.1]$ pp on Hard and $[-4.7, +12.5]$ pp on Real-Clean. For the within-block PCP-Core contrast (one direct versus nine pragmatic variants per scenario-context block), per-system means span $+0.2$ to $+2.1$ pp across the ten near-ceiling systems, with block-level bootstrap 95% CIs that exclude zero for nine of them, and $+10.0$ pp for the lower-accuracy BioMistral-7B; the pooled mean is $+1.72$ pp (block-level bootstrap 95% CI $[1.3, 2.2]$, 1,000 resamples over the 300 scenario-context blocks).

Table 15: Statistical uncertainty diagnostics. CI rows give bootstrap 95% confidence intervals on recognition accuracy (1,000 paired resamples per cell). McNemar rows give paired two-sided tests; p -values are unadjusted.

Diagnostic	Model / comparison	Split	n	Estimate	Uncertainty / test
<i>Bootstrap 95% CIs on PCP-Hard-469 (11 systems)</i>					
CI	GPT-5.5	Hard	469	90.6%	[88.1, 93.2]
CI	Claude Opus 4.7	Hard	469	89.8%	[87.0, 92.3]
CI	GLM-5.1	Hard	469	89.3%	[86.4, 92.1]
CI	Qwen3-30B-A3B-Instruct	Hard	469	89.1%	[86.4, 91.7]
CI	Mistral-Small-3.2-24B-Instruct	Hard	469	87.4%	[84.4, 90.4]
CI	Gemini 3.1 Pro	Hard	469	87.2%	[84.0, 90.0]
CI	MiMo-V2.5-Pro	Hard	469	87.0%	[84.0, 89.8]
CI	MedGemma-27B-it	Hard	469	85.7%	[82.5, 88.7]
CI	Gemma-3-27B-it	Hard	469	84.0%	[80.8, 87.2]
CI	MiniMax-M2.7	Hard	469	83.4%	[79.7, 86.6]
CI	BioMistral-7B	Hard	469	77.2%	[73.3, 81.2]
<i>Bootstrap 95% CIs on PCP-Real-Clean-64 (11 systems)</i>					
CI	GPT-5.5	Real-Clean	64	90.6%	[82.8, 96.9]
CI	MedGemma-27B-it	Real-Clean	64	90.6%	[82.8, 96.9]
CI	MiniMax-M2.7	Real-Clean	64	89.1%	[81.2, 96.9]
CI	Claude Opus 4.7	Real-Clean	64	87.5%	[79.7, 95.3]
CI	Gemma-3-27B-it	Real-Clean	64	87.5%	[78.1, 95.3]
CI	MiMo-V2.5-Pro	Real-Clean	64	85.9%	[76.6, 93.8]
CI	GLM-5.1	Real-Clean	64	84.4%	[75.0, 92.2]
CI	Qwen3-30B-A3B-Instruct	Real-Clean	64	84.4%	[75.0, 92.2]
CI	BioMistral-7B	Real-Clean	64	82.8%	[73.4, 92.2]
CI	Mistral-Small-3.2-24B-Instruct	Real-Clean	64	82.8%	[71.9, 92.2]
CI	Gemini 3.1 Pro	Real-Clean	64	79.7%	[70.3, 89.1]
<i>Paired McNemar tests (representative pairs)</i>					
McNemar	MedGemma-27B-it vs. Gemma-3-27B-it	Hard	469	+1.7 pp	$p = 0.374$
McNemar	MedGemma-27B-it vs. Gemma-3-27B-it	Real-Clean	64	+3.1 pp	$p = 0.727$
McNemar	GPT-5.5 vs. MiMo-V2.5-Pro	Hard	469	+3.6 pp	$p = 0.028$
McNemar	Claude Opus 4.7 vs. Gemma-3-27B-it	Hard	469	+5.8 pp	$p = 0.001$
McNemar	GPT-5.5 vs. MiniMax-M2.7	Hard	469	+7.2 pp	$p < 10^{-4}$
McNemar	GPT-5.5 vs. BioMistral-7B	Hard	469	+13.4 pp	$p < 10^{-6}$

E Per-Cell Recognition Breakdowns on PCP-Hard-469

The main paper reports aggregate Hard accuracy and the direct-vs-pragmatic split. The underlying per-cell audit decomposes Hard accuracy by clinical domain (Table 16), primary pragmatic phenomenon (Table 17), and utterance variant (Table 18), with all values computed by majority vote across ten recognition runs per probe under the final inclusion mask. Table 19 additionally summarizes per-split domain-level recognition across the full release. Because several cells are small, these analyses are diagnostic rather than ranking-stable.

Domain-level spread is widest in Pediatric and Specialist items: on Hard, BioMistral-7B reaches 45.6% on Pediatric, while Mistral-Small-3.2 reaches 100.0% on Geriatric. Across the eleven evaluated systems and the full release, Emergency is the strongest domain (mean 97.9%), while Specialist is the weakest (mean 92.0%). Phenomenon-level cells show the intended direct-vs-pragmatic gap: direct controls range from 82.0–96.0%, whereas pragmatic cells vary more across Hedging, Indirect, Urgency, and Politeness. Variant-level diagnostics show that the hardest utterance forms are not uniform across models; BioMistral-7B is

especially weak on casual/proxy and third-party variants, while MedGemma-27B-it loses most on casual and deflecting variants. These patterns motivate the main paper’s conservative interpretation of per-cell findings.

Table 16: Per-model recognition accuracy on PCP-Hard-469 by clinical domain (%). Column n counts Hard items in each domain.

Model n	Emerg. 95	Prim. 85	Spec. 85	Mental 88	Pediat. 57	Geriat. 59
GPT-5.5	98.9	89.4	88.2	88.6	78.9	96.6
Claude Opus 4.7	97.9	88.2	84.7	86.4	86.0	94.9
Gemini 3.1 Pro	96.8	85.9	85.9	81.8	87.7	83.1
MiMo-V2.5-Pro	90.5	85.9	78.8	86.4	87.7	94.9
MiniMax-M2.7	85.3	82.4	77.6	81.8	78.9	96.6
GLM-5.1	98.9	88.2	83.5	83.0	86.0	96.6
MedGemma-27B-it	84.2	85.9	75.3	92.0	89.5	89.8
BioMistral-7B	83.2	76.5	87.1	81.8	45.6	78.0
Gemma-3-27B-it	81.1	88.2	76.5	88.6	82.5	88.1
Mistral-Small-3.2	90.5	94.1	83.5	81.8	73.7	100.0
Qwen3-30B-A3B	96.8	81.2	84.7	89.8	84.2	98.3

Table 17: Per-model recognition accuracy on PCP-Hard-469 by primary pragmatic phenomenon (%). Row n gives the number of Hard items for each phenomenon (Direct row covers the 50 direct controls).

Model n	Direct 50	Hedging 182	Indirect 141	Urgency 49	Politeness 47
GPT-5.5	94.0	90.7	88.7	87.8	95.7
Claude Opus 4.7	96.0	88.5	88.7	85.7	95.7
Gemini 3.1 Pro	92.0	87.4	87.2	81.6	87.2
MiMo-V2.5-Pro	96.0	83.0	87.2	85.7	93.6
MiniMax-M2.7	96.0	79.1	83.0	81.6	89.4
GLM-5.1	94.0	88.5	87.9	89.8	91.5
MedGemma-27B-it	94.0	80.8	85.8	85.7	95.7
BioMistral-7B	82.0	73.6	75.9	79.6	87.2
Gemma-3-27B-it	86.0	81.3	87.2	77.6	89.4
Mistral-Small-3.2	96.0	86.8	83.7	85.7	93.6
Qwen3-30B-A3B	94.0	87.4	87.9	85.7	97.9

Table 18: Per-model recognition accuracy on PCP-Hard-469 by utterance variant (%). Row n gives the number of Hard items for each variant. Dir. is the direct control row; the remaining nine variants are pragmatic transformations.

Model	Dir.	Hed.	Ind.	Urg.	Pol.	Csl.	Norm.	3rd.	Euph.	Defl.
n	50	47	47	49	47	47	42	48	46	46
GPT-5.5	94.0	95.7	87.2	87.8	95.7	83.0	90.5	91.7	93.5	87.0
Claude Opus 4.7	96.0	93.6	87.2	85.7	95.7	78.7	90.5	91.7	91.3	87.0
Gemini 3.1 Pro	92.0	91.5	91.5	81.6	87.2	83.0	88.1	87.5	87.0	82.6
MiMo-V2.5-Pro	96.0	83.0	83.0	85.7	93.6	78.7	83.3	95.8	87.0	82.6
MiniMax-M2.7	96.0	78.7	80.9	81.6	89.4	83.0	76.2	89.6	78.3	78.3
GLM-5.1	94.0	91.5	91.5	89.8	91.5	80.9	90.5	87.5	91.3	84.8
MedGemma-27B-it	94.0	85.1	89.4	85.7	95.7	74.5	81.0	91.7	82.6	76.1
BioMistral-7B	82.0	78.7	85.1	79.6	87.2	63.8	76.2	68.8	76.1	73.9
Gemma-3-27B-it	86.0	78.7	89.4	77.6	89.4	87.2	73.8	89.6	84.8	82.6
Mistral-Small-3.2	96.0	89.4	85.1	85.7	93.6	89.4	85.7	83.3	82.6	82.6
Qwen3-30B-A3B	94.0	87.2	91.5	85.7	97.9	89.4	83.3	85.4	89.1	87.0

Table 19: Exploratory domain-level recognition analysis. Counts are final-release items by split; mean and range are aggregate recognition accuracy across the eleven reported model rows, grouped by domain.

Domain	Core n	Hard n	Real n	Mean	Range
Emergency	600	95	5	97.9	94.0–99.9
Geriatric	360	59	6	97.7	85.9–99.8
Mental health	540	88	5	94.6	73.1–97.6
Pediatric	360	57	0	96.6	87.5–98.6
Primary care	600	85	38	97.2	93.6–98.1
Specialist	540	85	10	92.0	65.0–98.1

F Auxiliary Tasks

F.1 Phenomenon Classification

The auxiliary classification task asks the model to predict the primary pragmatic phenomenon label (DIRECT, HEDGING, INDIRECT, URGENCY, POLITENESS, IMPLICATURE) rather than the speaker-action answer. The primary-phenomenon gold distribution is dominated by Hedging and Indirect in Core and Hard; IMPLICATURE primary items occur in the source-grounded Real-Clean split (33 of 64 items), so the classification label set matches the six-option prompt template in Appendix H. Table 20 reports per-model primary and lenient phenomenon-classification accuracy. Ten of the eleven evaluated systems land at 56.6–64.5% aggregate primary-phenomenon accuracy, with a lenient “any annotated phenomenon” rate of 59.9–68.4%. BioMistral-7B falls to 21.4% primary accuracy and 22.8% lenient accuracy. The same ten systems sit at 95.5–98.4% aggregate recognition (Table 3), so naming the pragmatic mechanism is consistently 32–40 pp harder than selecting the intended clinical action under our protocol.

F.2 Distractor-Type Pickup on Hard Pragmatic Errors

Each PCP-Hard pragmatic item has three close distractors with explicit type tags: `same_target_wrong_action`, `adjacent_target_same_action`, and `wrong_pragmatic_focus`. Table 21 reports, per model, how wrong picks on the 419 pragmatic Hard items distribute across the three distractor types. Models differ in which failure mode dominates. BioMistral-7B and MiniMax-M2.7 mostly choose the `same_target_wrong_action` distractor (45–49% of errors), meaning they identify the right clinical target but mirror the speaker’s framing into the wrong action. GLM-5.1 and Qwen3-30B-A3B-Instruct mostly choose `wrong_pragmatic_focus` (45.8–47.9%), meaning their failures concentrate on literal-content readings that miss the hedge or aside. Mistral-Small-3.2 unusually concentrates on `adjacent_target_same_action` (45%). These distributions show that Hard errors do not collapse to a single generic pattern. The per-model wrong-pick totals here are computed by the per-cell pipeline and may differ by a few items (under 1 pp) from the Hard-pragmatic accuracy in Table 3 owing to tie-breaking on a small number

of probes; the type distribution is unaffected.

F.3 Generation Judge Detail

Generation analysis is a secondary diagnostic. We follow a three-judge protocol (Claude Opus 4.7, GPT-5.5, Gemini 3.1 Pro) over Core and Hard generation outputs for the six API-accessed systems; Real-Clean is not a generation split. Table 22 reports per-model success rates and 1–5 quality scores, and Table 23 reports pairwise judge agreement and the drop-self ablation. Aggregate overall-success rates span 77.6% for MiniMax-M2.7 to 94.5% for Claude Opus 4.7. Mean intent-match scores are 4.11–4.62, clinical-action scores are 3.82–4.25, and pragmatic-cue scores are 3.77–4.68. Pairwise Cohen’s κ on binary `overall_success` ranges from 0.20 to 0.53 between judge pairs, and two of three drop-self moves exceed 2 pp, so generation rankings are treated as suggestive rather than primary evidence. We report two distinct κ summaries that should not be conflated: the 0.20–0.53 range above is computed *between judge pairs* (pooled over models), whereas the κ column of Table 22 is a *per-model* mean over judge pairs and can fall slightly below the pooled range (e.g. 0.18 for Claude Opus 4.7).

Table 20: Final auxiliary phenomenon-classification results. Primary accuracy uses `phenomena_primary` (or `DIRECT` for direct controls); Any reports whether the prediction matches any annotated phenomenon.

Model	Core P	Core Any	Hard P	Hard Any	Real P	Real Any	Agg. P	Agg. Any
GPT-5.5	60.4	66.1	58.8	65.5	37.5	46.9	59.8	65.7
Claude Opus 4.7	59.9	66.1	60.1	66.3	59.4	68.8	59.9	66.2
Gemini 3.1 Pro	63.3	68.7	61.2	68.0	46.9	56.2	62.8	68.4
MiMo-V2.5-Pro	61.3	66.5	61.8	67.4	40.6	51.6	61.0	66.4
MiniMax-M2.7	57.9	60.7	53.9	56.5	37.5	50.0	57.0	59.9
GLM-5.1	58.8	63.4	58.2	64.4	46.9	54.7	58.5	63.3
MedGemma-27B-it	57.1	62.1	56.1	61.4	37.5	45.3	56.6	61.7
BioMistral-7B	21.2	22.5	23.5	24.5	15.6	20.3	21.4	22.8
Gemma-3-27B-it	57.8	63.0	58.4	63.8	45.3	54.7	57.6	62.9
Mistral-Small-3.2-24B-Instruct	57.8	60.7	55.0	58.0	45.3	51.6	57.2	60.1
Qwen3-30B-A3B-Instruct	65.1	68.0	65.0	68.0	29.7	35.9	64.5	67.4

Table 21: Distractor-type pickup distribution among PCP-Hard pragmatic errors per model. *Wrong* is the total number of pragmatic-Hard items the model got wrong; the three percentage columns show what fraction of those errors landed on each tagged close-distractor type.

Model	Wrong	Same Tgt	Adj. Tgt	Wrong P.F.
GPT-5.5	40	32.5	30.0	37.5
Claude Opus 4.7	47	21.3	36.2	42.6
Gemini 3.1 Pro	58	34.5	29.3	36.2
MiMo-V2.5-Pro	59	44.1	30.5	25.4
MiniMax-M2.7	76	48.7	25.0	26.3
GLM-5.1	48	16.7	35.4	47.9
MedGemma-27B-it	64	34.4	37.5	28.1
BioMistral-7B	100	45.0	18.0	35.0
Gemma-3-27B-it	69	34.8	34.8	30.4
Mistral-Small-3.2	60	38.3	45.0	16.7
Qwen3-30B-A3B	48	33.3	20.8	45.8

Table 22: LLM-judge evaluation of free-text generation outputs on Core and Hard only. Overall is the mean $\pm$ sample standard deviation of binary success rates across judges when an ensemble is used; κ is the aggregate mean pairwise Cohen’s κ over `overall_success`. Intent, Action, and Cue are aggregate 1–5 means.

Model	Core Overall	Hard Overall	Agg. Overall	κ	Intent	Action	Cue	Unsup.	Unsafe
GPT-5.5	93.8 $\pm$ 5.3	95.1 $\pm$ 3.8	94.0 $\pm$ 5.1	0.20	4.58	4.23	4.55	1.5	0.1
Claude Opus 4.7	94.4 $\pm$ 5.5	95.4 $\pm$ 4.3	94.5 $\pm$ 5.3	0.18	4.62	4.25	4.62	2.1	0.1
Gemini 3.1 Pro	82.6 $\pm$ 14.2	93.8 $\pm$ 6.1	84.1 $\pm$ 12.9	0.27	4.31	3.85	4.29	10.0	0.5
MiMo-V2.5-Pro	88.0 $\pm$ 9.5	89.4 $\pm$ 7.8	88.2 $\pm$ 9.3	0.32	4.39	3.99	4.38	5.8	0.3
MiniMax-M2.7	77.3 $\pm$ 9.0	79.4 $\pm$ 10.7	77.6 $\pm$ 9.3	0.49	4.11	3.82	3.77	10.2	1.9
GLM-5.1	89.1 $\pm$ 10.9	90.8 $\pm$ 9.1	89.3 $\pm$ 10.7	0.23	4.49	4.09	4.68	13.6	0.8

Table 23: Generation-judge agreement and drop-self ablations. Cohen’s κ is computed on binary `overall_success`. Drop-self reports aggregate success after removing a judge that matches the evaluated model family.

Section	Comparison	n	κ / All	Drop-self
Pairwise	Claude Opus 4.7 vs Gemini 3.1 Pro	62442	0.276	–
Pairwise	Claude Opus 4.7 vs GPT-5.5	62442	0.533	–
Pairwise	Gemini 3.1 Pro vs GPT-5.5	62442	0.202	–
Drop-self	Claude Opus 4.7 (drop self)	–	94.5	94.3
Drop-self	Gemini 3.1 Pro (drop self)	–	84.1	76.8
Drop-self	GPT-5.5 (drop self)	–	94.0	96.3

G Evaluation Protocol and Model Metadata

G.1 Recognition Protocol

Each recognition probe is evaluated with five deterministic answer-order shuffles (seeds 101, 203, 307, 409, 503) under two prompt formats (letter-only and JSON-answer), giving ten recognition runs per probe. Predictions are canonicalized by answer text rather than answer position, no-parse probe-level majorities are counted as incorrect, and final accuracy is reported after applying the 3,533-row inclusion mask. The shared prompt-template SHA-256 hash and per-run task hashes are stored in retained run metadata.

G.2 API-Accessed Models

API-accessed identifiers recorded in the frozen run manifest are `gpt-5.5`, `claude-opus-4-7`, `gemini-3.1-pro-preview`, `mimo-v2.5-pro`, `MiniMax-M2.7`, and `z-ai/glm-5.1`. Access times are between 2026-05-13 23:37 and 2026-05-15 01:49 UTC. Recognition max-token settings are 128 for GPT-5.5 and Claude, 512 for Gemini and GLM, 3072 for MiMo, and not recorded for MiniMax. These API identifiers are mutable provider family aliases rather than dated snapshots, so the API rows are reproducible in protocol and pinned only by the recorded access timestamps above; the open-weight rows are version-pinned by the Hugging Face revision hashes below.

G.3 Open-Weight Models

Open-weight systems are served through OpenAI-compatible local vLLM endpoints in bf16 with no quantization. Recognition max tokens are 20 and classification max tokens are 12. Retained Hugging Face revision prefixes are `MedGemma-27B-it 2d3e00ea38b5`, `BioMistral-7B 9a11e1ffa817`, `Gemma-3-27B-it 005ad3404e59`, `Mistral-Small-3.2 95a6d26c4bfb`, and `Qwen3-30B-A3B 0d7cf23991f4`. BioMistral ran under vLLM 0.10.2 on RTX 4090 hardware; the other open-weight systems used vLLM 0.11.2 on RTX PRO 6000 hardware.

H Prompt Templates

Prompt templates use a shared recognition body with clinical context, speaker utterance, the question “What is the speaker’s most likely underlying

communicative intention?”, and four speaker-centered answer choices.

Recognition (letter style). The letter-style prompt asks the model to return only one capital letter: A, B, C, or D.

Recognition (JSON style). Same task body, with the instruction replaced by a request for a JSON object whose `answer` field contains the selected letter.

Auxiliary classification. The classification prompt uses the same context and utterance fields and asks which pragmatic phenomenon best describes the utterance, with options `DIRECT`, `HEDGING`, `INDIRECT`, `URGENCY`, `POLITENESS`, and `IMPLICATURE`.

Generation. Generation prompts ask for the speaker’s likely intended clinical action and the pragmatic cue, using `concise`, `clinical`, and `json` styles.