# Same Context, Different Clinician: Recoverable Pragmatic Signatures of Language Models in Clinical History-Taking

**Sara Mirabi**
School of Information Technology
Deakin University, Australia
s222496341@deakin.edu.au

**Bahadorreza Ofoghi**
School of Information Technology
Deakin University, Australia
b.ofoghi@deakin.edu.au

## Abstract

Large language models are increasingly tested as clinical interlocutors, but most evaluations ask whether their answers are medically reliable, although a consultation is built turn by turn. From 272 simulated Objective Structured Clinical Examination (OSCE) consultations, we select 1,500 points where a clinician spoke next and ask six language models to continue from the same preceding dialogue. Treating the clinician's turn as one reasonable continuation rather than the correct answer, we characterise each turn with model-independent measures and analyse them with consultation-clustered regressions. First, models do not converge on a shared style: pragmatic features alone recover the source of a turn well above chance. Second, the assumption that models pack more questions into each turn holds only for some: GPT and, more modestly, Claude issue more information requests per turn than the clinician, whereas Gemini, Qwen and the medically fine-tuned models issue no more or fewer, a pattern that blind human coding broadly reproduces. Models differ more consistently in how much language surrounds a question and whether they acknowledge the patient. Finally, medical fine-tuning does not guarantee good interaction: MedLlama frequently produces no information request and departs prematurely from history-taking.

## 1 Introduction

Large language models (LLMs) are increasingly proposed as clinical interlocutors and evaluated on medical question answering, diagnostic accuracy, and information-gathering completeness (Singhal et al., 2023; Tu et al., 2025), where current models perform well. A consultation, however, is an interaction: history is built turn by turn as the clinician acknowledges the patient, asks focused follow-ups, paces disclosure, and decides when to move from information gathering to interpretation. A model may therefore be clinically sensible in what it asks while differing in how it organises the exchange, a pragmatic dimension that remains comparatively under-evaluated.

One common assumption concerns bundling. Communication-skills training discourages compound, multi-part questions in favour of one focused question at a time because combining requests can rush the patient and yield a poorer history (Silverman et al., 2016). Language models tend to respond at length, a tendency linked to weaker multi-turn performance (Laban et al., 2026), and may combine several requests in one turn. This motivates the hypothesis that models compress interaction relative to clinicians, folding into one turn what clinicians distribute across several. For chest pain, a clinician may ask about site, radiation, and aggravating factors across successive turns, whereas a model may ask all three at once ("Where exactly is the pain, does it spread anywhere, and is it worse on exertion?"). Both gather useful information but organise the interaction differently.

We test whether such differences are systematic and model-specific using a position-matched design. From simulated Objective Structured Clinical Examination (OSCE) consultations (Fareez et al., 2022), we identify history-taking points where a clinician spoke next and compare the actual turn with model-generated turns from the same preceding dialogue. All sources therefore face the same conversational state, isolating how they proceed rather than the situation they encounter. We treat the clinician's turn as one reasonable continuation, not a gold standard, and examine differences in the shape and pragmatics of each move. We assemble 1,500 position-matched contexts from 272 OSCE consultations and generate the next clinician turn with six language models spanning proprietary and open-weight, general-purpose and medically fine-tuned systems. We ask whether these models

organise turns differently from the clinician and one another, and whether interactional compression separates them. Each turn is characterised using deterministic, model-independent measures of length, information-request density, question bundling, acknowledgement, discourse phase, and alignment with the patient's language, and analysed using consultation-clustered regressions. To our knowledge, this is the first position-matched pragmatic comparison of multiple model families against clinicians in history-taking. We make our code and data publicly available.[1]

This work makes three contributions:

1. **Position-matched evaluation.** A controlled framework comparing models and clinicians at identical dialogue states, treating the clinician as a reference rather than a gold standard and separating pragmatic organisation from the situation.

2. **Recoverable pragmatic signatures.** The evaluated configurations occupy distinct regions of pragmatic space, recoverable from interpretable pragmatic features alone rather than converging on a shared LLM style.

3. **Context-dependent divergence.** Differences can grow as dialogue history accumulates, giving the comparison a multi-turn rather than single-shot character.

## 2 Related Work

Our work draws on three areas: clinical evaluation of language models, which has largely focused on what models say; history-taking as a structured communicative skill; and the pragmatic behaviour and distinctive signatures of language models. We combine them to examine how models organise clinical dialogue when the conversational state is held constant.

**Evaluating clinical language models.** Most evaluation of clinical LLMs targets the correctness of what a model says. The MultiMedQA benchmark and Med-PaLM assess factuality, reasoning and possible harm in medical question answering, with human raters scoring answer quality (Singhal et al., 2023). AMIE, a system optimised for diagnostic dialogue, was compared against primary care physicians in OSCE-style consultations on axes such as history-taking, communication and empathy (Tu et al., 2025), and interactive benchmarks test whether models gather sufficient information from simulated patients (Johri et al., 2025; Schmidgall et al., 2024) or ask questions when information is missing (Li et al., 2024). These lines of work ask whether a system reaches the right information or matches clinicians on rated quality. We ask a descriptive question instead: given an identical dialogue state, how each of several models organises the next turn, relative to the clinician and to one another. Our aim is not to build or rank a diagnostic system but to characterise pragmatic behaviour under matched conditions.

**History-taking and communication skills.** Clinical communication is a taught skill with an explicit structure. The Calgary-Cambridge guides codify the process of the consultation, including incremental information gathering and a preference for focused questions over compound ones (Silverman et al., 2016). This literature is the source of the intuition we test: that packing several requests into a single turn is dispreferred. More broadly, conversation analysis has long treated the organisation of turns, not only their content, as consequential for interaction (Sacks et al., 1974). Our measures operationalise a few of these organisational properties in a reproducible way.

**Pragmatics and signatures of language models.** A growing body of work probes the pragmatic competence of LLMs, for instance their ability to resolve implicature, and finds that behaviour varies with model and training strategy (Ruis et al., 2023; Hu et al., 2023). These studies evaluate the comprehension of pragmatic phenomena; we instead measure the pragmatics a model produces when continuing a dialogue. Separately, work on authorship attribution shows that LLM outputs carry model-specific regularities from which their source can be identified (Kumarage and Liu, 2023; Sun et al., 2025); we test whether such signatures also appear in interactional choices rather than only in wording. Methodologically, our design follows the logic of matched evaluation, which holds a problem constant to expose behaviour that aggregate accuracy hides (Wu et al., 2024); here the dialogue state is held constant and only the source of the next turn varies.

[1] https://github.com/saramirabi/osce-next-turn-ALTA2026

## 3 Methods

### 3.1 Data and position-matched instances

We use $D = 272$ simulated OSCE consultations. A *candidate target* is a clinician turn immediately following a patient turn, yielding $|\mathcal{C}| = 13{,}111$. For candidate $i$, $C_i$ is the preceding dialogue (context), $p_i$ the patient's last turn, $H_i$ the clinician's next utterance (human target), and $\pi_i \in \{\text{EARLY}, \text{MIDDLE}, \text{LATE}\}$ its position. All sources see the same $C_i$, so differences reflect continuations of the same conversational state.

Candidates are eligible if a deterministic, rule-based filter identifies history-taking as their primary function and at least two exchanges precede them, leaving $|\mathcal{E}| = 11{,}573$. This is an eligibility filter, not a validated dialogue-act scheme; it does not use punctuation or wh-words as proxies, preserving elliptical turns such as "Any fevers?". We sample $1{,}500$ contexts from $\mathcal{E}$ using a dialogue-aware, position-stratified scheme. Since $1{,}500/272 \approx 5.51$, each consultation contributes five or six targets: 132 contribute five and 140 six ($132 \times 5 + 140 \times 6 = 1{,}500$). Every consultation contributes two early, two middle and one late target, and the sixth target, where present, is drawn from the late stratum, yielding 544 early, 544 middle and 412 late contexts ($272 + 140$). Sixth-target consultations are selected by fixed random seed rather than dialogue ID to avoid confounding with specialty prefixes. After quality control (Appendix C), the sample was frozen before model evaluation.

### 3.2 Sources and generation

Sources are the human clinician and $M = 6$ instruction-tuned models: four general-purpose (GPT, Gemini, Claude, Qwen) and two medically fine-tuned (MedGemma, MedLlama). Checkpoints, developers, access modes, and decoding settings are in Table 4 (Appendix B). All models receive the same minimal prompt with the setting and $C_i$, requesting one next clinician turn. It withholds $H_i$, specialty, position and labels, and gives no instruction to continue history-taking, ask one question, be brief, or imitate the clinician, so moves into management are model-generated. Decoding is as deterministic as supported (temperature zero; reasoning budgets disabled or minimised; Table 4). We retain raw outputs and a normalised $\nu(\cdot)$ that removes role labels and trailing scaffolding and retains only the first utterance of hallucinated multi-turn exchanges. Linguistic measures use $\nu$; the normalisation flags feed the compliance analysis (Section 4.3), which is reported separately so normalisation does not conceal poor instruction-following.

### 3.3 Pragmatic measures

Each turn is represented by a model-independent feature map $\phi(\nu(y))$. The measures are designed to capture observable interactional properties of the continuation rather than clinical correctness or similarity to an assumed gold response.

**Expansion.** We measure the number of words and sentences in each turn, together with words per information request. These features capture how much interactional material a source produces at the same dialogue point.

**Information-request density.** We use a deterministic information-request counter. Let $\nu(y) = \{s_1, \ldots, s_m\}$ denote the sentences in the normalised turn. A sentence is treated as interrogative when it ends in a question mark, opens with a wh-word or auxiliary, or matches a restricted set of elliptical question forms such as "Any …". Let $\mathbb{1}_Q(s_j)$ indicate that sentence $s_j$ is interrogative, and let $c(s_j)$ denote the number of additional interrogative clauses coordinated onto it, where a coordinated clause is one introduced by a comma, semicolon or coordinator (*and*, *but*), immediately followed by a wh-word or auxiliary, and preceded in the same sentence by a wh-word or auxiliary. The last condition prevents an opening acknowledgement ("OK, do you …") from creating a second request. The request count is

$$r(y) = \sum_{j=1}^{m} \mathbb{1}_Q(s_j)\big(1 + c(s_j)\big).$$

An option list serving a single request ("arm, jaw or back") therefore counts once, whereas separate requests joined by coordination ("…, and is it worse on exertion?") count separately. We also record the binary bundling indicator $\mathbb{1}[r(y) \geq 2]$.

Because information-request count is an outcome, we compare the counter with blind human coding. Of these $1{,}050$ coded turns, a stratified subset of 150 turns was independently double-coded and adjudicated ($\kappa = .759$ for request count, linear weighted $\kappa = .772$; $\kappa = .773$ for discourse function). Human coding reproduces the main clinician-relative pattern of the counter (Appendix E). Human coders counted each item of a listed checklist

("any fever, chills, or cough?") as a separate request, whereas the counter counts one question as one request; the counter is therefore a question-level and conservative measure of request density. Against human counts it has high precision (.89) and lower recall (.67), with under-counting concentrated in GPT's checklist questions (recall .47).

**Discourse phase.** We use a binary lexical indicator of departure from history taking. Specifically,

$$\mathrm{sw}(y) = \mathbb{I}\big[\nu(y) \text{ matches the phase-switch lexicon}\big].$$

Thus, $\mathrm{sw}(y) = 1$ indicates lexical evidence of a transition from history taking into examination, diagnosis, interpretation, or management, whereas $\mathrm{sw}(y) = 0$ indicates that no such cue is present. This is a deliberately coarse operational measure rather than a full multi-class dialogue-act classifier. Against human primary-function coding collapsed to the same binary contrast, it agrees at $\kappa = .751$ (1,050 turns; accuracy $\approx .92$).

**Interpersonal marking.** We record whether a turn begins with an acknowledgement and count hedges using a fixed lexicon. These are surface interactional markers rather than direct measures of empathy, uncertainty, or communication quality.

**Lexical and semantic alignment.** We measure lexical content overlap with the immediately preceding patient turn $p_i$. For descriptive semantic alignment, a single frozen sentence encoder computes $\cos_H(y)$ between the embeddings of $\nu(y)$ and the observed clinician continuation $H_i$, and $\cos_p(y)$ between $\nu(y)$ and $p_i$. Embeddings are L2-normalised, and the encoder is used only for scoring, never for generation or fine-tuning. For the observed clinician continuation, $\cos_H = 1$ by construction and is therefore excluded; $\cos_p$ is reported for all sources. Both embedding-based measures are excluded from the source-recovery classifier (Section 4.5), which uses pragmatic features alone.

**Compliance.** Finally, we record malformed outputs and zero-request turns separately so that pragmatic differences are not conflated with failures to produce a usable clinician continuation. No generative language model participated in human coding, eligibility filtering, adjudication, or feature computation; the only learned component in scoring is the frozen sentence encoder used for the descriptive alignment measures.

# 4 Results

We compare six models with the human clinician (seven sources). Inferential contrasts use consultation-clustered generalised estimating equations with the clinician as reference (Appendix G); comparisons between models are descriptive. Table 1 reports per-source means and clinician contrasts with 95% confidence intervals.

## 4.1 Expansion is common; compression is model-specific

The clearest differences between the sources are in how much language they spend. Qwen is the most expansive, producing turns nearly three times the clinician's length (+23.7 words). GPT, MedGemma and Claude produce turns roughly 1.7 to 1.9 times the clinician's length (+10.3, +11.5 and +8.4 words); Gemini is terser than the clinician (−2.0 words), and MedLlama is slightly longer (+2.1 words). But length conceals what the words are doing. Measured as words per information request, Gemini sits at the clinician's level (median 9.0 against 9.5) and MedGemma slightly above (11.80), whereas Claude and GPT (13.3 and 14.0) and especially Qwen (19.0) spend more words on each request: Qwen's extra length is elaboration around a request, whereas GPT's combines elaboration with additional requests.

The common intuition that language models issue more information requests per turn than clinicians holds only for some models. The clinician produces 1.16 automatically detected requests per turn and asks a single request in most turns (14.5% multi-request). GPT is the clearest exception, with 1.40 requests per turn (RR 1.21, 95% CI [1.17, 1.25]) and the highest multi-request share (37.9%), followed more modestly by Claude (1.26; RR 1.09, [1.06, 1.12]; 21.9%). Gemini, Qwen, MedGemma and MedLlama are below the clinician (Table 1, Figure 3). Blind human coding of 1,050 turns reproduces this clinician-relative pattern (Appendix E): GPT and Claude exceed the clinician, Gemini and MedLlama fall below it, and Qwen and MedGemma do not differ from it. Because human coders counted each item of a listed checklist as a separate request, GPT's excess is larger still under human coding (RR 2.13): it frequently packs several symptoms into a single question ("Have you had any fever, chills, sore throat, or cough?"), a form of bundling that a question-level count understates.

| | Words | | Info-req. | | Multi-q. | Words/ | Ack. | | Switch | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | mean | $\beta$ [95% CI] | mean | RR [95% CI] | % | req. | % | OR [95% CI] | % | $\Delta$pp |
| Human | 12.4 | *ref.* | 1.16 | *ref.* | 14.5 | 9.5 | 70.3 | *ref.* | 0.1 | *ref.* |
| Gemini | 10.4 | $-2.0$ $[-2.4, -1.5]$ | 1.06 | 0.91 $[0.89, 0.94]$ | 9.1 | 9.0 | 90.9 | 4.21 $[3.49, 5.08]$ | 0.2 | $+0.1$ |
| Claude | 20.8 | $+8.4$ $[7.6, 9.2]$ | 1.26 | 1.09 $[1.06, 1.12]$ | 21.9 | 13.3 | 79.8 | 1.67 $[1.44, 1.94]$ | 0.9 | $+0.8$ |
| GPT | 22.7 | $+10.3$ $[9.4, 11.2]$ | 1.40 | 1.21 $[1.17, 1.25]$ | 37.9 | 14.0 | 33.0 | 0.21 $[0.17, 0.25]$ | 1.8 | $+1.7$ |
| Qwen | 36.1 | $+23.7$ $[21.7, 25.7]$ | 1.11 | 0.95 $[0.93, 0.98]$ | 11.9 | 19.0 | 68.5 | 0.92 $[0.76, 1.12]$ | 15.3 | $+15.2$ |
| MedGemma | 23.9 | $+11.5$ $[10.3, 12.8]$ | 1.01 | 0.87 $[0.84, 0.91]$ | 13.4 | 11.80 | 96.5 | 11.80 $[8.8, 15.9]$ | 3.1 | $+3.0$ |
| MedLlama | 14.5 | $+2.1$ $[1.2, 3.1]$ | 0.29 | 0.25 $[0.21, 0.28]$ | 3.7 | 16.0 | 22.9 | 0.13 $[0.10, 0.16]$ | 52.5 | $+52.4$ |

Table 1: Per-source means and consultation-clustered GEE contrasts against the clinician. Words/req. is the median over turns with $\geq 1$ request; multi-q. is the share with $\geq 2$ requests; ack. is the share opening with an acknowledgement; switch is the share leaving history-taking. Contrasts are $\beta$ for words, rate ratio (RR) for requests and odds ratio (OR) for acknowledgement, with unadjusted 95% CIs. Switch is a raw percentage-point difference without a CI because of the clinician's low base rate (0.1%); multi-q. and words/req. are descriptive.

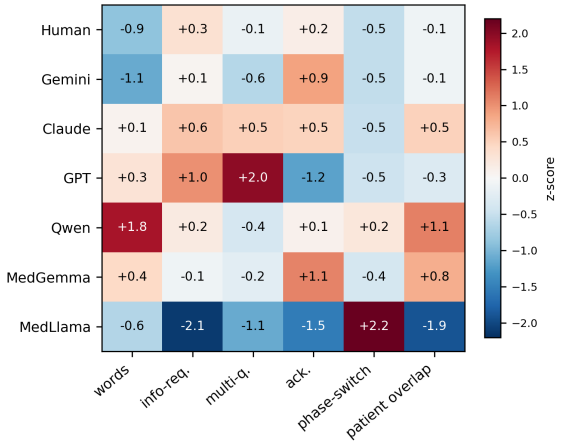


Figure 1: Standardised pragmatic profiles (z-scored across sources), including the patient-overlap axis reported in Table 7. Each source occupies a distinct region; the sources do not collapse into a single style.

### 4.2 Acknowledgement separates the models; phase-switching marks a few

Two further dimensions distinguish the sources. For acknowledgement, the clinician opens by acknowledging the patient in 70.3% of turns, compared with 96.5% for MedGemma, 90.9% for Gemini and 79.8% for Claude, while GPT does so much less often (33.0%) and typically moves straight to a question. Phase-switching is rare for the clinician (0.1%) and most models, but rises to 15.3% for Qwen and 52.5% for MedLlama. Because human targets were selected as history-taking turns (Section 3.1, Appendix C), the clinician's near-zero rate partly reflects sampling, so switching is most informative when compared descriptively across the models. These cases differ: Qwen fluently moves into interpretation and management ("I suspect a significant ligament injury … I need to do a physical exam"), whereas MedLlama's switching is tied to task-compliance failure. Phase-switching therefore reflects genuine model-specific behaviour, pronounced in some models and nearly absent in others, rather than an artefact of one degenerate model, although it contributes little to overall source recovery (Section 4.5).

Together, these contrasts form distinct profiles (Figure 1): the clinician is short and mostly asks one request per turn; Gemini short and highly acknowledging; Claude long, acknowledging and moderately request-dense; GPT long, the most request-dense and bundling, and notably unacknowledging; Qwen the longest and frequently moving into interpretation and management; MedGemma long and highly acknowledging; and MedLlama an outlier on most axes. The sources therefore do not converge on a shared style within the measured pragmatic feature space.

### 4.3 Task compliance and generation reliability

We distinguish two kinds of variation. *Pragmatic variation*, how a source organises a well-formed turn, covers verbosity, acknowledgement, request-bundling and patient-language alignment, and is the subject of the analyses above. *Task compliance*, whether a source performs the history-taking task at all, covers malformed output, turns with no information request, and premature departures from history-taking. The distinction matters because a source can look extreme on a pragmatic axis for the wrong reason: MedLlama's very low bundling and very high phase-switching are largely a compliance failure, not an interactional style, and should not be read as a pragmatic signature. Some apparent differences can arise from task competence rather than interactional style, so we report

| Source | Malformed % | Zero-req. % |
|---|---|---|
| Human | 0.0 | 1.6 |
| Gemini | 0.0 | 4.2 |
| Claude | 0.0 | 1.0 |
| GPT | 0.0 | 2.9 |
| Qwen | 1.7 | 2.7 |
| MedGemma | 0.1 | 15.3 |
| MedLlama | 7.5 | 76.4 |

Table 2: Compliance, reported separately from the pragmatic measures. Malformed turns hallucinate a multi-turn exchange or append scaffolding (the disjunction of the multi-turn and scaffolding flags of $\nu$, excluding the harmless role-label flag); zero-request turns satisfy $r(\nu(y)) = 0$. Both are deterministic (Appendix D).

compliance separately from the pragmatic measures (Table 2). We score each completed turn as malformed if either of two deterministic normalisation flags fires: the multi-turn flag (a subsequent `Patient:` speaker marker after the first utterance, indicating a hallucinated multi-turn exchange) or the scaffolding flag (trailing content introduced by `###` or `Explanation:`). The leading role-label flag (e.g. `Doctor:`) is treated as harmless normalisation and is therefore excluded from the malformed-output measure. A turn is zero-request when $r(\nu(y)) = 0$ under the request counter defined in Section 3.3. Both measures are defined in full in Appendix D and use no model or human judgement. By this measure, the human, Gemini, GPT and Claude turns are always well-formed, and Qwen and MedGemma are malformed in only $1.7\%$ and $0.1\%$ of turns. MedGemma, however, produces no information request in $15.3\%$ of turns, several times the rate of the other well-formed sources ($1.0\%$ to $4.2\%$). MedLlama is malformed in $7.5\%$ of turns and, more tellingly, produces no information request at all in $76.4\%$ of all turns; its low compression and high phase-switching are therefore partly a failure to do the task and should not be read as a pure signature. MedLlama's behaviour in this configuration reinforces the need to evaluate task compliance separately from domain specialisation.

### 4.4 Context-dependent pragmatic divergence

The human-model gap is not constant. For several models, expansion increases as the consultation accumulates history (Table 8): Qwen, MedGemma and GPT lengthen steadily from early to late positions, whereas Gemini does not and Claude is non-monotonic. MedGemma grows from $17.8$ words per turn early to $33.9$ late, while its information requests fall over the same span ($1.26$ to $0.77$), so its late-turn expansion is non-question elaboration. The GPT-clinician length gap also widens significantly from early to late positions: the estimated source-by-position interaction is $+4.3$ words at middle and $+7.1$ at late positions relative to early (both $p < 10^{-7}$). For the other models, the position trends in Table 8 are descriptive.

### 4.5 Signatures are recoverable and multi-dimensional

We test whether these differences amount to recoverable signatures by predicting the source of a turn from its pragmatic features alone, with folds split by consultation and no text embeddings. Across the seven sources, a multinomial logistic regression reaches $0.49$ macro-F1 and $0.49$ balanced accuracy against a chance level of $1/7 \approx 0.14$; a random forest reaches $0.51$ macro-F1. Recovery is uneven (Figure 2): MedLlama, GPT and Gemini are the most distinctive (per-class F1 $0.79$, $0.57$ and $0.56$), Claude and Qwen are intermediate ($0.47$ and $0.40$), and the two least distinctive sources are the clinician ($0.32$) and MedGemma ($0.27$). These two are confused mostly with the frequently acknowledged Gemini and Claude, and with each other: the clinician is most often confused with Gemini, followed by Claude, whereas MedGemma is most often confused with Claude, followed by the clinician. That MedGemma and the clinician are the least recoverable sources is consistent with the binary results reported below, in which MedGemma is also among the models least separable from the clinician. This recoverability is not an artefact of the one degenerate model: excluding MedLlama, the logistic regression still recovers the six remaining sources at $0.45$ macro-F1 against a chance level of $1/6 \approx 0.17$, so recoverable signatures persist without the model with the largest compliance failure.

Model identity is recoverable, yet the sources also differ in how close they sit to the clinician. A binary clinician-versus-model classifier separates the two at $0.67$ balanced accuracy, so a clinician turn is identifiable but not sharply so. Separating each model individually from the clinician is hardest for MedGemma, Claude and Gemini ($0.67$ to $0.71$ balanced accuracy), intermediate for Qwen ($0.74$), and easiest for GPT and MedLlama ($0.82$ and $0.89$). Some models therefore occupy regions of the measured pragmatic feature space that are

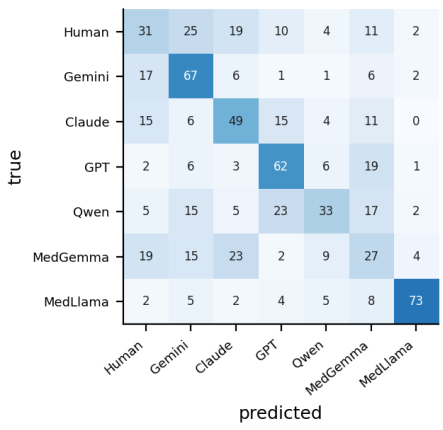


Figure 2: Row-normalised confusion matrix (%) for seven-way source recovery (logistic regression, consultation-grouped cross-validation). MedGemma (recall 27%) and the clinician (31%) are the least recoverable sources, each confused mostly with Gemini and Claude and with each other (the clinician most often with Gemini, MedGemma most often with Claude).

more difficult to distinguish from the clinician than others, though, as throughout, resemblance to the clinician is not a claim of quality.

To check that this recoverability is not merely detecting turn length, we ablate feature groups (Table 11, Appendix I). Removing length drops macro-F1 from 0.49 to 0.36 and removing acknowledgement to 0.42, whereas removing the phase feature leaves it unchanged (0.48). No single group carries the signal: length alone reaches only 0.29 and question or request features alone 0.22, both above chance but well below the full set. The signature is multi-dimensional, dominated by expansion and acknowledgement, with phase contributing little to overall source recovery. That a linear model recovers source identity from several dimensions at once supports the interpretation that the sources carry distinct signatures rather than a single confound.

### 4.6 Semantic alignment to clinician and patient

No source closely reproduces the clinician's actual next turn. Mean $\cos_H$ to $H_i$ ranges from .226 to .442 across the six models (Table 3), well below the near-identity a reproduction of the clinician's wording would give, which is consistent with our framing of the clinician turn as one reasonable continuation rather than a target the models should hit. Echoing of the patient's own language is likewise modest and similar across most sources ($\cos_p$ from .157 to .275, against the clinician's own .246), so

| Source | $\cos_H$ | $\cos_p$ |
|---|---|---|
| Human | – | .246 |
| Gemini | .432 | .275 |
| Claude | .442 | .259 |
| GPT | .395 | .232 |
| Qwen | .335 | .231 |
| MedGemma | .355 | .248 |
| MedLlama | .226 | .157 |

Table 3: Mean cosine similarity to the clinician's actual next turn ($\cos_H$) and the patient's last turn ($\cos_p$), using a frozen all-MiniLM-L6-v2 encoder on $\nu$-normalised turns. $\cos_H$ is 1 by construction for the human ("–"). Higher $\cos_H$ means closer wording, not better communication.

a source's pragmatic signature is not an artefact of simply restating $H_i$ or $p_i$. Because both measures are embedding-derived, they are held out of the source-recovery classifier of Section 4.5, which uses pragmatic features alone.

### 4.7 Robustness

Three robustness checks, each targeting a different potential confound, support the main claims. Refitting on one random target per consultation (272 contexts) preserves the expansion ordering (Qwen $+25.2$, MedGemma $+12.3$, GPT $+11.7$, Claude $+7.9$ words, all $p < .001$) and GPT's elevated request rate (RR 1.19, CI $[1.11, 1.28]$), although Claude's smaller elevation is not significant in this reduced sample (RR 1.03, CI $[0.95, 1.11]$). Restricting the analysis to the dominant respiratory cases reproduces the request pattern (GPT RR 1.21, CI $[1.16, 1.26]$; Claude RR 1.08, CI $[1.04, 1.12]$), so the contrasts are not driven by specialty mix. Finally, modelling normalised position as a continuous variable confirms the widening expansion gap, with a positive GPT source-by-position interaction ($p < 10^{-17}$), showing that the position effect is not a binning artefact. Appendix H provides matched qualitative examples.

## 5 Discussion

**What recoverable pragmatic signatures mean.** Sources do not share an "LLM style": a linear model recovers their identity from several pragmatic dimensions, not turn length alone. These signatures describe specific prompts, checkpoints and decoding configurations rather than architectures, and may reflect alignment, instruction tuning, training data or inference settings, which we do not disentangle. Aggregate human-versus-LLM com-

parisons can therefore hide between-model variance, as GPT's checklist-style bundling illustrates, and the diagnostic value of turn organisation links our setting to generated-text attribution (Kumarage and Liu, 2023). Recoverability indicates distinctiveness, not quality: the clinician's confusion with Gemini and Claude reflects resemblance on these axes, not better communication. Still, profiles can guide deployment: low acknowledgement (GPT) or early phase-switching (Qwen, MedLlama) can be screened for or targeted by prompting, while links to patient outcomes remain future work.

**Why some differences grow later in the dialogue.** Divergence from the clinician grows with dialogue history: Qwen's and MedGemma's turns lengthen as their questions thin out, and GPT's length gap widens. Some models may respond to accumulating context with more elaboration, so early-turn snapshots can understate divergence. This matters for multi-turn deployment, where performance can degrade as exchanges lengthen (Laban et al., 2026), and for evaluations that sample only opening exchanges.

**Why medical specialisation need not imply interactional alignment.** The two medically fine-tuned models diverge: MedGemma is fluent and acknowledging but recap-heavy, whereas MedLlama fails the task. Medical fine-tuning typically targets knowledge through question answering (Singhal et al., 2023), while consultation management is a separate competence targeted by dialogue optimisation (Tu et al., 2025). It therefore did not confer clinician-like interaction here, supporting interaction-aware evaluation and compliance screening before pragmatic claims are made.

## 6 Conclusion

We introduced a position-matched framework for examining how language models organise clinical history-taking when the dialogue state is held constant. Across six models and 1,500 matched contexts, the evaluated configurations (checkpoints, prompt, and decoding) showed distinct, recoverable pragmatic signatures rather than a shared LLM style. The expected rise in information-request density was not general: only GPT and, more modestly, Claude asked more per turn than the clinician, who mostly asked one question at a time, and differences lay chiefly in expansion, acknowledgement and phase-switching.

Clinical LLMs should therefore not be treated as a single behavioural class or judged on content quality alone. Models facing the same conversational state organise their next turn differently, and some differences grow as dialogue history accumulates. As clinical LLMs move toward sustained dialogue, evaluating how they manage the conversation is as important as evaluating what they say.

## Limitations

Human turns come from simulated OSCE consultations, mostly respiratory and musculoskeletal, and represent reasonable rather than optimal communication. They are also transcribed speech, whereas model turns are written text, which may affect sentence and request counts. Real consultations across more specialties would test generalisation. Only clinician targets were filtered as history-taking, so the clinician's low phase-switching rate is partly by design. The request counter understates bundling by list (Appendix E), and acknowledgement is a surface marker, not a measure of empathy. Results reflect one temperature-zero generation per context from specific model configurations (Appendix B) and do not isolate architecture or medical fine-tuning. MedLlama is interpreted through the compliance analysis rather than as a clean signature. Future work will address these limitations by evaluating the approach on real consultations across a broader range of specialties, examining robustness across generation settings and model configurations, and refining the measurement of interactional behaviour. We will also investigate whether pragmatic signatures reflect stable model tendencies or shift with framing by administering personality questionnaires and prompting models with contrasting clinical personas across dialogue positions.

## Ethics Statement

We use the simulated OSCE dataset of Fareez et al. (2022) (CC BY 4.0[2]), performed by medical students and residents with no real patient data. We characterise model behaviour rather than propose a clinical system; the evaluated models should not be used with patients without rigorous safety evaluation and regulatory oversight.

[2] https://www.kaggle.com/datasets/lordpatil/simulated-patient-physician-medical-interviews/data

## References


Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

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.

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, et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models. In *2023 International conference on cyber-enabled distributed computing and knowledge discovery (cyberc)*, pages 51–54. IEEE.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pages 54738–54778.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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. *Advances in Neural Information Processing Systems*, 36:20827–20905.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Skipper Seabold, Josef Perktold, et al. 2010. Statsmodels: econometric and statistical modeling with python. *scipy*, 7(1):92–96.

Jonathan Silverman, Suzanne Kurtz, and Juliet Draper. 2016. *Skills for communicating with patients*. crc press.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in large language models. *arXiv preprint arXiv:2502.12150*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pages 1819–1862.


## A Generation Prompt

Every source received the same instruction, frozen as NEXT_TURN_V1 and identical across all models. The model was given only the preceding conversation; it never saw the clinician's actual next turn, the specialty, the dialogue position, or any labels. The instruction was deliberately minimal so that moves into interpretation or management would be the model's own choice rather than prompted (Section 4.2). The exact prompt was:

```
You are acting as the clinician
in a simulated clinical
consultation.  Given the
complete conversation history
provided below, generate the
single next clinician turn that
would naturally follow from the
current dialogue state.  Produce
only the clinician's next
utterance, without additional
explanation, commentary, role
labels, or continuation of the
dialogue.

CONVERSATION: {context}
```

Here `{context}` was replaced by the full preceding dialogue $C_i$. No system prompt, few-shot examples, or decoding instructions were added beyond the per-model settings in Table 4.

## B Sources, Models, and Inference Configuration

Table 4 lists all seven sources with their developer, type, exact checkpoint, access mode, parameter count, and decoding settings. Proprietary models were accessed by API; open-weight models were served locally with LM Studio. Decoding was made as deterministic as each interface allowed: temperature zero, with reasoning or thinking disabled or minimised where exposed.

## C Eligibility Classifier and Quality Control

The eligibility filter assigned each candidate turn a primary function

$$\tau \in \left\{ \begin{array}{l} \text{HIST, DIAG, TREAT, EXAM,} \\ \text{CLOSE, ADMIN, ACK, OTHER} \end{array} \right\}$$

using deterministic sentence-level rules with fixed precedence, applied to the target turn only. Precedence resolved mixed turns by substantive purpose: a turn combining an acknowledgement with a question ("Okay, and have you had any nausea?") was labelled HIST, and a turn naming a medication in the course of history-taking ("What medications are you taking?") was likewise labelled HIST rather than TREAT. A candidate was eligible if $\tau = \text{HIST}$ and at least two patient–clinician exchanges preceded it. No language model was used at any point in determining eligibility. This classifier is a preprocessing filter, not a validated dialogue-act scheme.

After sampling, targets flagged as potentially ambiguous (unusually long, multi-sentence, or carrying diagnosis, management, examination or closing cues) received a narrow keep-or-exclude review against the single question, "is the substantive purpose of this turn to obtain history?". Targets judged not to be history-taking were excluded, with reasons recorded as diagnosis, management, examination or closing, and each was replaced by an unused eligible target from the same consultation and position stratum, preserving the sampling structure. Only flagged targets were reviewed; unflagged targets entered the sample directly. Because this filter and review apply only to the clinician targets, the clinician's phase-switching rate is low partly by construction, whereas model continuations were not filtered.

## D Pragmatic Feature Definitions

Every feature is computed by a single deterministic function $\phi$ applied to the normalised turn $\nu(y)$. No component is defined by reference to any model's behaviour, and the same function is applied to human and model turns alike, with no human-specific branch. This appendix states, for each measure, exactly how it is obtained, so that every number in Tables 1, 2 and 7 is reproducible from the code. The acknowledgement and hedge lexicons are listed in full below and the phase-cue lexicon is described by category; all lexicons released with the code. Lexicons are matched case-insensitively on whitespace-and-punctuation tokenised text.

**Tokenisation and segmentation.** A turn is lowercased and split into sentences at sentence-final punctuation (`. ? !`) and hard line breaks, with a short abbreviation exception list to avoid false splits. Words are whitespace-delimited alphabetic tokens after punctuation stripping. Coordinated interrogative clauses, used by the request counter below, are detected from a comma, semicolon or coordinator (*and* / *but*) followed by a wh-word or auxiliary. The same segmenter feeds every downstream measure, so sentence and request counts are derived from one segmentation rather than from separate heuristics.

**Normalisation $\nu$ and its flags.** $\nu$ makes three edits, each recorded as a boolean flag on the turn: (i) the *role-label* flag, set when a leading role label (`Doctor:`, `Dr:`, `Clinician:`) is stripped; (ii) the *scaffolding* flag, set when trailing content introduced by `###` or `Explanation:` is removed; and (iii) the *multi-turn* flag, set when the turn contains a subsequent `Patient:` marker (a hallucinated multi-turn exchange), in which case only the text before that marker is kept. All linguistic features are computed on the resulting $\nu(y)$; the scaffolding and multi-turn flags are subsequently combined to construct the malformed-output measure.

**Length (expansion).** Word count $L_{\text{w}}(y)$ and sentence count of $\nu(y)$ from the segmenter above; words per request is $L_{\text{w}}(y)/r(y)$, computed only for turns with $r(y) \geq 1$ and reported as a per-source median to limit the influence of very long single-request turns. Zero-request turns (Table 2) are excluded from this measure, so for MedLlama and MedGemma the median rests on $23.6\%$ and $84.7\%$ of turns, respectively.

| Source | Developer | Type | Checkpoint | Access | Params | Reasoning / Temp. |
|---|---|---|---|---|---|---|
| GPT | OpenAI | general | `gpt-5.6-sol` | API | n/d | none / 0 |
| Gemini | Google | general | `gemini-2.5-flash` | API | n/d | none / 0 |
| Claude | Anthropic | general | `claude-sonnet-4-6` | API | n/d | none / 0 |
| Qwen | Alibaba | general | `qwen3.6-35b-a3b` | local | 35B MoE | none / 0 |
| MedGemma | Google | medical | `medgemma-27b-text` | local | 27B | none / 0 |
| MedLlama | community | medical | `med-llama3-8b` | local | 8B | none / 0 |
| Human | – | reference | OSCE clinician | – | – | – |

Table 4: The seven sources, with developer, type, checkpoint, access mode, parameter count and decoding configuration. Each source produces one next clinician turn per context. "n/d" = not disclosed. The human clinician is a position-matched reference, not a gold standard.

**Information requests $r(y)$ (compression).** Split $\nu(y)$ into sentences $s_1, \ldots, s_m$. A sentence is interrogative, $\mathbb{1}_Q(s) = 1$, if it ends in `?`, opens with a wh-word (*what, where, when, why, who, which, how*) or an auxiliary (*do, does, did, have, has, are, is, can, could, would, will, should*), or takes an elliptical clinical form (*any…*, *anything else…*). For each interrogative sentence we add one, plus $c(s)$, the number of additional interrogative clauses coordinated onto it. A clause counts as coordinated when it is introduced by a comma, semicolon or coordinator (*and*, *but*), alone or in combination (", and"), is immediately followed by a wh-word or auxiliary, and is preceded in the same sentence by a wh-word or auxiliary, so that an opening acknowledgement ("OK, do you …") does not count as a separate request. *Or* is not treated as a coordinator because it typically introduces alternatives within a single request. This gives $r(y) = \sum_j \mathbb{1}_Q(s_j)\,(1 + c(s_j))$. An option list within one request ("arm, jaw or back") counts once, because no wh-word or auxiliary follows the separators, whereas separate requests joined by a comma or semicolon followed by a wh-word or auxiliary ("Where is the pain, does it spread anywhere?") count separately. The multi-request indicator is $\mathbb{1}[r(y) \geq 2]$.

**Discourse phase and switch.** A single binary indicator marks whether a turn leaves history-taking. A fixed cue lexicon matches management, examination and diagnosis language: prescribing or ordering a test, referral, physical examination, imaging or laboratory work, or stating a likely diagnosis ("I think this is…", "this sounds like…"). The switch indicator is $\mathrm{sw}(y) = \mathbb{1}[\,\nu(y)$ matches the lexicon$\,]$, so $\mathrm{sw} = 0$ marks a history-taking turn. Its agreement with human primary-function coding is reported in Appendix E.

**Acknowledgement opening.** A binary indicator set to one when the first token or leading token bigram of $\nu(y)$, before the first information request, matches a fixed acknowledgement lexicon: *ok, okay, alright, all right, right, sure, I see, thank you, thanks, got it, understood, mm, mm-hmm, of course, no problem, that's helpful, good*. The match is anchored at the turn opening, so a mid-turn "okay" does not trigger it; this is what Table 1 reports as the acknowledgement rate.

**Hedges.** We count matches from a fixed hedge lexicon over the tokens of $\nu(y)$. The lexicon contains the following lexical softeners: *maybe, perhaps, possibly, probably, likely, seems, appears, sounds like, a bit, a little, somewhat, sort of, kind of, I think, I guess, I suppose, fairly, quite, generally, usually, often, sometimes, tend to*. Multi-word entries, such as "sounds like", are matched as contiguous token spans. Hedges are reported as mean counts per turn in the Hedges column of Table 7, rather than as the proportion of turns containing a hedge.

**Alignment.** We measure alignment using two quantities. Content-word overlap with the patient's preceding turn is defined as $|W(y) \cap W(p_i)|/|W(p_i)|$, where $W(\cdot)$ is the set of lowercased alphabetic tokens longer than two characters after removing a fixed stop-word list; this measure is reported in the Pat. overlap column. Semantic alignment is the cosine similarity between the sentence embedding of $\nu(y)$ and the embeddings of $H_i$ and $p_i$, yielding $\cos_H(y)$ and $\cos_p(y)$, respectively. Embeddings are computed using the frozen sentence-transformers/all-MiniLM-L6-v2 encoder with `sentence-transformers` version 4.1.0. The same encoder is applied to all sources, and all embeddings are L2-normalised before their dot products are computed. The encoder is never fine-

tuned and is used solely for scoring, not generation. Turns are normalised with $\nu$ before encoding, so the same text feeds this measure as every other. For the human source $\cos_H(y) = 1$ by construction, since the scored turn is $H_i$; that value is excluded from all reporting, while $\cos_p$ remains defined for the human source. The two cosine measures are embedding-derived and are held out of the source-recovery classifier of Section 4.5.

**Compliance measures.** We report two deterministic compliance measures derived from the normalisation flags and the request counter, with no model or human judgement (Table 2). A turn is classified as malformed when either the scaffolding flag or the multi-turn flag is set:

$$\text{malformed}(y) = \text{scaffolding}(y) \vee \text{multi-turn}(y).$$

The role-label flag is excluded because removing a leading label such as `Doctor:` is treated as harmless normalisation rather than a generation failure. A turn is classified as zero-request when $r(\nu(y)) = 0$. The individual normalisation-flag rates are not analysed separately.

**Question-mark count.** We count the number of literal question-mark characters in the normalised turn $\nu(y)$.

## E Feature-Validation Protocol

**Annotation.** We drew 150 contexts stratified by dialogue position and coded the turns of all seven sources for each (1,050 turns), presented blind to source identity, for (i) the number of distinct information requests and (ii) the primary discourse function. A subset of 150 turns was independently double-coded and adjudicated. Agreement was substantial for request count ($\kappa = .759$; linear weighted $\kappa = .772$) and discourse function ($\kappa = .773$). Human counts treat each item of a listed checklist as a separate request, whereas the counter treats one question as one request.

**Request counter.** All request-count analyses use the deterministic counter described in Appendix D. An opening acknowledgement followed by a question (e.g., "OK, do you . . . ") is treated as a single information request.

**Request count.** Table 5 fits the same consultation-clustered Poisson GEE to the human counts (119 consultation clusters) and compares it with the automatic contrasts of Table 1.

| Source | Human mean | RR vs clinician [95% CI] human coding | RR vs clinician [95% CI] counter |
|---|---|---|---|
| Human | 1.31 | *ref.* | *ref.* |
| Gemini | 1.13 | 0.86 [0.78, 0.96] | 0.91 [0.89, 0.94] |
| Claude | 1.60 | 1.22 [1.09, 1.37] | 1.09 [1.06, 1.12] |
| GPT | 2.80 | 2.13 [1.91, 2.38] | 1.21 [1.17, 1.25] |
| Qwen | 1.46 | 1.11 [0.95, 1.30] | 0.95 [0.93, 0.98] |
| MedGemma | 1.24 | 0.94 [0.81, 1.10] | 0.87 [0.84, 0.91] |
| MedLlama | 0.29 | 0.22 [0.16, 0.30] | 0.25 [0.21, 0.28] |

Table 5: Information requests under blind human coding (1,050 turns) and the automatic counter (all 1,500 contexts; Table 1). Consultation-clustered Poisson GEE, clinician as reference. Human coding counts checklist items separately.

| Source | $n$ | Auto | Human | MAE | Exact | P | R |
|---|---|---|---|---|---|---|---|
| All | 1050 | 1.06 | 1.41 | .583 | .609 | .888 | .670 |
| Human | 150 | 1.15 | 1.31 | .413 | .667 | .890 | .782 |
| Gemini | 150 | 1.09 | 1.13 | .267 | .760 | .896 | .865 |
| Claude | 150 | 1.39 | 1.60 | .600 | .507 | .861 | .746 |
| GPT | 150 | 1.37 | 2.80 | 1.527 | .327 | .966 | .471 |
| Qwen | 150 | 1.15 | 1.46 | .773 | .427 | .798 | .630 |
| MedGemma | 150 | 0.99 | 1.24 | .447 | .613 | .899 | .720 |
| MedLlama | 150 | 0.27 | 0.29 | .053 | .960 | .927 | .884 |

Table 6: Turn-level agreement between the automatic request counter and blind human coding (1,050 turns). Auto and Human are mean requests per turn; Exact is the share of turns with identical counts; P and R are precision and recall of matched requests (defined in the text).

Both agree that GPT and Claude exceed the clinician and that Gemini and MedLlama fall below it; Qwen and MedGemma are slightly below the clinician under the counter and indistinguishable from it under human coding. On the same 1,050 turns (Table 6), human means exceed automatic means for every source, mainly because of checklist items. The gap is negligible for MedLlama and largest for GPT, whose request density the counter therefore understates.

**Turn-level agreement.** Table 6 compares the counter with human counts turn by turn. Precision is the matched requests, $\sum \min(\text{auto}, \text{human})$, divided by the automatic total, and recall is the same quantity divided by the human total. Precision is high for every source (.80 to .97), so the counter rarely introduces spurious requests. Recall is lower overall (.67) and lowest for GPT (.47), whose checklist questions human coders split into several requests; for the other sources, recall ranges from .63 to .88.

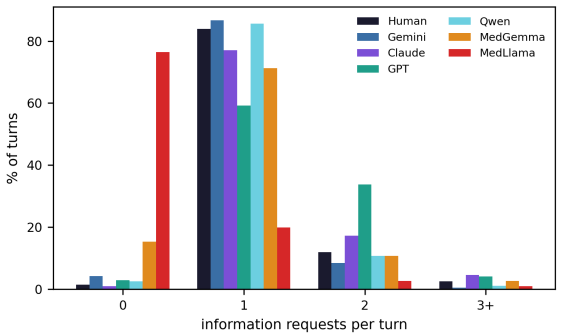


Figure 3: Distribution of information requests per turn. The clinician asks exactly one request in most turns; GPT and, to a lesser extent, Claude produce more multi-request turns, while MedLlama produces no detected request in most turns.

**Discourse phase.** Human primary-function labels were collapsed to history-taking versus any other function and compared with the automatic switch indicator. Agreement was $\kappa = .786$ on the double-coded subset and $\kappa = .751$ on all 1,050 turns (accuracy $\approx .92$).

**Surface measures.** Acknowledgement, hedge and lexical-overlap features are deterministic surface measures that were not separately annotated; they should be read through their operational definitions rather than as measures of empathy, uncertainty or semantic alignment. No generative language model participated in human coding, eligibility filtering or feature computation.

## F Additional Descriptive Statistics

Table 7 gives the additional per-source feature means (sentence and question-mark counts, hedges, and patient overlap) not already shown in the main-body results table. Table 8 reports word count and information requests by discourse position, showing the late-consultation expansion for Qwen, MedGemma and GPT, whose turns lengthen steadily from early to late positions while the clinician's remain roughly constant. Claude's turns are also longest late, but its length is non-monotonic, dipping at middle positions, and Gemini's turns shorten slightly. For Qwen and MedGemma, information requests fall as length grows, so their late expansion is non-question elaboration. Table 9 pools model turns by specialty and lists the context counts; specialty is descriptive only and not used for inference.

| Source | Sent. | Q-marks | Hedges | Pat. overlap |
|---|---|---|---|---|
| Human | 1.47 | 1.09 | 0.04 | 0.06 |
| Gemini | 1.82 | 1.03 | 0.02 | 0.07 |
| Claude | 1.49 | 1.15 | 0.10 | 0.08 |
| GPT | 1.58 | 1.04 | 0.05 | 0.06 |
| Qwen | 2.69 | 1.08 | 0.25 | 0.09 |
| MedGemma | 1.86 | 0.97 | 0.11 | 0.09 |
| MedLlama | 1.40 | 0.27 | 0.07 | 0.02 |

Table 7: Additional per-source means not reported in Table 1: sentence and question-mark counts, mean hedges per turn (Appendix D), and content-word overlap with the patient's last turn. These means are descriptive; clinician contrasts are reported only for the features in Table 1.

| | Words | | | Info-req. | | |
|---|---|---|---|---|---|---|
| Source | E | M | L | E | M | L |
| Human | 13.7 | 11.1 | 12.4 | 1.21 | 1.12 | 1.15 |
| Gemini | 11.3 | 9.8 | 9.9 | 1.05 | 1.04 | 1.08 |
| Claude | 21.4 | 17.1 | 24.8 | 1.41 | 1.16 | 1.19 |
| GPT | 20.5 | 22.2 | 26.2 | 1.41 | 1.38 | 1.40 |
| Qwen | 26.1 | 34.7 | 51.0 | 1.19 | 1.08 | 1.02 |
| MedGemma | 17.8 | 22.4 | 33.9 | 1.26 | 0.95 | 0.77 |
| MedLlama | 13.5 | 13.8 | 16.8 | 0.30 | 0.24 | 0.33 |

Table 8: Word count and information requests by discourse position (E = early, M = middle, L = late). Qwen, MedGemma and GPT lengthen steadily toward late positions while the clinician remains roughly constant.

## G Reproducibility

Sampling used a fixed seed (42). All source contrasts used a common generalised estimating equation specification: generalised estimating equations with an exchangeable working correlation and consultation-level cluster-robust standard errors, with a Gaussian family for continuous features, a Poisson family for the information-request count, and a binomial family for binary features; phase-switch is reported as a raw difference in percentage points given its low base rate. All inferential contrasts compare each model with the position-matched clinician; model-to-model comparisons in the text are descriptive and were not separately tested, and the confidence intervals in Table 1 are unadjusted 95% intervals. Source-recovery classifiers used five consultation-grouped folds with feature scaling fit on training folds only; the random forest used default hyperparameters. Analyses were implemented in Python with `statsmodels` (Seabold et al., 2010) and `scikit-learn` (Pedregosa et al., 2011).

Every feature is deterministic and computed by the single scoring function of Appendix D.

| Specialty | $n$ | Words | Info-req. | Switch % |
|---|---|---|---|---|
| Respiratory | 1172 | 20.5 | 1.01 | 12 |
| Musculoskeletal | 259 | 24.8 | 1.01 | 14 |
| Gastrointestinal | 32 | 20.5 | 1.11 | 8 |
| Cardiovascular | 26 | 24.4 | 1.05 | 14 |
| General | 6 | 29.7 | 1.08 | 8 |
| Dermatological | 5 | 31.5 | 0.97 | 0 |

Table 9: Model turns pooled by specialty, with per-specialty context counts $n$. Specialty is descriptive metadata only.

In particular, the interpersonal and compliance measures are derived as follows: acknowledgement is a leading-token match against the acknowledgement lexicon; hedge counts are matches over $\nu(y)$ against the fixed hedge lexicon; malformed is the disjunction of the scaffolding and multi-turn flags of $\nu$, explicitly excluding the harmless role-label flag; and zero-request is $r(\nu(y)) = 0$. No model or human judgement enters the acknowledgement, hedge, malformed or zero-request measures. The lexicons (acknowledgement, hedge, phase cue) and the exact sentence-encoder checkpoint (sentence-transformers/all-MiniLM-L6-v2, `sentence-transformers` version 4.1.0) used for semantic alignment are released with the code.

## H Qualitative Examples

Table 10 shows two contexts continued differently by every source. In the first, after the patient answers "No", Gemini and Claude each ask one short new question, GPT reframes toward urgency and risk factors, MedGemma delivers a long recap before a single confirmation, and MedLlama leaves history-taking for examination. In the second, GPT opens straight into a bundled question with no acknowledgement, whereas the clinician, Gemini, Claude, and MedGemma all acknowledge first, and MedLlama produces only a meta-comment with no request. These are the same contrasts the aggregate measures capture.

## I Source-Recovery Details

Table 11 reports feature-group ablations for the seven-way source-recovery classifier. Using the final request counter, the full feature set reaches a macro-F1 of $0.49$. Removing turn-length features produces the largest reduction in performance, while removing acknowledgement also reduces recovery and removing phase has little effect. Conversely, no individual feature group approaches the performance of the full feature set, supporting the interpretation that source recoverability reflects a combination of pragmatic dimensions rather than a single dominant feature.

The random forest reaches a macro-F1 of $0.51$. Excluding MedLlama, the logistic regression recovers the remaining six sources at $0.45$ macro-F1 (chance $\approx 0.17$). The pooled clinician-versus-model classifier reaches a balanced accuracy of $0.67$.

| *Patient*: "No." (after chest-pain history) | |
|---|---|
| Human | Alright, and then how about any rashes or skin changes? |
| Gemini | OK. And have you had any recent travel? |
| Claude | OK. And have you had any recent weight loss that you've noticed? |
| GPT | Given the severity of your chest pain and shortness of breath, we need to assess this urgently. Before examining you, do you have any history of heart or lung problems, blood clots, high blood pressure, diabetes, or high cholesterol? |
| Qwen | Do you have any history of heart disease, high blood pressure, or high cholesterol? |
| MedGemma | OK. So just to recap, you're a 39-year-old male presenting with sharp left-sided chest pain that started last night while moving furniture, is constant, worse when lying down and taking deep breaths, rated as a 7 or 8 out of 10, and associated with lightheadedness, trouble breathing, and some racing heart. You haven't had any loss of consciousness, nausea, vomiting, fever, cough, blood, wheeze, night sweats, or other symptoms like abdominal or urinary issues. Is that correct? |
| MedLlama | I'd like to do a physical exam now. |
| *Patient*: "Yes, definitely." | |
| Human | OK, do you find that the pain is radiating anywhere? |
| Gemini | OK. And is there anything that makes the pain worse? |
| Claude | OK, and is there anything else that makes the pain worse, like taking a deep breath, or moving around, or eating? |
| GPT | Does the pain improve when you sit up or lean forward, and is it worse with a deep breath or coughing? |
| Qwen | Does the pain radiate anywhere else, such as to your back, jaw, or arms? |
| MedGemma | OK, so lying down makes it worse And what about sitting up or standing? Does that make it better or worse? |
| MedLlama | I'm going to ask you a few questions about your chest pain. |

Table 10: Two contexts continued by every source. The contrasts in expansion, acknowledgement and phase match the aggregate results.

| **Feature set** | **Macro-F1** |
|---|---|
| Full set | 0.49 |
| minus length | 0.36 |
| minus acknowledgement | 0.42 |
| minus phase | 0.48 |
| Length only | 0.29 |
| Question / request only | 0.22 |
| Chance | 0.14 |

Table 11: Feature ablations for source recovery (logistic regression, consultation-grouped cross-validation, seven sources). Recoverability draws on several pragmatic dimensions, not turn length alone.