CAPE: A Benchmark for Schema-Grounded Clinical Action Plan Extraction from doctor-patient Dialogue

Anonymous ACL submission

Abstract

Doctor-patient conversations contain management decisions that are implicit, revised, and scattered, yet existing benchmarks evaluate note generation rather than faithful plan recovery. We introduce **CAPE**, a benchmark for schema-grounded treatment-plan extraction from clinical dialogue: 1,900 physician-annotated synthetic conversations spanning 21 ICD-10 chapters and seven management domains, with evaluation combining surface fidelity, semantic similarity, and structured error counts validated against physician judgments. Across 36 models, general instruction-following systems outperform medical-specific models; matched-scale comparisons confirm schema adherence, temporal reasoning, and hallucination control are bottlenecks unaddressed by domain specialization. Fine-tuning reduces hallucinations by 62–68% and enables 7–8B models to outperform larger medical-specific baselines. CAPE provides a leaderboard and testbed for structured clinical plan extraction.

1 Introduction

doctor-patient conversations encode the management decisions that determine post-encounter care, yet these decisions are negotiated across turns, revised after clarification, and mixed with past or conditional actions. Existing clinical dialogue benchmarks primarily evaluate note generation and summarization (Abacha et al., 2023b; Yim et al., 2023; Abacha et al., 2023a), which does not directly test whether models preserve plan structure: fluent clinical text can still contain omissions, unsupported additions, or clinically consequential errors (Abacha et al., 2025; Agarwal et al., 2024).

We study **treatment-plan extraction**: converting a doctor-patient dialogue into a hierarchical representation of the clinician's intended management plan. The task requires recovering actionable decisions across investigations, monitoring, treatment, follow-up, referral, patient education, and return precautions, while distinguishing active plans from past, negated, or merely discussed actions. This schema-grounded framing aligns with the broader move from free-text documentation toward structured, interoperable clinical representations (HL7 International, 2019a,c,b).

We introduce **CAPE**, a benchmark of 1,900 synthetic doctor-patient dialogues paired with physician-annotated structured treatment plans spanning 21 ICD-10 chapters. Ten resident physicians annotated the dataset under two attending clinicians, with qualification testing and adjudication supporting annotation quality. We validate evaluation against physician judgments on 300 instances, showing that chrF (Popovic, 2015) and MoverScore (Zhao et al., 2019) capture complementary aspects of quality; we report both alongside explicit counts of omissions, hallucinations, schema misplacement, and temporal-scope errors.

Across 36 models, the strongest general instruction-following systems outperform the strongest evaluated medical-specific systems, and matched-scale comparisons confirm that schema adherence, temporal reasoning, and hallucination control matter more than domain specialization alone. Supervised fine-tuning reduces hallucinated items by 62–68% and enables 7–8B open-source models to outperform larger medical-specific baselines.

Our contributions are: (1) a definition of treatment-plan extraction as a schema-grounded clinical dialogue task; (2) CAPE, a 1,900-case physician-annotated benchmark with broad ICD-10 coverage; (3) a physician-validated evaluation framework combining chrF, MoverScore, and clinically meaningful error counts; and (4) a 36-model benchmark and fine-tuning analysis showing that schema adherence, temporal scoping, and hallucination control remain central challenges for clinically reliable extraction.

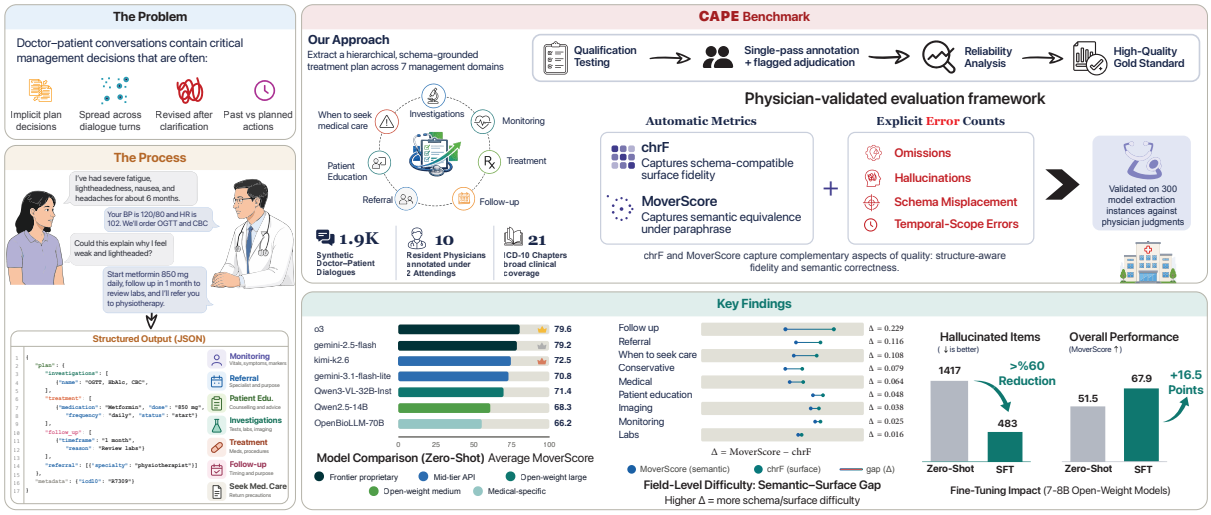

Figure 1: Overview of CAPE. The figure summarizes the motivation for schema-grounded clinical treatment-plan extraction, the CAPE benchmark construction and physician-validated evaluation framework, and the main empirical findings across zero-shot models, fine-tuning, and field-level difficulty.

2 Related Work

Clinical dialogue benchmarks such as MTS-Dialog, ACI-BENCH, and MEDIQA-Chat evaluate note generation and summarization from doctor-patient conversations (Abacha et al., 2023b; Yim et al., 2023; Abacha et al., 2023a), but do not require recovering management decisions as structured objects with field boundaries and temporal scope. The concurrent MEDIQA-OE 2025 shared task (Abacha et al., 2023a) targets medical order extraction from dialogue; CAPE differs in its hierarchical seven-domain schema, physician-validated metric framework, and 36-model diagnostic analysis. Entity-level clinical IE from written records is well established through shared tasks like i2b2/n2c2 (Henry et al., 2020) and systems such as MedCAT (Kraljevic et al., 2021), yet these target flat concept lists rather than nested plan structures. Temporal clinical IE resources such as THYME (Styler IV et al., 2014) address event temporality in text but not the problem of distinguishing active plans from past, conditional, or discussed actions in spoken dialogue.

Standard generation metrics align poorly with clinical quality; (Wang et al., 2023) show that automated metrics for medical summarization can be anti-correlated with human judgments, while embedding-based metrics like BERTScore (Zhang et al., 2019) can over-credit semantically related but clinically distinct items. CAPE validates its metrics against physician judgments and reports complementary surface and semantic scores alongside explicit error counts.

LLMs have shown strong clinical QA and documentation performance (Singhal et al., 2025; Nori et al., 2023), but evaluations predominantly test factual recall or free-text generation rather than schema-grounded structured extraction. CAPE targets this gap: faithful recovery of a clinician's intended plan into a hierarchical, FHIR-aligned representation from dialogue.

3 Dataset

We introduce **CAPE**, a benchmark of 1,900 synthetic doctor-patient dialogues paired with physician-annotated structured treatment plans. CAPE is built from *MedSynth* (Mianroodi et al., 2025), a publicly available corpus of synthetic doctor-patient dialogues, where each dialogue is associated with a unique ICD-10 diagnostic code. The final dataset spans 21 ICD-10 chapters and is split into 1,400 training and 500 test cases using stratified sampling by ICD-10 chapter, with no dialogue appearing in both partitions. Dataset statistics are in Appendix B and Table 7.

3.1 Schema

We define treatment-plan extraction as recovering the clinician's intended management plan from a dialogue into a structured JSON object. Formally, given a doctor-patient dialogue $d = (u_1, \ldots, u_n)$, where each u_i is a speaker-attributed utterance, the task is to produce a structured plan $y \in \mathcal{Y}$

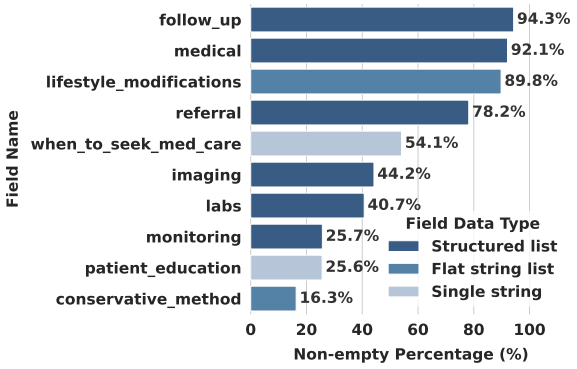

Figure 2: Prevalence of annotated fields and subfields in the CAPE dataset.

that instantiates the CAPE schema. The output should include only clinician-intended future or ongoing management actions explicitly supported by the dialogue, while excluding past, negated, conditional, or merely discussed actions unless they are part of the active plan. The schema contains seven management domains: INVESTIGATIONS, MONITORING, TREATMENT, FOLLOW-UP, REFERRAL, PATIENT EDUCATION, and WHEN TO SEEK MEDICAL CARE. These domains cover tests to order, parameters to monitor, treatments to initiate or modify, referrals, follow-up plans, patient-facing instructions, and safety-netting advice.

The schema is designed to make clinically actionable distinctions explicit: for example, separating medication decisions from conservative procedures and lifestyle advice, follow-up timing from follow-up purpose, and referral specialty from referral aim. It follows the Plan component of SOAP-style documentation (Weed et al., 1968; Podder et al., 2023) and is structurally aligned with FHIR CarePlan, ServiceRequest, and MedicationRequest resources (HL7 International, 2019a,c,b). Absent list-valued fields are annotated as empty lists, while unstated attributes are annotated as `null`; annotators were instructed not to infer plausible but unverbalised details. Full schema definitions, controlled vocabularies, and field-to-standard mappings are provided in Appendix C.

Figure 2 summarizes the prevalence of the main annotated fields and subfields in CAPE. The dataset contains both near-universal fields, such as `follow_up` and treatment-related `medical` plans, and sparse fields, such as `conservative_method`. This heterogeneity reflects the variable clinical relevance of different plan components across diagnoses and motivates our separate tracking of item-level and subfield-level omissions and hallucinations in the evaluation protocol.

3.2 Annotation and Reliability

Annotation was performed by ten resident physicians with outpatient documentation experience and supervised by two attending clinicians with expertise in internal medicine and clinical informatics. Annotators completed training and a supervised 100-case qualification assessment before production; each was required to achieve at least 0.80 field-level syntactic F1 on every field against a supervisor-adjudicated gold standard. Production annotation used a sharded initial single-pass design, with each case assigned to one qualified annotator and with source-quality exclusions replaced to maintain the 1,900-case target. To strengthen the evaluation reference set, all 500 test cases were subsequently reviewed by a five-physician panel before leaderboard scoring, with flagged disagreements adjudicated under the same annotation guidelines. Full dual annotation of all 1,900 cases was not feasible under the physician-annotation budget; instead, we prioritized broad ICD-10 coverage while enforcing a per-field qualification gate, interface-level validation, test-set review, and flag-based adjudication for uncertain or low-quality cases.

Reliability was assessed using two held-out reliability checks: independent annotations of a 100-case qualification set by all annotators and a separate 50-case production-overlap subset annotated independently by both supervisors. These two reliability sets were disjoint from each other and from the final training/test cases used for model evaluation, and the qualification set was used only for assessment rather than as practice material. The five-physician test-set review was used for reference-set quality control rather than for estimating these reliability coefficients. Metrics were selected by field type: Fleiss' κ for structured list fields, Coverage F1 for flat string-list fields, and BERTScore F1 for free-text fields. Overall annotator-level Krippendorff's α was 0.812, while supervisor-level agreement reached mean pairwise agreement of 0.888 and Krippendorff's $\alpha = 0.881$, providing an empirical ceiling for the task. We therefore view the initial single-pass production design as a coverage-oriented benchmark construction choice supported by explicit qualification, validation, test-set review, and adjudication rather than as a substitute for exhaustive dual annotation. Qualitative

disagreement review further indicated that most residual disagreements reflected clinically meaningful schema-boundary cases, especially conservative management versus lifestyle advice, item splitting, and whether unstated attributes should remain null. Full annotation protocol, per-field reliability results, and disagreement analysis are reported in Appendices D and E.

4 Evaluation

4.1 Empirical Metric Validation

To justify our metric choices empirically, we correlated six candidate metrics against physician judgment on 300 model-extraction instances.

Study design. We sampled 100 cases from the 500-case CAPE test set and collected predictions from three pre-specified anchor models: *gemini-3-flash-preview*, *gpt-5-mini*, and *Llama-3.1-8B-Instruct*, yielding 300 case-model instances. The anchors were fixed before leaderboard scoring, per-field analysis, or metric comparison, and were chosen only to produce heterogeneous outputs for metric validation: two closed-source models accessed via API and one smaller open-source general instruction-tuned model. Group labels throughout the paper describe source availability and domain specialization; they do not imply capability tiers or architectural parity. We did not include a medical-specific anchor because this study evaluates metric-human alignment rather than model-family performance.

The human-judgment study was independent of benchmark scoring. Three resident physicians rated each extraction against the gold plan on five dimensions: field completeness, factual accuracy, hallucination absence, field placement, and overall clinical utility. They were blinded to model identity, chrF, MoverScore, leaderboard rankings, and aggregate automatic scores. Ratings were averaged into a single human clinical score per instance; inter-rater reliability was $\kappa = 0.76$. The validation subset was used only to estimate metric–human correlations, not to tune prompts, select models or outputs, set weights or thresholds, calibrate normalization constants, or alter the schema. The final leaderboard was scored once using the fixed protocol; thus, validation evaluates metrics, whereas the leaderboard evaluates models.

Candidate metrics and leakage control. We evaluated six metrics: ROUGE-L, ROUGE-1, mean-pooled sentence cosine similarity using `clinical-sts-S-PubMedBert-MS-MARCO`, BERTScore with the same encoder, chrF ($\beta = 1$, $n = 6$), and our clinical adaptation of MoverScore using IDF-weighted optimal transport. The metric set, field-level aggregation rule, and reporting protocol were frozen before scoring the 36-model benchmark. All metrics were computed at the field level and averaged using the same protocol as the main benchmark. No model-specific parameters, field weights, leaderboard-dependent normalization, thresholds, or post-hoc ensembles were fitted. MoverScore was selected because it had the strongest correlation with blinded physician judgments, not because it favored any anchor model.

Results. Table 1 reports Pearson (r) and Spearman (ρ) correlations with physician clinical scores. ROUGE metrics align weakest with physician judgment ($r = 0.31$–0.34), consistent with failures on short clinical strings and abbreviation mismatches (e.g., ROUGE-L $= 0.00$ for *"CBC"* vs. *"complete blood count"*). Sentence cosine similarity performs better on free-text fields ($\rho = 0.49$) than structured fields ($\rho = 0.35$), because mean-pooled embeddings can over-credit clinically distinct but nearby items such as *warfarin* and *metformin*. chrF is the strongest syntactic metric ($r = 0.61$, $\rho = 0.59$), benefiting from character n-gram matching over abbreviations, partial matches, and morphological variants. MoverScore achieves the highest overall correlation ($r = 0.71$, $\rho = 0.69$) and the most consistent performance across field types; despite using the same encoder as BERTScore ($r = 0.52$), its optimal-transport alignment over token distributions is more effective than greedy token matching, especially for structured fields.

Robustness to anchor-model choice. To test whether the metric conclusion depended on the particular anchor models used in the physician-rating study, we repeated the metric–human correlation analysis under leave-one-model-out resampling. For each run, we removed all outputs from one anchor model and recomputed the correlation between each automatic metric and the physician clinical score on the remaining 200 instances. The ordering of the two strongest metrics was unchanged: MoverScore remained the highest-correlating metric in every leave-one-model-out run, and chrF remained the strongest surface-sensitive metric. This

Metric	Overall		Structured		Free-text	
	r	ρ	r	ρ	r	ρ
ROUGE-L	0.31±.09	0.29±.09	0.33±.09	0.31±.09	0.22±.09	0.20±.09
ROUGE-1	0.34±.09	0.32±.09	0.36±.09	0.34±.09	0.24±.09	0.22±.09
Cosine (sent.)	0.41±.08	0.39±.08	0.37±.08	0.35±.09	0.51±.07	0.49±.07
BERTScore	0.52±.07	0.50±.07	0.54±.07	0.52±.07	0.48±.08	0.46±.08
chrF	0.61±.06	0.59±.06	0.63±.06	0.61±.06	0.54±.07	0.52±.07
MoverScore	**0.71**±.05	**0.69**±.05	**0.72**±.05	**0.70**±.05	**0.68**±.05	**0.66**±.06

Table 1: Pearson (r) and Spearman (ρ) correlation of six metrics with physician clinical scores on 300 (case, model) instances ($p < 0.01$, two-tailed). MoverScore outperforms chrF significantly (Steiger $Z = 3.56$, $p = 0.0004$); chrF outperforms BERTScore significantly (Steiger $Z = 2.83$, $p = 0.005$). Structured fields: `labs`, `imaging`, `medical`, `monitoring`, `referral`, `follow_up`. Free-text fields: `patient_education`, `when_to_seek_medical_care`.

analysis addresses the main circularity concern: the metric choice is not driven by the presence of a single high-performing frontier model, nor by the inclusion of the lower-performing open-source model. The validation study therefore supports the metric protocol as a model-agnostic evaluation choice rather than as a post-hoc choice tailored to the final leaderboard.

4.2 Encoder Selection for MoverScore

MoverScore requires a token-level encoder whose contextual embeddings accurately reflect clinical term similarity. We benchmarked 24 encoders across four categories (clinical STS fine-tuned, biomedical pretrained, general-domain, and large instruction-tuned); full benchmark design and results appear in Appendix F.2.

Selected encoder. We select *clinical-sts-S-PubMedBert-MS-MARCO* because it provides the best trade-off between separating non-equivalent clinical terms and preserving high similarity for equivalent clinical paraphrases. On our 300-pair clinical term benchmark, it obtains Spearman = 0.848 and AUROC-EN = 0.960. PubMedBERT-family encoders dominate the top rankings, confirming that biomedical pretraining matters more than parameter scale: large general encoders such as *e5-mistral-7b-instruct* and *bge-m3* score substantially lower.

Despite the highest AUROC-EN achieved by the raw `pubmedbert-base-embeddings`, we do not select it because its token-level optimal-transport scores are less well calibrated for equivalent paraphrases. Appendix F.2 compares the selected encoder with `pubmedbert-base-embeddings` on equivalent, related/confusable, and unrelated clinical term pairs. The raw checkpoint assigns lower similarity to equivalent pairs and produces a larger equivalent–related separation than the STS-tuned encoder, which would over-penalize clinically valid surface variation such as abbreviation expansion and paraphrase. We therefore use STS-tuned PubMedBERT encoder as the MoverScore backbone. Table 11 gives results for all 24 encoders.

5 Results and Analysis

We evaluate 36 models on CAPE after fixing the metric set, aggregation protocol, and metric-validation design. We report MoverScore for semantic equivalence, chrF for schema-compatible surface fidelity, and $\Delta = \text{MoverScore} - \text{chrF}$ as a diagnostic gap: large positive values indicate outputs that are semantically close but surface- or schema-divergent.

Overall, the results show that CAPE primarily tests *plan grounding* rather than medical knowledge alone. The strongest general instruction-following models achieve the highest scores, while medical specialization by itself does not guarantee strong schema-grounded extraction. Across model families, the main remaining failures involve temporal scope, schema placement, controlled vocabulary, hallucination control, and subfield completeness.

5.1 Main Leaderboard

Table 2 reports a representative subset of the 36-model leaderboard; the full leaderboard with parse rates and error counts is provided in Appendix G.2.

The top systems are closed-source general-purpose models: GPT-o3 obtains the highest MoverScore (0.796), followed closely by Gemini-2.5-Flash (0.792), Gemini-3-Flash-Preview (0.765), GPT-5-Mini (0.765), and Grok-4.20 (0.760). However, even the best models remain below perfect performance, mainly because of missing subfields, unsupported attribute completion, and schema-boundary errors.

Open-source general models form a competitive second group. Qwen3-VL-32B (0.714) and Qwen2.5-32B (0.700) approach several closed-source systems, while smaller open-source models vary substantially: Qwen2.5-14B reaches 0.683, whereas Llama-3.1-8B-Instruct reaches 0.456. This variation suggests that CAPE is sensitive not only to parameter scale, but also to instruc-

Model	Category	MoverScore ↑	chrF ↑	Δ
GPT-o3 (OpenAI, 2025b)	Closed-source	0.796±0.014	0.730±0.016	+0.066±0.009
Gemini-2.5-Flash (Comanici et al., 2025)	Closed-source	0.792±0.014	0.709±0.017	+0.083±0.009
Gemini-3-Flash-Preview (Google DeepMind, 2025a)	Closed-source	0.765±0.016	0.698±0.017	+0.067±0.009
GPT-5-Mini (OpenAI, 2025a)	Closed-source	0.765±0.015	0.706±0.016	+0.059±0.009
Grok-4.20 (xAI, 2025)	Closed-source	0.760±0.016	0.698±0.017	+0.062±0.009
Kimi-k2.6 (Team et al., 2025)	Open-source	0.725±0.017	0.644±0.018	+0.081±0.011
Llama-4-Maverick-17B (Meta, 2025)	Open-source	0.706±0.018	0.620±0.019	+0.086±0.011
Command-R-Plus (Cohere, 2024)	Open-source	0.680±0.019	0.581±0.020	+0.099±0.012
Mistral-Small-2603 (Mistral AI, 2025)	Open-source	0.678±0.019	0.589±0.020	+0.089±0.012
GLM-5.1 (Zhipu AI, 2025)	Open-source	0.648±0.021	0.583±0.021	+0.066±0.011
Qwen3-VL-32B-Instruct (Qwen Team, Alibaba Cloud, 2025)	Open-source	0.714±0.018	0.629±0.019	+0.085±0.011
Qwen2.5-32B (Yang et al., 2024)	Open-source	0.700±0.019	0.599±0.020	+0.100±0.013
Llama-3.1-70B-Instruct (Grattafiori et al., 2024)	Open-source	0.686±0.019	0.601±0.020	+0.086±0.012
Gemma-3-27b-it (Google DeepMind, 2025b)	Open-source	0.646±0.021	0.576±0.021	+0.070±0.012
Qwen2.5-14B (Yang et al., 2024)	Open-source	0.683±0.019	0.589±0.020	+0.094±0.012
Qwen3-VL-8B-Instruct (Qwen Team, Alibaba Cloud, 2025)	Open-source	0.634±0.021	0.556±0.022	+0.078±0.013
Qwen2.5-7B-Instruct (Yang et al., 2024)	Open-source	0.573±0.023	0.485±0.023	+0.089±0.013
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Open-source	0.456±0.025	0.350±0.024	+0.106±0.015
OpenBioLLM-70B (Goel et al., 2024)	Medical-specific	0.662±0.021	0.586±0.021	+0.076±0.012
MedGemma-27b-it (Sellergren et al., 2025)	Medical-specific	0.658±0.020	0.608±0.020	+0.051±0.010
HuatuoGPT-3-8B (Zhang et al., 2023)	Medical-specific	0.525±0.024	0.414±0.024	+0.111±0.015
MedGemma-4b-it (Sellergren et al., 2025)	Medical-specific	0.548±0.023	0.454±0.023	+0.095±0.013

Table 2: **Representative CAPE leaderboard.** The table reports a representative subset of evaluated models. Δ denotes MoverScore−chrF, with intervals computed from paired per-case differences (MoverScore − chrF). The full leaderboard with error counts appears in Appendix G.2. Model versions are listed in Table 6.

tion tuning and structured-output reliability.

Medical-specific models do not dominate the benchmark. OpenBioLLM-70B (0.662) and MedGemma-27b-it (0.658) are competitive with some open-source general models, but remain below the strongest closed-source and open-source general systems. Smaller medical models perform worse, including MedGemma-4b-it (0.548) and HuatuoGPT-3-8B (0.525). This supports the conclusion that domain specialization alone is insufficient for faithful structured plan extraction.

5.2 Field-Level Difficulty

Figure 3 shows that field difficulty is highly non-uniform. The largest semantic–surface gap occurs in `follow_up` ($\Delta = +0.229$), followed by `referral` and `when_to_seek_care`. These fields often require models to separate compact dialogue phrases into timing, purpose, specialty, and safety-netting subfields.

In contrast, `labs` and `monitoring` have near-zero Δ values, indicating that models usually either recover these structured items in near-reference form or miss them.

The `conservative` field remains difficult across model families because it encodes a clinically meaningful schema boundary: clinician-directed conservative interventions, such as splinting or phlebotomy, must be separated from patient-directed lifestyle advice, such as rest, diet, or activity modification. This is also one of the lower-agreement fields in the human reliability analysis, suggesting that the difficulty reflects an inherently ambiguous but important boundary rather than a purely model-specific failure.

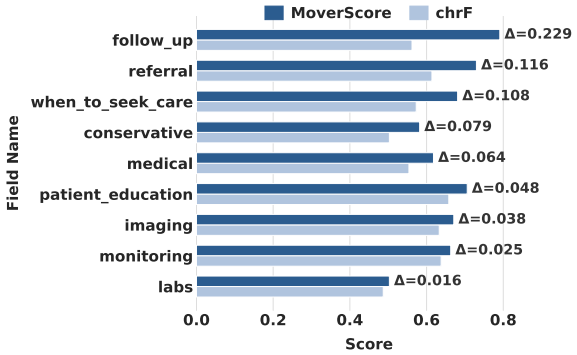

Figure 3: Per-field chrF and MoverScore averaged across all models. Delta annotations indicate the semantic–surface gap.

5.3 Error Analysis

The dominant errors are missing plan content, hallucinated plan content, subfield-boundary violations, and controlled-vocabulary deviations. Smaller schema-unstable models often omit ac-

tionable aims for follow-up or referral. Weaker medical-specific models more often add clinically plausible but unsupported plans, such as unstated tests, referrals, monitoring intervals, or medication attributes. High-scoring models more often recover the right content but place it in the wrong schema location, for example confusing `conservative_method` with `lifestyle_habit_modifications` or merging imaging modality with anatomical site. Controlled-vocabulary errors also occur when models output values such as "new" or "prescribed" instead of the allowed `treatment_status` value `start`.

These patterns explain why MoverScore and chrF must be reported together. MoverScore rewards semantic overlap and valid paraphrase, but can over-credit schema-invalid outputs. chrF rewards schema-faithful surface forms, but can over-penalize harmless paraphrases. Their gap therefore exposes cases where a model captures the approximate clinical intent but fails to express it in the required structured form. Detailed examples of temporal qualifier truncation, follow-up aim absorption, register compression, and schema misplacement are provided in Appendix G.

5.4 Model Families and Scale

At matched parameter scales, general instruction-tuned models outperform medical-specific models on average. At the 7–11B scale, general models average 0.502 MoverScore compared with 0.288 for medical models; at the 70–90B scale, the gap narrows but remains, 0.650 versus 0.569 (Table 3). Scaling improves extraction, but does not eliminate CAPE's hardest failures: temporal scoping, field placement, and schema-boundary discrimination remain bottlenecks, especially for `conservative`, `follow_up`, and `referral`.

Scale	General mean	Medical mean	Δ
7–11B	0.502±0.022	0.288±0.025	+0.214±0.034
70–90B	0.650±0.020	0.569±0.023	+0.081±0.029

Table 3: Medical vs. general models at matched parameter scales. Scores are reported as mean MoverScore ± 95% bootstrap confidence-interval half-width, computed by non-parametric bootstrap over the 500 CAPE test cases with 10,000 resamples. Δ denotes the paired bootstrap estimate of General−Medical after averaging models within each scale group.

Some model families show local non-monotonicity: Qwen2.5-14B outperforms larger Qwen2.5 variants overall, and Qwen3-VL-8B slightly underperforms Qwen3-VL-4B before the 32B model improves again. This suggests that CAPE is sensitive not only to scale, but also to instruction tuning and structured-output reliability.

6 Generalization and Fine-tuning

The results in §5 are based on CAPE. We therefore examine whether the main findings show preliminary transfer to real clinical conversations and whether task-specific supervision improves schema-grounded extraction.

6.1 Generalization to Real Clinical Dialogues

CAPE is built from *MedSynth* (Mianroodi et al., 2025), a publicly available corpus of synthetic doctor-patient dialogues whose diagnoses, management decisions, and conversational structure originate from curated medical resources, while the dialogue surface form is synthetically generated. This choice reflects practical constraints: real clinical conversations are expensive to collect, legally restricted to de-identify, and difficult to annotate at scale with broad ICD-10 coverage. Synthetic dialogue enables controlled coverage of 21 ICD-10 chapters and 1,900 unique diagnostic codes, which would be infeasible with comparable real-encounter annotation cost.

To test whether findings from synthetic dialogue transfer directionally to real conversations, we conduct a limited external check on PRIMOCK57 (Korfiatis et al., 2022), a corpus of 57 real UK primary-care consultations. Seven models spanning closed-source systems (*gemini-3-flash-preview*, *gpt-5-mini*), (*command-r-plus*, *deepseek-v4-flash*), and smaller or medical-specific models (*Llama-3.1-8B-Instruct*, *HuatuoGPT-3-8B*, *MedGemma-4b-it*) were evaluated using the same schema, annotation guidelines, and qualification threshold as CAPE. Because this check covers only seven models and 57 consultations, we treat it as directional evidence rather than a definitive generalization test.

Model rankings are largely preserved on PRIMOCK57: *gemini-3-flash-preview* and *gpt-5-mini* obtain the highest MoverScores (0.715 and 0.714), while *MedGemma-4b-it* and *Llama-3.1-8B-Instruct* remain near the bottom (0.498 and 0.419). The Spearman rank correlation between CAPE and PRIMOCK57 MoverScores across the seven models is $\rho = 0.991$ (exact two-sided permutation test, $p = 0.0008$, over $7! = 5{,}040$ rank assign-

Base model	Objective	chrF	Mover	Miss. item	Miss. subf.	Hall. item	Mispl.
Llama-3.1-8B	Base	0.350	0.456	1706	1340	1791	148
Llama-3.1-8B	DPO	0.461	0.517	1006	1332	1397	59
Llama-3.1-8B	IPO	0.566	0.624	1209	2147	818	141
Llama-3.1-8B	ORPO	0.577	0.614	1572	1301	641	162
Llama-3.1-8B	SFT	0.644	0.680	1661	1481	570	71
Qwen2.5-7B	Base	0.485	0.573	1752	1088	1043	60
Qwen2.5-7B	DPO	0.545	0.602	1034	1082	834	22
Qwen2.5-7B	IPO	0.602	0.652	1242	1743	527	57
Qwen2.5-7B	ORPO	0.608	0.647	1614	1056	434	66
Qwen2.5-7B	SFT	0.644	0.678	1877	1276	396	68

Table 4: **Effect of task-specific fine-tuning on CAPE.** chrF and Mover denote overall syntactic and semantic scores. SFT achieves the highest MoverScore for both base models; ORPO gives the strongest hallucination reduction among preference objectives.

ments). This near-identical ranking suggests that the main model-comparison conclusions are unlikely to be artifacts of synthetic dialogue alone.

Field-level patterns are also directionally consistent: `labs` separates stronger from weaker models, `follow_up` again shows a large MoverScore–chrF gap due to temporal qualifier truncation and aim absorption, and `conservative` remains unstable, consistent with the clinician-directed versus patient-directed boundary discussed in §5.2. One exception is `deepseek-v4-flash`, which drops sharply on `referral` (MoverScore 0.114) and `follow_up` (MoverScore 0.307) in PRIMOCK57. Inspection indicates near-complete omission of these fields rather than surface-form mismatch, suggesting sensitivity to real-consultation lexical variability when extracting care-continuity plans. Such omissions are clinically important because they remove actionable follow-up and referral decisions.

Overall, CAPE's model-ranking and field-difficulty patterns do not appear to be artifacts of synthetic dialogue, although larger real-dialogue evaluations across more consultations and models remain necessary. Full per-field results are provided in Appendix G.9.

6.2 Effect of Task-Specific Fine-tuning

We fine-tune *Llama-3.1-8B-Instruct* and *Qwen2.5-7B-Instruct* with LoRA under supervised fine-tuning (SFT), direct preference optimization (DPO) (Rafailov et al., 2023), identity preference optimization (IPO) (Azar et al., 2024), and odds-ratio preference optimization (ORPO) (Hong et al., 2024). All variants use the same 500-case test set, prompt, and decoding configuration as the main benchmark; hyperparameters and full per-field results are in Appendix H.

Table 4 shows that SFT gives the largest gains: Llama-3.1-8B improves from 0.456 to 0.680 MoverScore, and Qwen2.5-7B from 0.573 to 0.678. DPO improves both base models but remains below SFT; IPO yields larger semantic gains than DPO but increases missing subfields; and ORPO gives the strongest hallucination reduction among preference objectives.

Comparison	Δ
Llama-3.1-8B-SFT vs. OpenBioLLM-70B	$+4.5_{\pm 2.4}$
Qwen2.5-7B-SFT vs. OpenBioLLM-70B	$+3.9_{\pm 2.4}$
Llama-3.1-8B-SFT vs. MedGemma-27B	$+5.2_{\pm 2.3}$
Qwen2.5-7B-SFT vs. MedGemma-27B	$+4.7_{\pm 2.3}$

Table 5: Paired MoverScore gains, in percentage points, of fine-tuned 7–8B models over medical-specific baselines.

SFT also substantially reduces hallucination: Llama-3.1-8B-SFT reduces hallucinated items from 1,791 to 570 (-68.2%), and Qwen2.5-7B-SFT from 1,043 to 396 (-62.0%). Paired comparisons against the strongest medical-specific baselines show that both fine-tuned 7–8B models outperform OpenBioLLM-70B and MedGemma-27b-it by $+3.9$ to $+5.2$ MoverScore points (Table 5), indicating that task-specific supervision can outweigh medical-domain specialization for schema-grounded plan extraction on CAPE.

The lowest-gain field after fine-tuning is again `conservative`, consistent with human-agreement analysis: this field encodes a clinically meaningful but partly ambiguous boundary between clinician-directed intervention and patient-directed behaviour. Fine-tuning improves hallucination control more reliably than it resolves this schema-boundary ambiguity.

7 Conclusion

CAPE introduces a schema-grounded benchmark for extracting clinical action plans from doctor-patient dialogue. Evaluation across 36 models shows that reliable extraction depends on schema adherence, temporal reasoning, and hallucination control, rather than medical-domain specialization alone. Task-specific supervision reduces hallucinations by over 60% and allows 7–8B models to surpass larger medical-specific baselines, while a small PRIMOCK57 replication suggests preliminary transfer to real dialogue. CAPE offers a leaderboard and diagnostic testbed for clinically reliable structured extraction.

Limitations

CAPE is built primarily from synthetic doctor-patient dialogues, which enables broad coverage and controlled annotation but may not capture all variability of real clinical encounters. We partially address this through evaluation on PRIMOCK57, but the real-dialogue sample is small and should be viewed as preliminary evidence of transfer rather than a conclusive test of ranking preservation. Most production cases were initially annotated by one qualified physician; however, the 500-case test set was additionally reviewed by five physicians, with quality control through training, qualification testing, interface validation, test-set review, and adjudication of flagged cases. CAPE evaluates faithful extraction of stated plans, not the clinical correctness of those plans or downstream patient outcomes.

Ethics Statement

CAPE is intended for research on clinical information extraction and should not be used as a clinical decision-support system without expert oversight. The benchmark focuses on extracting clinician-stated plans and explicitly penalizes unsupported additions, omissions, and schema errors that could be harmful in clinical settings. The primary dataset is based on synthetic dialogues, reducing privacy risk, and the real-dialogue replication uses an existing de-identified research corpus. Because model errors may affect patient safety if deployed, CAPE results should be interpreted as diagnostic evidence of extraction behavior rather than proof of clinical readiness.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. 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.

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.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen-Yildiz, Fei Xia, and Thomas Lin. 2025. Medec: A benchmark for medical error detection and correction in clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550.

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2024. A general theoretical paradigm to understand learning from human feedback. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Cohere. 2024. Command R+: Enterprise-grade language model. https://cohere.com/blog/command-r-plus-microsoft-azure. Version evaluated: `cohere/command-r-plus`.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen K Wong, Graham Wills, Elliot First, Frank J Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2024. Evaluation of large language models for summarization tasks in the medical domain: A narrative review. *arXiv preprint arXiv:2409.18170*.

Ankit Goel, Usman Naseem, Subhajit Saha, Vu Minh Thang, Gholamreza Haffari, Priya Krishnaswamy, Hyeon-Gu Kang, Hyoung-Gyu Kim, Vivek Kashyap, and 1 others. 2024. OpenBioLLM: Open-source biomedical large language models. *arXiv preprint*. Versions evaluated: `aaditya/OpenBioLLM-70B`, `aaditya/Llama3-OpenBioLLM-8B`. HuggingFace: https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Google DeepMind. 2025a. Gemini 3 flash preview. https://deepmind.google/technologies/gemini/. Version evaluated: `google/gemini-3-flash-preview` (May 2025).

Google DeepMind. 2025b. Gemma 3: Open models. https://ai.google.dev/gemma/docs/gemma3. Versions evaluated: `google/gemma-3-4b-it`, `google/gemma-3-12b-it`, `google/gemma-3-27b-it`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

HL7 International. 2019a. HL7 FHIR Release 4: CarePlan resource. HL7 FHIR Release 4. Accessed 2024.

HL7 International. 2019b. HL7 FHIR Release 4: MedicationRequest resource. HL7 FHIR Release 4. Accessed 2024.

HL7 International. 2019c. HL7 FHIR Release 4: ServiceRequest resource. HL7 FHIR Release 4. Accessed 2024.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Meta. 2025. Llama 4 maverick 17b 128e instruct. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct. Model card and checkpoint. Released April 5, 2025. Accessed: 2026-05-25.

Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. Medsynth: Realistic, synthetic medical dialogue-note pairs. *arXiv preprint arXiv:2508.01401*.

Mistral AI. 2025. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1. Version evaluated: mistralai/mistral-small-2603.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2025a. GPT-5 mini: Efficient frontier language model. https://openai.com/index/gpt-5-mini-system-card/. Version evaluated: openai/gpt-5-mini (May 2025).

OpenAI. 2025b. o3: Reasoning model. https://openai.com/o3. Version evaluated: openai/o3 (May 2025).

World Health Organisation. 2020. Who package of essential noncommunicable (pen) disease interventions for primary health care.

Vivekanand Podder, Valerie Lew, and Sassan Ghassemzadeh. 2023. *SOAP Notes*. StatPearls Publishing, Treasure Island, FL. NCBI Bookshelf NBK482263. Updated 2023.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.

Qwen Team, Alibaba Cloud. 2025. Qwen3-VL: Visual language models. https://qwenlm.github.io/blog/qwen3-vl/. Versions evaluated: Qwen/Qwen3-VL-2B-Instruct, Qwen/Qwen3-VL-4B-Instruct, Qwen/Qwen3-VL-8B-Instruct, Qwen/Qwen3-VL-32B-Instruct.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy A Miller, Chen Lin, Guergana K Savova, and 1 others. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:9871 – 9889.

Lawrence L Weed and 1 others. 1968. Medical records that guide and teach. *N Engl J Med*, 278(11):593–600.

World Health Organization. 2014. WHO handbook for guideline development. Technical report, World Health Organization, Geneva. ISBN: 978-92-4-154896-0.

xAI. 2025. Grok-4. `https://x.ai/grok`. Version evaluated: `x-ai/grok-4.20` (May 2025).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578.

Zhipu AI. 2025. GLM-5.1: General language model. `https://zhipuai.cn`. Version evaluated: `z-ai/glm-5.1` (May 2025).

A Model Versions and API Identifiers

Table 6 lists the exact API identifiers and HuggingFace model IDs for all models evaluated on CAPE. Proprietary models were accessed via OpenRouter or the provider's native API and are identified by the string used at inference time. Open-source models are identified by their published HuggingFace checkpoint. This information is provided to support reproducibility; provider-hosted general-purpose models may be silently updated by their providers, so re-evaluation results may differ from those reported in Table 2.

Model (display name)	Exact identifier	Access	Access date
Closed-Source Models			
o3	openai/o3	Open Router	2026-04-10 UTC
Gemini 2.5 Flash	google/gemini-2.5-flash	Open Router	2026-04-10 UTC
Gemini 3 Flash Preview	google/gemini-3-flash-preview	Open Router	2026-04-11 UTC
GPT-5 mini	openai/gpt-5-mini	Open Router	2026-04-11 UTC
Grok-4.20	x-ai/grok-4.20	Open Router	2026-04-12 UTC
Kimi k2.6	moonshotai/kimi-k2.6	Open Router	2026-04-12 UTC
Gemini 3.1 Flash Lite	google/gemini-3.1-flash-lite	Open Router	2026-04-13 UTC
Open-Source, API Access			
Llama-4-Maverick-17B	meta-llama/Llama-4-Maverick-17B	Open Router	2026-04-13 UTC
Command R+	cohere/command-r-plus	Open Router	2026-04-14 UTC
Mistral Small 2603	mistralai/mistral-small-2603	Open Router	2026-04-14 UTC
MiniMax-M2.7	minimax/minimax-m2.7	Open Router	2026-04-15 UTC
GLM-5.1	z-ai/glm-5.1	Open Router	2026-04-15 UTC
DeepSeek-V4-Flash	deepseek/deepseek-v4-flash	Open Router	2026-04-20 UTC
Open-Source Large ($\geq$27B)			
Qwen3-VL-32B-Instruct	Qwen/Qwen3-VL-32B-Instruct	HuggingFace	2026-04-16 UTC
Qwen2.5-32B	Qwen/Qwen2.5-32B	HuggingFace	2026-04-16 UTC
Llama-3.1-70B-Instruct	meta-llama/Llama-3.1-70B-Instruct	HuggingFace	2026-04-16 UTC
Llama-3.2-90B-Vision	meta-llama/Llama-3.2-90B-Vision	HuggingFace	2026-04-17 UTC
Qwen2.5-72B	Qwen/Qwen2.5-72B	HuggingFace	2026-04-17 UTC
Gemma-3-27b-it	google/gemma-3-27b-it	HuggingFace	2026-04-17 UTC
Open-Source Medium (7B–14B)			
Qwen2.5-14B	Qwen/Qwen2.5-14B	HuggingFace	2026-04-18 UTC
Qwen3-VL-8B-Instruct	Qwen/Qwen3-VL-8B-Instruct	HuggingFace	2026-04-18 UTC
Gemma-3-12b-it	google/gemma-3-12b-it	HuggingFace	2026-04-18 UTC
Qwen2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct	HuggingFace	2026-04-18 UTC
Llama-3.2-11B-Vision	meta-llama/Llama-3.2-11B-Vision	HuggingFace	2026-04-20 UTC
Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct	HuggingFace	2026-04-20 UTC
Open-Source Small (<7B)			
Qwen3-VL-4B-Instruct	Qwen/Qwen3-VL-4B-Instruct	HuggingFace	2026-04-18 UTC
Gemma-3-4b-it	google/gemma-3-4b-it	HuggingFace	2026-04-20 UTC
Qwen3-VL-2B-Instruct	Qwen/Qwen3-VL-2B-Instruct	HuggingFace	2026-04-20 UTC
Llama-3.2-3B-Instruct	meta-llama/Llama-3.2-3B-Instruct	HuggingFace	2026-04-20 UTC
Medical-Specific Models			
OpenBioLLM-70B	aaditya/OpenBioLLM-70B	HuggingFace	2026-04-19 UTC
MedGemma-27b-it	google/medgemma-27b-it	HuggingFace	2026-04-19 UTC
HuatuoGPT-3-32B	FreedomIntelligence/HuatuoGPT-3-32B	HuggingFace	2026-04-19 UTC
MedGemma-4b-it	google/MedGemma-4b-it	HuggingFace	2026-04-19 UTC
HuatuoGPT-3-8B	FreedomIntelligence/HuatuoGPT-3-8B	HuggingFace	2026-04-19 UTC
Llama3-OpenBioLLM-8B	aaditya/Llama3-OpenBioLLM-8B	HuggingFace	2026-04-20 UTC

Table 6: Model identifiers and access dates for the evaluated models. *Access* indicates the inference backend used: OR = OpenRouter, API = provider native API, HF = HuggingFace. API-accessed systems are reported using the endpoint identifier used at inference time; open-weight systems are reported using the HuggingFace checkpoint identifier. Because hosted API endpoints and HuggingFace repositories may change over time, the reported results should be interpreted with respect to the endpoint or checkpoint state available on the recorded April 2026 access date.

B Dataset Statistics

Table 7 summarizes the construction, coverage, split, and annotation statistics for CAPE.

Statistic	Value	Notes
Initial candidate pool	1,905	Prior to quality exclusions
Excluded cases	5	Source dialogue quality issues
Total dialogues	1,900	Final cases after replacements
Unique ICD-10 codes	1,900	One code per dialogue
ICD-10 chapters	21 / 22	All except Chapter U
Training split	1,400	73.7%; stratified by ICD-10 chapter
Test split	500	26.3%; stratified by ICD-10 chapter
Avg. dialogue turns	18.4 ± 5.1	Doctor and patient turns combined
Test-set review	5 physicians	All 500 test cases reviewed before model evaluation
Avg. tokens/dialogue	$1{,}847 \pm 412$	cl100k tokenizer; median 1,764
Avg. medications/case	2.3 ± 1.1	Non-empty `medical[]` only
Avg. lab tests/case	2.7 ± 1.4	Non-empty `labs[]` only
Annotators	10 residents	Supervised by 2 attending clinicians
Qualification threshold	≥ 0.80 F1	Per-field syntactic F1; all fields required
Source dataset	MedSynth	Synthetic doctor-patient dialogues

Table 7: CAPE dataset statistics. The initial candidate pool of 1,905 dialogues yielded 1,900 final cases after quality exclusions and replacements. Train/test splits are stratified by ICD-10 chapter.

C Schema Details

This appendix provides the schema details supporting §3.1. The main text describes the schema at a high level; here we provide the clinical rationale, standards grounding, and additional field-level details used during annotation.

C.1 Clinical and Standards Grounding

The CAPE schema was designed to represent a treatment plan as a set of clinically actionable commitments rather than as a prose summary. Its design follows three complementary sources of grounding.

Clinical documentation logic. The schema operationalises the Plan component of SOAP-style documentation (Weed et al., 1968; Podder et al., 2023). In particular, it separates distinctions that are often compressed in clinical prose but matter for downstream actionability: medication decisions versus non-pharmacological interventions, clinician-directed conservative procedures versus patient-directed lifestyle modifications, follow-up timing versus follow-up purpose, and referral target versus referral aim.

Standards alignment. The schema is structurally aligned with common FHIR resource families while remaining intentionally task-oriented. Investigations correspond conceptually to ServiceRequest-like actions; medication plans to MedicationRequest-like actions; monitoring and ongoing management to CarePlan-like structures; follow-up to appointment-like scheduling; and patient education or safety-netting advice to patient-directed communication. We use these standards as conceptual anchors rather than claiming one-to-one implementation equivalence. Field-to-standard mappings are reported in Table 8.

Guideline-style action structure. Clinical guidelines commonly specify what action should be taken, for whom, under what conditions, with what monitoring, and with what follow-up (World Health Organization, 2014). WHO PEN similarly organises primary-care management around actionable interventions, monitoring, referral, and follow-up (Organisation, 2020). CAPE mirrors this action-oriented structure by separating investigations, monitoring, treatment, referral, follow-up, patient education, and return precautions.

C.2 Top-Level Domains

The schema contains seven top-level domains:

- INVESTIGATIONS: tests to be ordered, pending, or planned, including laboratory tests, imaging, nerve/muscle conduction studies, and tissue sampling.
- MONITORING: ongoing surveillance of clinical parameters, symptoms, or disease markers, distinct from one-time investigations.
- TREATMENT: pharmacological and non-pharmacological management, including medication decisions, conservative procedures, lifestyle modifications, equipment, procedures, and other therapeutic actions.

CAPE schema field	Standards anchor	Rationale for mapping
`labs`	FHIR ServiceRequest-like action	Laboratory tests ordered, planned, or pending as part of diagnostic or reassessment workup.
`imaging`	FHIR ServiceRequest-like action	Imaging studies requested for diagnosis, staging, reassessment, or monitoring of clinical status.
`monitoring`	FHIR CarePlan-like activity	Longitudinal surveillance of symptoms, parameters, disease markers, or treatment response.
`medical`	FHIR MedicationRequest-like action	Medication initiation, continuation, discontinuation, or adjustment, including drug-specific attributes.
`conservative_method`	FHIR CarePlan-like activity	Clinician-directed non-pharmacological management such as splinting, wound care, compression, or phlebotomy.
`lifestyle_habit_modifications`	FHIR CarePlan-like activity / patient instruction	Patient-directed behavioural recommendations such as diet, exercise, rest, hydration, or smoking cessation.
`follow_up`	Appointment-like scheduling / CarePlan review	Planned reassessment, decomposed into timing and clinical purpose.
`referral`	FHIR ServiceRequest-like action	Request for evaluation or management by another specialty or clinician, including referral aim.
`patient_education`	Patient-directed communication	Substantive counselling or explanatory advice explicitly provided to the patient.
`when_to_seek_medical_care`	Patient-directed communication / safety-netting	Return precautions or urgent-care instructions explicitly stated in the dialogue.

Table 8: Conceptual mapping between CAPE schema fields and standards-inspired clinical information structures. These mappings are used as conceptual anchors rather than claims of one-to-one implementation equivalence.

- FOLLOW-UP: planned reassessment, decomposed into timing and purpose.
- REFERRAL: referral target and clinical aim.
- PATIENT EDUCATION: substantive patient-facing counselling explicitly stated in the dialogue.
- WHEN TO SEEK MEDICAL CARE: safety-netting or return precautions explicitly stated in the dialogue.

These domains were selected to cover the management actions most commonly expressed in outpatient doctor-patient conversations while preserving clinically meaningful field boundaries.

C.3 Key Schema Distinctions

Several schema distinctions were emphasised during annotator training because they are clinically meaningful and frequently confusable.

Investigations vs. monitoring. A one-time test ordered for diagnostic or reassessment purposes is annotated under INVESTIGATIONS; repeated or longitudinal tracking is annotated under MONITORING. For example, "check HbA1c today" is an investigation, whereas "monitor HbA1c every three months" is a monitoring plan.

Conservative method vs. lifestyle modification. `conservative_method` captures clinician-directed or procedural non-pharmacological management, such as wound dressing, splinting, compression, or phlebotomy. In contrast, `lifestyle_habit_modifications` captures patient-directed behaviour, such as diet changes, rest, hydration, exercise, or smoking cessation. This boundary is clinically important because it separates interventions performed or arranged by the clinician from behaviours the patient is instructed to carry out.

Follow-up time vs. follow-up aim. `scheduled_follow_up_time` records when the patient should return, whereas `aim_of_follow_up` records why the follow-up is needed. For example, "return in six weeks to assess treatment response" contains both a time and an aim. Separating these fields allows the benchmark to detect aim absorption, where a model recovers the timing but loses the clinical purpose.

Referral target vs. referral aim. `specialty_or_doctor` records the referral destination, while `aim_of_referral` records the clinical purpose. The same specialty may be consulted for diagnosis, medication optimisation, procedural management, or risk stratification, so the specialty alone is not sufficient to represent the plan.

Controlled medication status. Medication actions use a controlled `treatment_status` vocabulary: `start`, `continue`, `stop`, and `adjust`. This prevents models from receiving full credit for extracting a medication name while failing to distinguish whether it is newly initiated, continued, discontinued, or modified.

C.4 Null and Empty-Field Policy

CAPE uses explicit missing-value conventions. If a list-valued field is absent from the dialogue, it is annotated as an empty list. If an item is present but a specific attribute is not stated, the missing attribute is annotated as `null`. Annotators were instructed not to infer clinically plausible but unverbalised information.

This policy is central to the benchmark design. It allows evaluation to distinguish between omitted plan content, hallucinated plan items, unsupported attribute completion, and correct recognition that information was not stated.

C.5 Subfield-Level Population

Structured list fields contain both dense and sparse extraction targets. For example, medication name and treatment status are frequently stated, whereas medication route, dosage form, and drug class are often not verbalised and are therefore annotated as `null`. Similarly, follow-up and referral aims are usually recoverable when the corresponding item is present, while some investigation attributes are often implicit in natural dialogue.

These patterns motivate the evaluation protocol in §4, which tracks item-level and subfield-level omissions and hallucinations separately rather than collapsing all errors into a single aggregate score.

D Annotation Protocol

This appendix provides additional annotation details supporting §3.2. The main text reports the overall annotation process and reliability results; here we describe the training, qualification, interface, and adjudication procedures.

D.1 Annotator Training

Annotation was conducted by ten resident physicians with prior exposure to outpatient clinical documentation and treatment planning. Before annotation, all annotators completed a structured training session led by the supervising clinicians.

The training covered four components:

1. the clinical motivation and design principles of the schema;
2. field-level decision rules, including common confusable boundaries such as `conservative_method` versus `lifestyle_habit_modifications`, and `monitoring` versus `follow_up`;
3. a walkthrough of the annotation interface, including validation rules and case flagging; and
4. a fully worked example reviewed jointly by both supervising clinicians.

Annotators also received written guidelines containing field definitions, decision rules, controlled vocabularies, and worked examples for ongoing reference during annotation.

D.2 Qualification Assessment

Before production annotation, each annotator completed a supervised qualification assessment on 100 cases sampled across ICD-10 chapters and schema complexity levels. Performance was evaluated against a supervisor-adjudicated gold standard using field-level syntactic F1.

A minimum score of 0.80 on every field was required to proceed to production. This per-field requirement was chosen instead of a single overall threshold to prevent annotators from passing with strong performance on easier fields while underperforming on clinically important fields such as `medical`, `follow_up`, or `referral`. All ten annotators passed on their first attempt. Per-annotator scores are reported in Table 9.

Following qualification, the supervising clinicians reviewed each annotator's errors and provided feedback on field placement, omitted subfields, and schema-boundary interpretation before production annotation began.

D.3 Annotation Interface

Annotators used a custom Gradio-based interface that displayed each doctor-patient dialogue alongside the structured CAPE schema. The interface enforced required fields, expected data types, list formatting, permitted `null` values, and controlled vocabulary values such as `treatment_status` $\in$ {`start`, `continue`, `stop`, `adjust`}. Annotators could also flag ambiguous or low-quality cases for supervisor review. Figure 4 shows a screenshot of the annotation interface.

Annot.	Labs	Imag.	Med.	Cons.	Life.Mod.	Mon.	Follow	Ref.	Pat.Ed.	WTSMC	Overall
A-01	0.87	0.85	0.83	0.82	0.84	0.86	0.84	0.88	0.91	0.89	0.861
A-02	0.91	0.88	0.86	0.84	0.86	0.89	0.87	0.90	0.93	0.91	0.888
A-03	0.84	0.83	0.81	0.80	0.81	0.83	0.82	0.85	0.89	0.87	0.838
A-04	0.89	0.86	0.85	0.83	0.84	0.87	0.85	0.89	0.92	0.90	0.873
A-05	0.92	0.90	0.88	0.86	0.87	0.91	0.89	0.91	0.94	0.92	0.903
A-06	0.86	0.84	0.82	0.81	0.82	0.85	0.83	0.87	0.90	0.88	0.851
A-07	0.90	0.87	0.85	0.84	0.85	0.88	0.86	0.90	0.92	0.91	0.881
A-08	0.85	0.83	0.81	0.80	0.81	0.84	0.82	0.86	0.90	0.88	0.843
A-09	0.88	0.86	0.84	0.83	0.84	0.87	0.85	0.89	0.91	0.90	0.870
A-10	0.93	0.91	0.89	0.87	0.88	0.92	0.90	0.92	0.95	0.93	0.913
Mean	**0.885**	**0.863**	**0.844**	**0.830**	**0.842**	**0.872**	**0.853**	**0.887**	**0.917**	**0.899**	**0.872**
Std	0.029	0.028	0.027	0.022	0.024	0.029	0.027	0.022	0.017	0.019	0.025

Table 9: Per-annotator qualification scores on the 100-case qualification set, evaluated against the supervisor-adjudicated gold standard using field-level syntactic F1. **Cons.** = `conservative_method`; **Life.Mod.** = `lifestyle_habit_modifications`; **WTSMC** = `when_to_seek_medical_care`. All annotators met the required threshold of $\geq$0.80 on every field.

D.4 Production Annotation and Adjudication

Production annotation used a sharded initial single-pass design: each case was first assigned to exactly one qualified annotator after qualification. After the train/test split was fixed, all 500 test cases underwent an additional five-physician review before model evaluation, with flagged corrections or disagreements adjudicated under the same schema guidelines. Full dual annotation of all cases would have approximately doubled the physician-annotation burden and would have reduced either dataset size or diagnostic coverage. We therefore chose a coverage-oriented design for the released benchmark, while preserving quality control through four mechanisms: a per-field qualification gate, interface-level validation, flag-based adjudication by supervising clinicians, and an independent 50-case supervisor-overlap reliability check disjoint from the qualification set.

A total of 1,905 candidate cases were assigned during production. Annotators flagged 23 cases. Of these, 18 were adjudicated by the supervising clinicians and retained, while 5 were excluded due to source-dialogue quality issues and replaced with additional annotated cases from the same source pool. The final dataset therefore contains 1,900 cases.

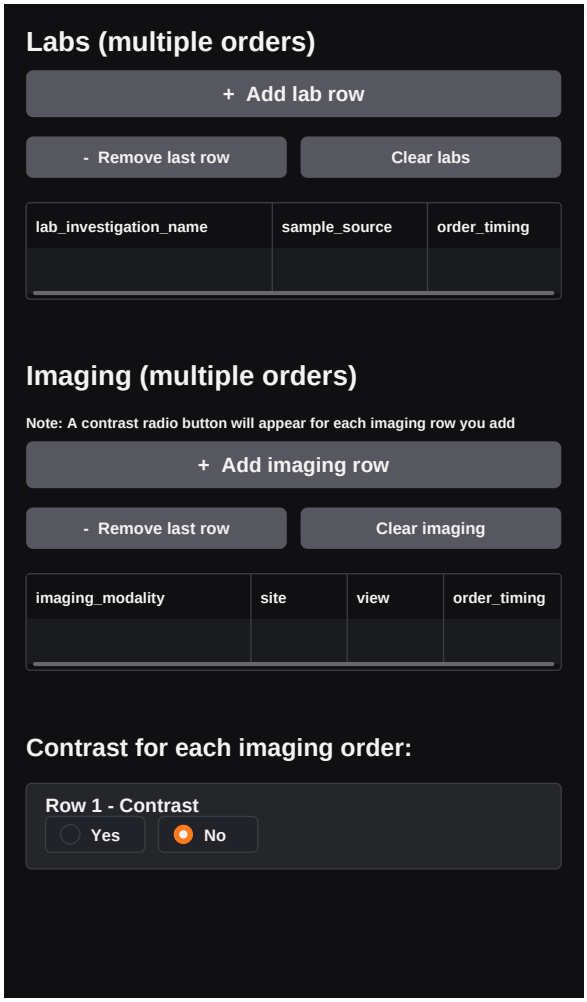

Figure 4: One tab of the CAPE annotation interface, showing structured fields and repeated-row controls for investigation orders.

E Inter-Annotator Agreement

This appendix provides the metric justification and detailed agreement results supporting §3.2. We report agreement at two levels: annotator-level agreement on the 100-case qualification set and supervisor-level agreement on a separate 50-case production overlap subset.

E.1 Agreement Setup

All ten annotators independently annotated the 100-case qualification set, allowing pairwise agreement to be estimated across all annotator pairs. In addition, both supervising clinicians independently annotated a randomly selected 50-case subset of production cases, disjoint from the qualification set. This second analysis provides an approximate empirical ceiling for the task.

Because CAPE contains heterogeneous field types, we use field-appropriate agreement metrics rather than forcing all fields into a single metric.

Structured list fields. For `labs`, `imaging`, `medical`, `monitoring`, `referral`, and `follow_up`, pairwise agreement is computed as the mean item-level F1 across all $C(10, 2) = 45$ annotator pairs. Items are matched by identifying key using chrF-based fuzzy matching. Fleiss' κ is also computed over item-presence decisions, indicating whether each canonical clinical concept is present or absent.

Flat string-list fields. Agreement for the two flat string-list fields, `conservative_method` and `lifestyle_habit_modifications`, is measured using Coverage F1, matching the evaluation protocol in §4. We do not report Fleiss' κ for these fields because canonical item-presence units are not well defined for open-ended string lists.

Free-text fields. For `patient_education` and `when_to_seek_medical_care`, agreement is measured using BERTScore F1 (Zhang et al., 2019) with BioClinicalBERT (Alsentzer et al., 2019). Fleiss' κ is not reported for these fields because they do not naturally reduce to discrete item-presence decisions.

Overall reliability is reported using Krippendorff's α across all field units, computed with the `krippendorff` Python package (Hayes and Krippendorff, 2007). We interpret these values in the context of known limitations of fixed κ thresholds and the differing difficulty of each field type.

E.2 Agreement Results

Table 10 reports agreement by schema field. Agreement is strongest for structured fields with clear identifying keys, such as `labs` and `referral`. Agreement is lower for fields that require more interpretive judgment, including patient education, return precautions, and `conservative_method`. Across all field units, annotator-level Krippendorff's α is 0.812.

Supervisor agreement is higher than annotator agreement, with mean pairwise agreement of 0.888 and supervisor-level Krippendorff's $\alpha = 0.881$. We use this supervisor-level agreement as an approximate empirical ceiling for schema-grounded treatment-plan extraction from dialogue.

E.3 Disagreement Analysis

We qualitatively reviewed 100 sampled disagreement instances from the qualification set. Three

Field	Fleiss $\kappa^{\dagger}$	Pairwise F1 / BERTScore‡	Supervisor Agr.	Annot. vs. Gold	Field Type
`labs`	0.847	0.831	0.912	0.91	Structured list
`imaging`	0.823	0.809	0.895	0.88	Structured list
`monitoring`	0.831	0.818	0.903	0.90	Structured list
`medical`	0.814	0.802	0.887	0.87	Structured list
`follow_up`	0.803	0.791	0.878	0.86	Structured list
`referral`	0.836	0.822	0.908	0.89	Structured list
`conservative_method`	—	0.748	0.854	0.82	Flat string list
`lifestyle_modifications`	—	0.763	0.869	0.83	Flat string list
`patient_education`	—	0.729	0.841	0.80	Free text
`when_to_seek_med.`	—	0.744	0.852	0.81	Free text
Macro-average	**0.824***	**0.806**	**0.888**	**0.87**	All fields

Table 10: Inter-annotator agreement on the 100-case qualification set and supervisor agreement on the 50-case production overlap subset. †Fleiss' κ is computed for structured list fields only; — indicates fields where item-presence κ is not applicable. ‡Pairwise F1 is used for structured and flat list fields; BERTScore F1 is used for free-text fields. Supervisor agreement is pairwise agreement between the two attending clinicians. Annot. vs. Gold is the mean annotator score against the supervisor-adjudicated gold standard. *Macro-average κ covers structured list fields only. Overall Krippendorff's α across all field units is 0.812; supervisor-level Krippendorff's α is 0.881.

recurring patterns emerged.

Field placement disagreements. The most frequent pattern involved content that could plausibly be placed under either `conservative_method` or `lifestyle_habit_modifications`. For example, activity restriction or supportive measures were sometimes interpreted as clinician-directed conservative management and sometimes as patient-directed behavioural advice. This pattern accounted for approximately 41% of reviewed disagreements.

Item boundary disagreements. Annotators sometimes differed in whether related recommendations should be merged into one item or split into multiple items, especially in flat string-list fields. For example, one annotator might record "rest and apply ice packs" as a single item, while another records two separate items. This pattern accounted for approximately 34% of reviewed disagreements.

Subfield omission disagreements. Annotators occasionally differed in how strictly to treat unstated attributes as `null`. For example, some annotators inferred route of administration for familiar tablet medications, while others left route unstated in accordance with the protocol. This pattern accounted for approximately 25% of reviewed disagreements.

These patterns reflect genuine ambiguity in clinical dialogue annotation rather than systematic annotator failure. They also directly motivate the evaluation design: field-placement disagreements motivate explicit misplacement tracking, while subfield omission disagreements motivate separate missing-subfield and hallucinated-subfield counts.

F Evaluation Framework Details

F.1 Metric Validation: Extended Analysis

This appendix provides the full per-conclusion discussion for the metric validation study summarised in §4.1 of the main paper. The study design (100 stratified cases $\times$ 3 models $=$ 300 instances, three physician raters, $\kappa = 0.76$) and Table 1 are reported in the main paper and not repeated here.

ROUGE-L and ROUGE-1. Both ROUGE variants show the weakest correlation with physician judgment ($r = 0.31$ and 0.34 respectively), consistent with prior work on low metric–human alignment for clinical NLG (Croxford et al., 2024). The primary failure mode is abbreviation non-normalisation: a prediction of *"CBC"* against a gold of *"complete blood count"* receives ROUGE-L $= 0.00$ despite being clinically identical. This systematically depresses correlations for structured list fields, where standardised abbreviations are common. A secondary failure is that ROUGE operates on whole-word boundaries, making it insensitive to morphological variants and dosage-string differences (e.g., *"500 mg"* vs. *"500 milligrams"*).

Sentence cosine similarity. Mean-pooled sentence embeddings perform reasonably on free-text fields ($\rho = 0.49$) but poorly on structured fields ($\rho = 0.35$). The core problem is that a single sentence vector cannot discriminate between distinct clinical items that occupy similar embedding neighbourhoods. A model that hallucinates *"warfarin"* when the gold contains *"metformin"* still receives a high cosine score because both are cardiovascular medications. Similarly, hallucinating *"hepatitis B screening"* when the gold contains *"iron supplements"* receives moderate cosine credit because both are laboratory-adjacent concepts. This embedding-neighbourhood conflation is especially harmful for structured list fields, where the evaluation task requires item-level discrimination rather than document-level semantic proximity.

chrF. chrF is the best single syntactic metric ($r = 0.61$, $\rho = 0.59$), substantially outperforming both ROUGE variants. Its character n-gram formulation ($\beta = 1$, $n = 6$) handles abbreviations, partial matches, and morphological variants without relying on whole-word boundaries. For example, *CBC* and *complete blood count* share non-zero character overlap despite having no exact word match. chrF performs better on structured fields ($\rho = 0.61$) than on free-text fields ($\rho = 0.52$), reflecting its strength at controlled-vocabulary and schema-compatible phrasing.

MoverScore. MoverScore achieves the highest overall correlation ($r = 0.71$, $\rho = 0.69$) and the most consistent performance across both field types (structured: $\rho = 0.70$; free-text: $\rho = 0.66$). Crucially, it outperforms BERTScore ($r = 0.52$) despite using the same underlying encoder, confirming that optimal transport alignment over token distributions provides a stronger signal than greedy token matching. BERTScore matches each prediction token to the single best gold token, which can be misleading when clinical strings contain multiple semantically loaded tokens. MoverScore's earth mover's distance minimises the total transport cost across the full token distribution, capturing the aggregate alignment between prediction and reference rather than per-token peaks.

Complementarity and redundancy analysis. Pairwise Pearson correlation between chrF and MoverScore is $r = 0.68$—substantial but below the 0.80 threshold commonly used to declare metrics redundant. This confirms the two metrics capture non-overlapping signals. Figure 3 in the main paper maps each error type to its expected $\Delta = \text{MoverScore} - \text{chrF}$ signature; the same logic applies here: a high-Δ case is typically a valid paraphrase or vocabulary deviation, while a near-zero Δ case is either verbatim extraction or joint failure on both dimensions.

F.2 Encoder Benchmark: Full Results

Benchmark design. To select the MoverScore backbone encoder, we constructed a clinical term pair benchmark of 300 pairs sampled from the `labs` (`lab_investigation_name` subfield) and `lifestyle_modifications` annotations in CAPE—the two fields exhibiting the broadest lexical variation. Pairs were sampled by the same resident physicians who produced the treatment plan annotations and labelled with one of three similarity levels reflecting *clinical* equivalence, not surface form:

- **Equivalent** (1.0): same clinical concept expressed differently, requiring normalisation (e.g., *"CBC"* vs. *"complete blood count"*).
- **Related/confusable** (0.5): clinically related but distinct items that a naive semantic model may incorrectly equate (e.g., *"hepatitis B and C screenings"* vs. *"iron supplements"*).
- **Unrelated** (0.0): no meaningful clinical overlap (e.g., *"stop smoking and drinking alcohol"* vs. *"get plenty of rest"*).

The benchmark includes both *same-field* pairs (field_a = field_b) and *cross-field* pairs (field_a $\neq$ field_b). Cross-field pairs deliberately simulate misplacement scenarios in which content from one schema field leaks semantically into another. Each pair carries an *uncertain* flag for borderline annotations; these were retained but analysed separately in a sensitivity check. The benchmark was iteratively refined over eight annotation rounds (v8) to resolve pairwise disagreements.

For each encoder we computed MoverScore-style token-level optimal transport scores (Earth Mover's Distance over L2-normalised contextual token embeddings; cost = 1 − cosine similarity) and evaluated: Spearman rank correlation with human labels (*Spearman*), AUROC for equivalent vs. related/confusable pairs (*AUROC-ER*), and AUROC for equivalent vs. all non-equivalent pairs (*AUROC-EN*). All embeddings used each model's last hidden state at the token level; pooled sentence embeddings were not used, as MoverScore

requires per-token distributions.

Three diagnostic conclusions. (1) Domain specialisation matters more than parameter scale. The top three models are all PubMedBERT-family encoders (*clinical-sts-S-PubMedBert-MS-MARCO*, *S-PubMedBert-MS-MARCO*, *pubmedbert-base-embeddings*), with Spearman correlations of 0.848, 0.841, and 0.822 respectively. Large general-domain models with far more parameters—*e5-mistral-7b-instruct* (7B, Spearman = 0.658) and *bge-m3* (Spearman = 0.657)—fall considerably below the best biomedical encoders. The general *bge-large-en-v1.5* (Spearman = 0.788) is competitive with smaller biomedical models but still below the PubMedBERT cluster.

(2) Clinical STS fine-tuning provides a consistent but modest benefit. Our fine-tuned variant achieves the highest Spearman correlation (0.848 vs. 0.841 for its base checkpoint); however, the composite margin is only 0.0002, and on the harder AUROC-ER task the base checkpoint marginally outperforms the fine-tuned version (0.927 vs. 0.922). This confirms that biomedical pretraining is the primary driver of performance, with STS fine-tuning providing marginal additional calibration.

(3) The confusable-pair task (AUROC-ER) is the most diagnostic selection criterion. Several encoders rank highly on AUROC-EN but fall sharply on AUROC-ER. The most striking case is *monoelectra-large*, which achieves AUROC-EN = 0.705 but Spearman = 0.226, indicating cross-encoder-style scoring behaviour rather than a true token-level similarity function. We therefore prioritise AUROC-ER as the primary selection criterion: clinical extraction evaluation most critically requires distinguishing semantically similar but clinically distinct items (e.g., *"fasting glucose"* vs. *"random glucose"*).

Rejection of `pubmedbert-base-embeddings`. Although `pubmedbert-base-embeddings` achieves the highest raw AUROC-EN (0.964), it exhibits an anomalously large equivalent-vs-related gap score (0.206 vs. 0.057 for the selected encoder). We confirm that this over-separation persists at the token-level optimal-transport score level, not only at the raw embedding-cosine level: the equivalent-vs-related score gap for `pubmedbert-base-embeddings` is disproportionately large compared with the selected encoder, indicating that the OT alignment step does not fully correct the encoder's calibration issue at this granularity. The selected encoder's lower but more calibrated gap score is therefore preferable as a token-level similarity backbone.

Encoder	Spear.	AUC-ER	AUC-EN	Composite
clinical-sts-S-PubMedBert†	**0.848**	0.922	**0.960**	**0.910**
S-PubMedBert-MS-MARCO	0.841	**0.927**	0.962	0.910
pubmedbert-base-embeddings	0.822	0.929	0.964	0.905
clinical-sts-BAAI-bge-large	0.800	0.911	0.951	0.887
MedEmbed-small-v0.1	0.785	0.920	0.955	0.887
bge-large-en-v1.5	0.788	0.914	0.953	0.885
clinical-sts-MedEmbed-large	0.779	0.915	0.953	0.882
multilingual-e5-large-inst.	0.818	0.875	0.920	0.871
Qwen3-Embedding-0.6B	0.750	0.859	0.913	0.840
e5-large-v2	0.698	0.889	0.922	0.836
bge-m3	0.657	0.887	0.921	0.821
OctenEmbedding-8B	0.648	0.872	0.904	0.808
Qwen3-Embedding-4B	0.615	0.829	0.874	0.773
e5-mistral-7b-instruct	0.658	0.798	0.838	0.765
BidirLM-Omni-2.5B	0.658	0.765	0.815	0.746
BiomedNLP-BiomedBERT-large	0.510	0.748	0.787	0.682
gatortron-base	0.509	0.730	0.780	0.673
ms-marco-MiniLM-L6-v2	0.462	0.689	0.756	0.636
clinical-sts-broberta-large	0.446	0.601	0.701	0.582
monoelectra-large	0.226	0.731	0.705	0.554
quora-roberta-large	0.409	0.553	0.666	0.543
(3 additional models not shown; below composite 0.54)				

Table 11: Full encoder benchmark results on 300 clinical term pairs, sorted by composite score (unweighted mean of Spearman, AUROC-ER, AUROC-EN). Spear. = Spearman correlation with human similarity labels. AUC-ER = AUROC for equivalent vs. related/confusable pairs (harder task). AUC-EN = AUROC for equivalent vs. all non-equivalent pairs. †Selected encoder. **Bold** = best per column.

G Extended Results and Error Analysis

This appendix provides the detailed evidence supporting Section 5: the full 36-model leaderboard, representative per-field breakdowns, semantic–surface gap examples, error-count analyses, schema-misplacement examples, and scaling results.

G.1 Why MoverScore and chrF Must Be Reported Together

The MoverScore–chrF gap is a useful diagnostic signal rather than an artifact of evaluation. We observe three recurring high-Δ patterns. First, **temporal qualifier truncation**: a model extracts "2 weeks" but omits "or sooner if symptoms worsen". Second, **aim absorption**: a model places the follow-up aim inside `scheduled_follow_up_time` and leaves `aim_of_follow_up` null. Third, **register compression**: a model reduces patient-facing instructions such as "seek urgent care if you develop severe pain, fever, or new symptoms" to "severe pain, fever, or new symptoms".

These examples explain why neither metric is sufficient alone. MoverScore recognizes paraphrase and semantic overlap, but can over-credit outputs that violate field boundaries or controlled vocabularies. chrF rewards schema-faithful phrasing, but can over-penalize harmless wording differences such as "5 days" versus "for 5 days". Reporting both metrics makes these failure modes visible.

G.2 Full Model Leaderboard

Table 12 reports the complete CAPE leaderboard for all 36 models. Models are grouped by category and sorted by descending MoverScore within each group. In addition to MoverScore, chrF, and Δ, the table reports parsed-output counts, missing item and subfield counts, hallucinated item and subfield counts, and total misplacement flags when available. Here, *Parsed* denotes the number of model outputs, out of 500 test cases, that produced parseable schema-conformant JSON and could be included in automatic evaluation; it does not indicate clinically perfect extraction. Diagnostic error counts are computed over parsed outputs when available.

G.3 Additional Leaderboard Observations

One notable model-specific exception in the API-accessed group is DeepSeek-V4-Flash. Although evaluated under the same schema and decoding protocol as the other API-accessed systems, it obtains lower CAPE performance than nearby API-accessed models, with MoverScore 0.474 and chrF 0.428. It also produces only 389 parseable schema-conformant outputs out of 500 test cases, whereas most other API-accessed systems produce near-complete parse rates. This suggests that its low score reflects both clinical-content extraction errors and structured-output reliability failures. We therefore treat this as a model-specific exception rather than as representative of API-accessed systems as a class.

Some model families also show local non-monotonicity with scale. For example, Qwen2.5-14B outperforms larger Qwen2.5 variants overall, and Qwen3-VL-8B slightly underperforms Qwen3-VL-4B before the 32B model improves again. These patterns suggest that CAPE performance depends not only on model size, but also on instruction tuning, schema-following behavior, and structured-output control.

G.4 Representative Per-Field Breakdowns

Table 13 compares three representative models: o3, MedGemma-4b-it, and Llama-3.1-8B-Instruct. These models illustrate three distinct profiles: strong general extraction, hallucination-prone medical-specific extraction, and weaker open-weight structured extraction.

	o3		MedGemma-4b		Llama-3.1-8B	
Field	Mover	chrF	Mover	chrF	Mover	chrF
`labs`	0.784	0.744	0.166	0.163	0.116	0.109
`imaging`	0.742	0.741	0.579	0.565	0.545	0.521
`monitoring`	0.720	0.720	0.598	0.594	0.095	0.066
`medical`	0.826	0.712	0.505	0.425	0.558	0.461
`conservative`	0.741	0.720	0.418	0.399	0.123	0.073
`follow_up`	0.769	0.648	0.821	0.567	0.788	0.553
`referral`	0.769	0.641	0.755	0.565	0.722	0.531
`patient_edu.`	0.788	0.704	0.737	0.688	0.686	0.611
`when_to_seek`	0.704	0.788	0.606	0.420	0.467	0.226
Overall	**0.796**	**0.730**	**0.548**	**0.454**	**0.456**	**0.350**

Table 13: Per-field MoverScore/chrF for three representative models.

Llama-3.1-8B-Instruct exhibits a weaker open-weight structured-extraction profile. Its errors are not concentrated in a single anomalous field; instead, performance degrades across multiple schema components, reflecting difficulty with structured-output control, subfield completion, and field-boundary decisions. MedGemma-4b-it shows a different pattern: it achieves high semantic scores on some plan fields but suffers from hallucinated labs, monitoring entries, and unsupported subfields. Together, these models illustrate that CAPE distinguishes between strong schema-grounded extraction, hallucination-prone medical-specific extraction, and weaker general open-weight extraction.

G.5 Semantic–Surface Gap Examples

The examples in this section ground the interpretation of the gap metric Δ introduced in Section G.1. Recall that Δ measures the divergence between a model's MoverScore (which captures semantic proximity) and its chrF score (which captures surface-level character overlap). When Δ is large, the model has preserved the clinical meaning of a field but expressed it in a form that differs substantially from the gold annotation in wording, length, or structure. Crucially, this divergence is not random noise: across the 21 models evaluated, it clusters into a small number of recurring extraction patterns. The three patterns described below—temporal qualifier truncation, aim absorption, and register compression—each produce high MoverScore alongside low chrF in systematic, interpretable ways.

G.5.1 Temporal Qualifier Truncation

The most common source of high Δ in the `scheduled_follow_up_time` field is what we term *temporal qualifier truncation*. In this pattern, a model correctly identifies the core time expression (e.g., "2 weeks", "one week", "1 month") but discards the conditional or contingency clause that modifies it in the gold annotation. The clinical significance of this omission is real: a follow-up instruction of "in 2 weeks *or if you notice any worsening of symptoms, come in earlier*" carries a different patient-safety profile than the truncated prediction "2 weeks". The conditional phrase is precisely the actionable guidance a patient would need if their condition deteriorates between appointments.

Table 14 shows four representative cases drawn from our evaluation set. In each row, the gold annotation contains a compound time expression with a conditional qualifier; the corresponding prediction retains only the primary time unit. The Mover-

Score column confirms that the semantic core is preserved—scores range from 0.843 to 0.866—while the chrF column reveals the surface-level loss, with scores as low as 0.062. This MoverScore–chrF gap is precisely what Δ is designed to detect, and temporal qualifier truncation is its single largest contributor in the follow-up fields.

Gold	Prediction	Mover	chrF
"in 2 weeks or if you notice any worsening of symptoms, come in earlier"	"2 weeks"	0.846	0.062
"in a week or sooner if needed"	"one week"	0.852	0.141
"in 1 month, and then every 6 months"	"1 month"	0.866	0.161
"in a year unless anything changes"	"in a year"	0.843	0.207

Table 14: Temporal qualifier truncation in `scheduled_follow_up_time`. Models preserve the core time expression but omit conditional qualifiers, yielding high MoverScore alongside very low chrF.

G.5.2 Aim Absorption

A second recurring pattern involves the `aim_of_follow_up` subfield. In the gold annotations, the purpose of a follow-up appointment is recorded as a separate structured field from the time of the appointment. Several models, however, merge the two into a single composite string placed entirely within `scheduled_follow_up_time`, leaving `aim_of_follow_up` null. We refer to this as *aim absorption*: the purpose is absorbed into the time field rather than instantiated as a distinct extraction target.

The pattern is not random; it reflects an understandable compression of clinical language. Physicians often phrase follow-up instructions as unified sentences—"come back in six weeks to evaluate symptom control"—and models trained on such text naturally replicate this combined phrasing. The problem is that the schema explicitly separates time from aim, and a null `aim_of_follow_up` field propagates as a structural miss regardless of how semantically rich the time field is.

Table 15 illustrates this pattern across four cases spanning three different models. Note that in every case the predicted `scheduled_follow_up_time` is a valid clinical phrase: a reader would understand what it means. The structural error lies entirely in the null `aim_of_follow_up`, which makes the extraction invalid from a schema-compliance perspective. The examples show that aim absorption is not limited to a single model family: it appears in both frontier/API systems and smaller open-weight models, indicating a general schema-boundary failure mode.

Case	Gold time / aim	Predicted time / aim	Model
781	"in 6 weeks" / "to evaluate symptom control and adjust treatment"	"6 weeks to evaluate symptom control and adjust treatment" / *null*	Llama-3.1-8B
1112	"in 8 weeks" / "to assess response to treatment"	"8 weeks to assess response to treatment" / *null*	Llama-3.1-8B
425	"in 6 weeks" / "to assess how the treatment is going"	"6 weeks to assess how the treatment is going" / *null*	Gemini-2.5
1723	"in one week" / "to review MRI results and finalise the surgical plan"	"one week to review MRI results and finalise the surgical plan" / *null*	Kimi-k2.6

Table 15: Aim absorption: the model places the purpose of the follow-up inside `scheduled_follow_up_time` and leaves `aim_of_follow_up` null.

G.5.3 Register Compression

The third pattern, *register compression*, affects narrative fields such as `conservative_method` and the contingency clauses of monitoring and follow-up instructions. In these fields, gold annotations are written in full patient-facing language—they include the trigger condition, an explicit instruction verb, and often a target action (e.g., "if your symptoms worsen before then, please get in touch sooner"). Models tend to compress these phrases to their content core, stripping the imperative framing and conditional scaffolding. The result reads more like a clinical shorthand note than a patient instruction.

This compression is consequential in context. A treatment plan is ultimately communicated to a patient, and the instructional register of the gold annotation—"please get in touch", "seek immediate medical attention"—is part of the clinical content, not decorative phrasing. Its removal reduces actionability for the patient even if the semantic content scores reasonably under embedding-based metrics.

Table 16 provides five examples drawn from narrative fields. Each entry shows the full gold annotation alongside its compressed prediction. The pattern is consistent: the trigger keyword or medical term is retained, but the instruction verb, ur-

gency framing, and patient-directed phrasing are eliminated. In Case 55, the compression additionally fragments a unified instruction into two separate items ("include more fiber" and "reduce fat intake"), which also introduces a minor schema violation in fields that expect a single string value.

Case	Gold	Prediction
1467	"if your symptoms worsen before then, please get in touch sooner. Report any adverse reactions immediately."	"if symptoms worsen"
1879	"if you experience severe pain, fever, or any new symptoms, seek immediate medical attention"	"severe pain, fever, or any new symptoms"
779	"if you experience any severe side effects, please contact us immediately"	"severe side effects"
1647	"we'll administer IV fluids to help with rehydration"	"IV fluids"
55	"dietary changes, like including more fiber and reducing fat intake"	"include more fiber"; "reduce fat intake"

Table 16: Register compression examples: models preserve the clinical trigger term but remove patient-facing instruction and conditional framing.

Together, the three patterns documented above explain the bulk of high-Δ cases in the evaluation. They share a common structure: the model extracts the most salient semantic unit—the time, the aim keyword, the trigger term—while discarding the structural and pragmatic context that surrounds it. This is not a failure of clinical understanding; it is a failure of schema-faithful, register-aware extraction, and it motivates the metric design choices described in Section G.1.

G.6 Extended Error Analysis

The high-Δ patterns in the previous section represent a specific failure mode: content that is semantically present but structurally or linguistically incomplete. This section broadens the error taxonomy to cover four distinct error types observed across all 36 models: missing content (extractions that are absent or incomplete), hallucinated content (extractions that are fabricated or not grounded in the source dialogue), annotation-philosophy mismatches (cases where the model's extraction is clinically defensible but diverges from the annotators' conventions), and schema-boundary errors (cases where content is present but misplaced). For each type, we identify the models and fields most affected and interpret the underlying causes.

G.6.1 Missing Content

Missing content is a recurring failure mode in CAPE, especially for fields that require separating compact clinical phrases into multiple schema slots. The most common subfield omissions occur in follow-up, referral, and medication attributes. These errors often arise when a model extracts the existence of an action but leaves its clinically actionable purpose or status unspecified. For example, models may capture that a follow-up visit is needed while omitting `aim_of_follow_up`, or capture that a medication is prescribed while omitting `treatment_status`. This motivates our separate reporting of item-level and subfield-level omissions rather than relying only on aggregate semantic similarity.

Field	Miss. I	Miss. S	Dominant missing content
`conservative`	606	0	Lifestyle/conservative boundary items
`labs`	424	0	Item-level omissions
`medical`	136	925	`treatment_status` omissions
`imaging`	206	0	Item-level omissions
`referral`	25	362	`aim_of_referral`
`follow_up`	3	398	`aim_of_follow_up`
`monitoring`	116	0	Item-level omissions

Table 17: Missing content profile for a representative omission-heavy model. Miss. I counts item-level omissions (an entire list item absent); Miss. S counts subfield omissions within otherwise-present items. The profile is omission-heavy, particularly in conservative management and in the aim subfields of follow-up and referral.

G.6.2 Hallucinated Content

While stronger models tend toward omission, weaker models—particularly medical-specific smaller models—tend toward hallucination. To make this contrast concrete, Table 18 reports hallucination counts for a selected subset of models, again distinguishing item-level hallucinations (Hall. I: fabricated list items with no grounding in the dialogue) from subfield hallucinations (Hall. S: fabricated values within an otherwise-grounded item structure). The models are ordered to highlight the contrast between the highest-hallucination and lowest-hallucination ends of the range.

HuatuoGPT-3-8B shows the most severe hallucination burden overall: 2237 item-level and 3893 subfield hallucinations, dominated by fabricated referral entries and invented subfield values.

MedGemma-4b-it generates 1697 item-level and 2929 subfield hallucinations—concentrated in labs and monitoring—showing that domain pretraining does not confer reliable structured extraction behavior. Llama-3.1-8B and minimax-m2.7 occupy a middle tier with hallucination concentrated in labs and referral fields respectively.

At the low end, openai/o3 and gemini-2.5-flash show lower hallucination counts than weaker models, with errors distributed across fields rather than dominated by fabricated referral entries or invented monitoring values. This contrast shows that hallucination burden is strongly model-dependent and motivates reporting explicit hallucination counts alongside aggregate similarity metrics.

Model	Hall. I	Hall. S	Dominant pattern
HuatuoGPT-3-8B	2237	3893	Fabricated referrals/subfields
Llama-3.1-8B	1791	1912	Hallucinated labs and plan items
MedGemma-4b-it	1697	2929	Labs and monitoring hallucination
minimax-m2.7	983	1916	Referral and subfield overgeneration
openai/o3	586	579	Lower, distributed hallucination
gemini-2.5-flash	790	655	Lower, distributed hallucination

Table 18: Hallucination profiles for selected models. Hall. I counts item-level hallucinations (fabricated list items); Hall. S counts subfield hallucinations within otherwise-present items. Medical-specific smaller models show the highest hallucination burden; proprietary frontier models show substantially lower counts, while weaker models show substantially larger fabricated-item and fabricated-subfield burdens.

G.6.3 Condition-Treated Label Substitution

Not all errors reflect a failure of comprehension. Some are better described as annotation-philosophy mismatches: cases in which a model's extraction is clinically reasonable but systematically diverges from the convention adopted by annotators. The most consistent such mismatch occurs in the `medical.condition_treated` field.

The gold annotations in this field record the physician's stated functional purpose for prescribing a treatment—the clinical rationale as it would be explained to the patient. For example, a physician might explain that a corticosteroid is prescribed for "inflammation in your airways" rather than naming the diagnosis directly. Models, however, tend to substitute the diagnostic label: they extract "asthma" where the gold annotation reads "inflammation in your airways". This substitution is not wrong medically; it is a different representational choice.

Table 19 shows six examples of this substitution drawn from two models. Kimi-k2.6 accounts for four of the six and consistently extracts the diagnostic label ("asthma", "Crohn's disease", "cystic acne", "interstitial cystitis") rather than the stated mechanism or symptom cluster. Gemini-2.5 shows a similar tendency, though with less regularity. Taken together, these cases suggest that models default to diagnostic reasoning rather than patient-facing explanatory language—a preference that is appropriate in clinical documentation but penalized by the gold annotations, which privilege the physician's stated purpose.

Table 19: Condition label substitution in `medical.condition_treated`. Gold annotations record the physician's stated treatment purpose in patient-facing terms; models tend to substitute the diagnostic label instead. The error is an annotation-philosophy mismatch rather than a clinical error.

Case	Gold	Prediction	Model
1448	"inflammation in your airways"	"asthma"	Kimi-k2.6
856	"reduce inflammation in intestines"	"Crohn's disease"	Kimi-k2.6
1015	"inflammation and bacteria causing the acne"	"cystic acne"	Kimi-k2.6
1280	"pain and inflammation"	"interstitial cystitis"	Kimi-k2.6
438	"inflammation"	"Bell's palsy"	Gemini-2.5
1892	"help control blood pressure and reduce fluid buildup"	"reduce fluid buildup"	Gemini-2.5

G.7 Schema Misplacement Analysis

A fourth category of error is schema misplacement: the model extracts clinically relevant content but assigns it to the wrong field or subfield. This error type is particularly important because it can receive high semantic similarity scores—the content is there, but it is in the wrong place—while still producing a structurally invalid treatment plan. A downstream system consuming the extracted JSON would retrieve the content from the wrong slot, potentially with patient-safety consequences.

Table 20 provides an overview of the most frequent misplacement pairs observed in the evaluation. Each pair is characterized by a directionality pattern (absorption, swap, or merge) and an interpretation that explains why the boundary is structurally meaningful even though it may not always be intuitively obvious. The six pairs cover

all major structural boundaries in the schema: between conservative management and patient behavior, between preventive care and lifestyle change, between procedures and medications, between one-time tests and longitudinal monitoring, within imaging objects, and within medication objects. The subsections below discuss the three most consequential pairs in detail, with examples.

Table 20: Common schema-misplacement patterns identified across the 36-model evaluation. Absorption indicates content from one field is merged into another; swap indicates the two fields exchange their values; merge indicates two distinct subfields are concatenated into one.

Misplacement pair	Pattern	Interpretation
`conservative` ↔ `lifestyle`	Absorption/merge	Clinician intervention vs. patient behavior
`prevention` ↔ `lifestyle`	Absorption	Preventive care phrased as behavior change
`conservative` ↔ `medical`	Absorption/merge	Procedure parsed as medication object
`labs` ↔ `monitoring`	Absorption	One-time test vs. longitudinal monitoring
`imaging modality` ↔ `site`	Merge	Natural clinical phrase violates schema split
`route` ↔ `dosage_form`	Swap	Annotation guideline ambiguity

G.7.1 `conservative_method` vs. `lifestyle_habit_modifications`

The boundary between `conservative_method` and `lifestyle_habit_modifications` is the most frequently violated schema boundary in the evaluation, and it is also the most conceptually loaded. Conservative methods are clinician-directed interventions: procedures, devices, or protocols that require a healthcare professional to initiate or supervise (e.g., applying a cervical collar, aspirating a hydrocele). Lifestyle modifications are patient-initiated behaviors: changes to diet, exercise, sleep, or daily routine that the patient can implement independently (e.g., balanced diet, stress management, avoiding strenuous activity). The clinical distinction matters because the two categories imply different ownership, different supervision requirements, and different accountability structures.

Table 21 shows four cases where models violate this boundary. In Cases 72 and 159, the gold conservative field is empty—no clinician-directed intervention is mentioned—but models place lifestyle content (balanced diet, surgical site care) under conservative. In Case 1120, a genuine conservative item (cervical collar) coexists with genuine lifestyle items (ergonomic adjustment, stretching), but models merge both under conservative. In Case 1294, an invasive procedure (aspirate hydrocele, a conservative gold item) is accompanied by a lifestyle restriction (avoid strenuous activity), and the lifestyle item migrates to conservative. The common thread is that models lack a stable representation of the clinician-intervention boundary and default to treating any non-medication recommendation as conservative management.

Table 21: Representative conservative/lifestyle boundary violations. Gold columns show the correct schema placement; the rightmost column shows where models typically place the misattributed content. In all four cases, lifestyle content migrates to the conservative field.

Case	Gold conservative	Gold lifestyle	Typical prediction
72	Empty	Balanced diet, stress management, physical activity	Lifestyle content placed under conservative
159	Empty	Keep surgical site clean, limit lifting	Lifestyle content placed under conservative
1120	Cervical collar	Ergonomic adjustment, stretching	Both merged under conservative
1294	Aspirate hydrocele	Avoid strenuous activity	Lifestyle item placed under conservative

G.7.2 Procedure Parsed as Medication

A subtler misplacement occurs when models apply the structure of a `medical` medication object to entries that the schema classifies as `conservative_method`. The `medical` field expects a structured object with subfields such as drug name, dosage, frequency, and route. The `conservative_method` field expects a free-text description of a clinician-directed non-pharmacological intervention. Some procedures sit ambiguously between these categories: adjuvant chemotherapy, phlebotomy, and plasmapheresis are clinician-directed treatments that could plausibly be framed as either medical procedures or as structured medication events.

Gemini-2.5-Flash systematically predicts

medical objects for all three of these gold conservative_method entries. The predictions are medically intelligible—a clinician reading them would understand what is being described—but they violate the schema boundary between treatment procedures and pharmacological medication objects. This is another instance of the hallucination-adjacent behavior documented in Table 18, where the model over-generates structure in a field it treats as more tractable than the intended target.

G.7.3 Imaging Site Absorption

Within the imaging field, each item is structured with separate modality and site subfields—for example, modality "CT scan" and site "chest". This split reflects a deliberate schema design choice: it allows downstream systems to independently query by imaging type or body region. However, clinical language almost never separates the two. A physician says "chest CT scan" or "renal ultrasound" as a unified phrase, and models reproduce that natural phrasing, placing the combined expression in the modality subfield and leaving site null.

Table 22 shows four cases of this site absorption pattern. In each case, the gold annotation correctly separates modality from site, but the predicted modality value concatenates both into a single string. The error is structurally invalid under the schema even though the predicted string contains all the information needed to recover the correct annotation. This is a case where the schema imposes a granularity that diverges from natural clinical language, and the misplacement is a direct consequence of that divergence.

Table 22: Imaging site absorption. Gold annotations split modality and site into separate subfields; models concatenate the two into the modality field, leaving the site subfield null.

Case	Gold modality	Gold site	Predicted modality
1445	CT scan	chest	"chest CT scan"
1261	ultrasound	renal	"renal ultrasound"
1268	MRI	renal	"renal MRI"
464	CT scan	brain	"CT brain scan"

G.7.4 route vs. dosage_form

The final misplacement pair concerns two subfields within the medical medication object: route (the path by which a drug enters the body, e.g., intravenous, oral) and dosage_form (the physical form of the drug, e.g., injection, tablet, inhaler). The two concepts are clinically distinct—a drug can be administered intravenously (route) as a solution (form), or inhaled (route) as a dry powder (form)—but clinical language frequently uses them interchangeably. "Injection" describes both a route and a form; "inhaled" describes both a route and a form.

Table 23 shows three cases where models place the value in route rather than dosage_form as the gold annotation specifies. In all three cases, the gold route is null and the gold dosage_form is non-null ("injection", "inhaled", "injection"), but the predicted output swaps the assignment. Unlike the imaging site absorption cases, these swaps reflect genuine annotation guideline ambiguity: the decision to classify "injection" as a form rather than a route is a convention that reasonable annotators might resolve differently. Models penalized for this swap are being held to a guideline distinction that is not consistently recoverable from the clinical text itself.

Table 23: route–dosage_form swaps. Gold annotations place the value in dosage_form with null route; predictions reverse the assignment.

Case	Gold route	Gold form	Pred route	Pred form
1268	*null*	"injection"	"injection"	*null*
1448	*null*	"inhaled"	"inhaled"	*null*
1173	*null*	"injection"	"injection"	*null*

G.8 Scaling Analysis

The main results in Section 5 show that larger models generally outperform smaller ones on CAPE, but the relationship between parameter count and performance is not monotonic across all model families or all fields. This section examines scaling behavior in detail, using MoverScore as the primary lens because it is the metric most sensitive to semantic completeness rather than surface form.

Table 24 reports MoverScore at each available parameter scale for seven model families, along with the net gain from smallest to largest available size. Several patterns emerge. First, the absolute gain from scaling varies enormously across families: OpenBioLLM improves by 0.478 MoverScore points from 8B to 70B, while HuatuoGPT-3 improves by only 0.078 from 8B to 32B. This

difference cannot be explained by the parameter gap alone; it reflects underlying differences in pre-training data, instruction-tuning coverage, and schema-following capability. Second, two families exhibit local non-monotonicity, marked with † in the table: Qwen2.5 shows a regression from 32B (0.700) to 72B (0.654), and Qwen3-VL shows a regression from 4B (0.647) to 8B (0.634) before recovering to 0.714 at 32B. These regressions indicate that parameter count is not a reliable proxy for schema-faithful extraction—other factors, potentially including instruction tuning quality and output format conditioning, interact with scale in non-additive ways.

Table 24: MoverScore by model family and parameter scale. Small, Mid, and Large/XL columns report the best available representative at each size tier; Gain reports the net improvement from smallest to largest. The † symbol marks models where a larger version scores lower than a smaller one within the same family, indicating local non-monotonicity.

Family	Small	Mid	Large/XL	Gain
Gemma-3	0.507 (4B)	0.608 (12B)	0.646 (27B)	+0.139
Llama-3.1	0.456 (8B)	—	0.686 (70B)	+0.230
Llama-3.2-Vision	0.487 (11B)	—	0.683 (90B)	+0.196
OpenBioLLM	0.184 (8B)	—	0.662 (70B)	+0.478
HuatuoGPT-3	0.525 (8B)	0.603 (32B)	—	+0.078
Qwen2.5	0.573 (7B)	0.683 (14B)	0.700 (32B), 0.654 (72B)†	+0.127
Qwen3-VL	0.360 (2B)	0.647 (4B), 0.634 (8B)†	0.714 (32B)	+0.354

The fields where scaling gains are strongest are those that require the model to make principled schema-boundary decisions—precisely the fields where misplacement errors concentrate (Section G.7). This connection suggests that larger models do not simply learn more clinical vocabulary; they develop a more stable representation of the schema's structural constraints, allowing them to assign content to the correct field more reliably. However, no model family at any scale eliminates the conservative-management bottleneck documented in Table 17: conservative is systematically under-extracted at every parameter tier we evaluated.

Taken together, the scaling results suggest that CAPE measures capabilities that are correlated with scale but not reducible to it. Instruction following, temporal scoping, register-aware extraction, and schema-faithful field placement are all skills that improve with model size—but they improve at different rates, in different fields, and not always monotonically. This motivates continued evaluation with CAPE as models grow, rather than assuming that improvements on benchmark suites optimized for other tasks will translate automatically to structured clinical extraction.

G.9 Additional PriMock57 Results

This appendix reports the full real-dialogue validation results summarized in §6.1. We evaluate seven representative models on PRIMOCK57 (Korfiatis et al., 2022), a corpus of 57 real UK primary care consultations. Each consultation was annotated using the same schema and annotation protocol as CAPE. Table 25 reports per-field chrF, MoverScore, aggregate error counts, and misplacement flags.

The real-dialogue results provide preliminary evidence of similar patterns to those observed on CAPE: frontier general-purpose models have the highest point estimates overall, medical-specific models do not surpass frontier general models, and large semantic–surface gaps again appear in temporally complex fields such as follow_up. The main exception is `deepseek-v4-flash`, which drops sharply on `referral` and `follow_up` in PRIMOCK57; its `referral` MoverScore is only 0.114 and its `follow_up` MoverScore is 0.307. Qualitative inspection indicates that this reflects near-complete omission of these care-continuity fields rather than surface-form mismatch, making it the clearest model-specific real-dialogue sensitivity observed in this replication.

Table 12: Full model performance on CAPE with uncertainty estimates. ↑ = higher is better. Δ = MoverScore$-$chrF. Scores are reported as mean $\pm$ 95% bootstrap confidence-interval half-width, computed by non-parametric bootstrap over the 500 test cases with 10,000 resamples. For Δ, intervals are computed from paired per-case differences $(\mathrm{MoverScore} - \mathrm{chrF})$ within each bootstrap resample. *Parsed* = number of model outputs, out of 500, that produced parseable schema-conformant JSON. *Miss. I/S* = missing item/subfield counts. *Hall. I/S* = hallucinated item/subfield counts. *Misp.* = total misplacement flags. Model blocks follow provenance/access and size, rather than post-hoc performance profile.

Model	**Type**	**Mover ↑**	**chrF ↑**	**Δ**	**Parsed**	**Miss. I/S**	**Hall. I/S**	**Misp.**
Closed-Source Models								
openai/o3	General	$0.796_{\pm 0.014}$	$0.730_{\pm 0.016}$	$+0.066_{\pm 0.009}$	499	388 / 1033	586 / 579	19
google/gemini-2.5-flash	General	$0.792_{\pm 0.014}$	$0.709_{\pm 0.017}$	$+0.083_{\pm 0.009}$	498	439 / 439	790 / 655	33
google/gemini-3-flash-preview	General	$0.765_{\pm 0.016}$	$0.698_{\pm 0.017}$	$+0.067_{\pm 0.009}$	497	482 / 557	825 / 1077	23
openai/gpt-5-mini	General	$0.765_{\pm 0.015}$	$0.706_{\pm 0.016}$	$+0.059_{\pm 0.009}$	499	503 / 615	672 / 814	22
x-ai/grok-4.20	General	$0.760_{\pm 0.016}$	$0.698_{\pm 0.017}$	$+0.062_{\pm 0.009}$	500	597 / 804	563 / 920	35
moonshotai/kimi-k2.6	General	$0.725_{\pm 0.017}$	$0.644_{\pm 0.018}$	$+0.081_{\pm 0.011}$	499	621 / 1143	891 / 1384	24
google/gemini-3.1-flash-lite	General	$0.708_{\pm 0.018}$	$0.631_{\pm 0.019}$	$+0.077_{\pm 0.010}$	500	664 / 1089	878 / 1203	19
Open-Source Large (≥27B)								
meta-llama/Llama-4-Maverick-17B	General	$0.706_{\pm 0.018}$	$0.620_{\pm 0.019}$	$+0.086_{\pm 0.011}$	500	733 / 1294	967 / 698	33
cohere/command-r-plus	General	$0.680_{\pm 0.019}$	$0.581_{\pm 0.020}$	$+0.099_{\pm 0.012}$	478	1130 / 825	589 / 1401	47
mistralai/mistral-small-2603	General	$0.678_{\pm 0.019}$	$0.589_{\pm 0.020}$	$+0.089_{\pm 0.012}$	500	686 / 883	995 / 1736	14
minimax/minimax-m2.7	General	$0.662_{\pm 0.020}$	$0.571_{\pm 0.021}$	$+0.091_{\pm 0.013}$	494	1287 / 787	983 / 1916	95
z-ai/glm-5.1	General	$0.648_{\pm 0.021}$	$0.583_{\pm 0.021}$	$+0.066_{\pm 0.011}$	460	628 / 1472	599 / 443	26
deepseek/deepseek-v4-flash	General	$0.474_{\pm 0.025}$	$0.428_{\pm 0.024}$	$+0.046_{\pm 0.011}$	389	1643 / 1489	543 / 318	71
Qwen/Qwen3-VL-32B-Instruct	General	$0.714_{\pm 0.018}$	$0.629_{\pm 0.019}$	$+0.085_{\pm 0.011}$	497	459 / 1112	1001 / 2215	13
Qwen/Qwen2.5-32B	General	$0.700_{\pm 0.019}$	$0.599_{\pm 0.020}$	$+0.100_{\pm 0.013}$	475	817 / 1064	668 / 1567	29
meta-llama/Llama-3.1-70B-Instruct	General	$0.686_{\pm 0.019}$	$0.601_{\pm 0.020}$	$+0.086_{\pm 0.012}$	477	759 / 1929	1062 / 571	45
meta-llama/Llama-3.2-90B-Vision	General	$0.683_{\pm 0.019}$	$0.595_{\pm 0.020}$	$+0.088_{\pm 0.012}$	473	742 / 1914	1093 / 605	40
Qwen/Qwen2.5-72B	General	$0.654_{\pm 0.020}$	$0.557_{\pm 0.021}$	$+0.097_{\pm 0.013}$	500	896 / 1303	1073 / 1844	60
google/gemma-3-27b-it	General	$0.646_{\pm 0.021}$	$0.576_{\pm 0.021}$	$+0.070_{\pm 0.012}$	497	1569 / 1746	1141 / 706	87
Open-Source Medium (7B-14B)								
Qwen/Qwen2.5-14B	General	$0.683_{\pm 0.019}$	$0.589_{\pm 0.020}$	$+0.094_{\pm 0.012}$	500	761 / 1049	776 / 2348	19
Qwen/Qwen3-VL-8B-Instruct	General	$0.634_{\pm 0.021}$	$0.556_{\pm 0.022}$	$+0.078_{\pm 0.013}$	495	1657 / 1243	726 / 1725	150
google/gemma-3-12b-it	General	$0.608_{\pm 0.022}$	$0.532_{\pm 0.022}$	$+0.076_{\pm 0.013}$	448	1459 / 1790	1224 / 561	69
Qwen/Qwen2.5-7B-Instruct	General	$0.573_{\pm 0.023}$	$0.485_{\pm 0.023}$	$+0.089_{\pm 0.013}$	488	1752 / 1088	1043 / 2084	60
meta-llama/Llama-3.2-11B-Vision	General	$0.487_{\pm 0.025}$	$0.378_{\pm 0.024}$	$+0.108_{\pm 0.015}$	476	1686 / 1398	1888 / 1694	138
meta-llama/Llama-3.1-8B-Instruct	General	$0.456_{\pm 0.025}$	$0.350_{\pm 0.024}$	$+0.106_{\pm 0.015}$	486	1706 / 1340	1791 / 1912	148
Open-Source Small (<7B)								
Qwen/Qwen3-VL-4B-Instruct	General	$0.647_{\pm 0.021}$	$0.560_{\pm 0.022}$	$+0.086_{\pm 0.013}$	489	1680 / 911	728 / 2882	84
google/gemma-3-4b-it	General	$0.507_{\pm 0.025}$	$0.437_{\pm 0.025}$	$+0.070_{\pm 0.014}$	400	1389 / 759	608 / 2313	39
Qwen/Qwen3-VL-2B-Instruct	General	$0.360_{\pm 0.026}$	$0.246_{\pm 0.024}$	$+0.115_{\pm 0.016}$	342	1193 / 464	2177 / 2995	70
meta-llama/Llama-3.2-3B-Instruct	General	$0.279_{\pm 0.026}$	$0.217_{\pm 0.024}$	$+0.062_{\pm 0.014}$	287	930 / 666	1459 / 1994	129
Medical-Specific Models								
aaditya/OpenBioLLM-70B	Medical	$0.662_{\pm 0.021}$	$0.586_{\pm 0.021}$	$+0.076_{\pm 0.012}$	430	1676 / 1905	370 / 306	17
google/medgemma-27b-it	Medical	$0.658_{\pm 0.020}$	$0.608_{\pm 0.020}$	$+0.051_{\pm 0.010}$	498	1582 / 1636	999 / 880	114
FreedomIntelligence/HuatuoGPT-3-32B	Medical	$0.603_{\pm 0.022}$	$0.491_{\pm 0.023}$	$+0.111_{\pm 0.014}$	488	594 / 993	1804 / 2703	17
google/MedGemma-4b-it	Medical	$0.548_{\pm 0.023}$	$0.454_{\pm 0.023}$	$+0.095_{\pm 0.013}$	476	1815 / 738	1697 / 2929	19
FreedomIntelligence/HuatuoGPT-3-8B	Medical	$0.525_{\pm 0.024}$	$0.414_{\pm 0.024}$	$+0.111_{\pm 0.015}$	489	1164 / 803	2237 / 3893	83
aaditya/Llama3-OpenBioLLM-8B	Medical	$0.184_{\pm 0.025}$	$0.172_{\pm 0.023}$	$+0.013_{\pm 0.008}$	170	1048 / 208	96 / 466	4

Table 25: Per-field MoverScore (M), chrF (C), and overall performance on PRIMOCK57 (57 real primary care consultations). Per-field values are point estimates. Overall C and M are reported as mean ± 95% bootstrap confidence-interval half-width, computed by non-parametric bootstrap over the 57 consultations with 10,000 resamples. Overall = macro-average across all fields. Parsed. = fully successful cases. Miss. I/S = missing item / subfield counts. Hall. I/S = hallucinated item / subfield counts. Misp. = misplacement flags. Models sorted by descending overall MoverScore.

	labs		**imaging**		**monitor.**		**medical**		**conserv.**		**follow_up**		**referral**		**pat._edu.**		**wts_care**		**overall**		**Parsed.**	**Miss.I**	**Miss.S**	**Hall.I**	**Hall.S**	**Misp.**
Model	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M	C	M						
Closed-source																										
gemini-3-flash-preview	.704	.729	.693	.721	.565	.578	.607	.708	.546	.674	.751	.818	.635	.728	.698	.748	.634	.731	.648$_{\pm.050}$	**.715$_{\pm.047}$**	57	66	35	105	142	8
gpt-5-mini	.676	.688	.657	.686	.593	.617	.653	.746	.611	.726	.658	.776	.643	.714	.741	.756	.671	.721	.656$_{\pm.048}$	**.714$_{\pm.046}$**	57	62	53	94	74	10
Open-source																										
Llama-3.1-8B	.082	.089	.496	.521	.048	.078	.425	.524	.055	.106	.509	.751	.486	.683	.588	.666	.173	.421	.310$_{\pm.069}$	.419$_{\pm.068}$	56	111	28	215	162	7
command-r-plus	.528	.541	.564	.590	.638	.664	.561	.669	.280	.434	.525	.759	.526	.701	.676	.718	.503	.616	.531$_{\pm.060}$	.630$_{\pm.058}$	51	95	49	53	85	2
deepseek-v4-flash	.396	.550	.364	.364	.612	.634	.478	.523	.201	.276	.254	.307	.112	.114	.576	.597	.470	.514	.378$_{\pm.070}$	.424$_{\pm.069}$	50	99	84	55	62	9
Medical-specific																										
MedGemma-4b-it	.104	.128	.501	.511	.495	.529	.337	.431	.342	.358	.476	.722	.456	.659	.615	.662	.337	.517	.404$_{\pm.066}$	.498$_{\pm.065}$	56	123	20	108	236	3
HuatuoGPT-3-8B	.345	.360	.500	.536	.021	.026	.308	.419	.116	.253	.426	.714	.541	.713	.639	.696	.436	.613	.364$_{\pm.067}$	.474$_{\pm.066}$	53	94	16	162	217	4

H Fine-tuning Details

This appendix provides the training details and full results for the fine-tuning experiments summarized in §6.2. We fine-tune `Llama-3.1-8B-Instruct` and `Qwen2.5-7B-Instruct` under four objectives: supervised fine-tuning (SFT), direct preference optimization (DPO; Rafailov et al., 2023), identity preference optimization (IPO; Azar et al., 2024), and odds-ratio preference optimization (ORPO; Hong et al., 2024).

H.1 Training Setup

All fine-tuning runs use LoRA (Hu et al., 2022) applied to all attention and feed-forward projection matrices: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. We use rank r=16, scaling α=32, and dropout 0.05. All runs use bfloat16 precision, gradient checkpointing, cosine learning-rate scheduling, warmup ratio 0.1, weight decay 0.01, gradient clipping at norm 1.0, maximum sequence length 4096, seed 42, and 3 training epochs. The objective-specific hyperparameters are shown in Table 26.

Objective	LR	Eff. Batch	Key Config
SFT	2×10^{-5}	8 (2×4)	CE loss on assistant turn
DPO	5×10^{-7}	8 (1×8)	β=0.1, frozen reference model, chosen–rejected pairs
IPO	5×10^{-7}	8 (1×8)	β=0.1, implicit preference objective
ORPO	8×10^{-6}	8 (1×8)	λ=0.1, joint SFT+odds-ratio loss

Table 26: Fine-tuning hyperparameters. All runs share a cosine scheduler, warmup ratio 0.1, weight decay 0.01, maximum sequence length 4096, seed 42, three epochs, and LoRA rank r=16 with α=32.

H.2 Training Data

The SFT dataset, contains 1,400 training examples, one per CAPE training case. Each example is formatted as a three-turn message list: a system turn containing the schema and annotation guidelines, a user turn containing the clinical dialogue, and an assistant turn containing the gold treatment-plan JSON. The model's native chat template is applied internally by `SFTTrainer` (von Werra et al., 2020).

DPO, IPO, and ORPO use the same preference dataset of 2,181 chosen–rejected pairs. Chosen responses are physician-annotated gold plans. Rejected responses are model-generated predictions scoring below the 20th percentile on chrF against gold on the training split. This construction ensures that rejected outputs correspond to substantive extraction failures rather than minor paraphrases.

All preference-optimized models are initialized directly from the base instruction-tuned checkpoint rather than from the SFT checkpoint. This makes SFT, DPO, IPO, and ORPO directly comparable as independent training objectives.

H.3 Evaluation Protocol

All fine-tuned models are evaluated on the same 500-case CAPE test set using the same inference script, prompt, schema, and decoding parameters as the main benchmark. No prompt adaptation is applied to fine-tuned checkpoints. Thus, performance differences are attributable to training rather than evaluation-time prompt changes.

H.4 Full Fine-tuning Results

Table 27 reports full per-field scores and error counts for the base and fine-tuned variants. SFT produces the strongest aggregate results for both model families. Among preference-optimization objectives, DPO improves over the base models but remains weaker than IPO, ORPO, and SFT in overall MoverScore. IPO gives larger semantic gains than DPO but increases missing subfields, whereas ORPO gives the strongest hallucination–completeness tradeoff among the preference objectives.

H.5 Preference-Optimization Error Profiles

The aggregate MoverScore improvements from DPO, IPO, and ORPO hide different error profiles. DPO provides a useful standard preference-optimization baseline: for Llama-3.1-8B, it increases MoverScore from 0.456 to 0.517 and reduces hallucinated items from 1,791 to 1,397; for Qwen2.5-7B, it increases MoverScore from 0.573 to 0.602 and reduces hallucinated items from 1,043 to 834. However, DPO remains below IPO, ORPO, and SFT in overall semantic extraction quality.

IPO gives larger semantic gains than DPO but increases missing subfields: Llama-IPO produces 2,147 missing subfields compared with 1,340 for the base model and 1,481 for SFT; Qwen-IPO produces 1,743 missing subfields compared with 1,088 for the base model and 1,276 for SFT. This suggests that sequence-level preference optimization can reward recovering the correct plan item

	labs		imaging		monitor.		medical		conservative		follow_up		referral		pat._edu.		wts_care		Overall		Parsed.	Miss.I	Miss.S	Hall.I	Hall.S	Misp.
Model	S	M	S	M	S	M	S	M	S	M	S	M	S	M	S	M	S	M	S	M						
Llama-3.1-8B family																										
Base	.109	.116	.521	.545	.066	.095	.461	.558	.073	.123	.553	.788	.531	.722	.611	.686	.226	.467	.350	.456	486	1706	1340	1791	1912	148
+DPO	.149	.167	.561	.583	.345	.357	.491	.548	.228	.294	.611	.714	.590	.685	.729	.753	.446	.533	.461	.517	479	**1006**	1332	1397	1397	**59**
+SFT	.474	.501	.750	.777	.759	.803	.652	.699	.256	.293	.832	.884	.730	.783	.730	.732	.614	.644	**.644**	**.680**	494	1661	1481	**570**	**129**	71
+IPO	.376	.391	.589	.609	.649	.685	.533	.583	.226	.364	.746	.832	.650	.741	.738	.759	.592	.653	.566	.624	478	1209	2147	818	429	141
+ORPO	.269	.294	.659	.680	.740	.680	.612	.657	.215	.268	.786	.843	.641	.704	.706	.710	.566	.596	.577	.614	480	1572	**1301**	641	153	162
Qwen2.5-7B family																										
Base	.324	.340	.598	.622	.439	.474	.503	.600	.127	.166	.581	.815	.533	.707	.744	.788	.513	.647	.485	.573	488	1752	1088	1043	2084	60
+DPO	.363	.389	.613	.636	.566	.591	.525	.594	.238	.290	.629	.775	.548	.760	.778	.793	.565	.644	.545	.602	479	**1034**	1082	834	1535	**22**
+SFT	.680	.711	.686	.705	.755	.790	.640	.688	.259	.289	.815	.867	.583	.620	.779	.792	.604	.640	**.644**	**.678**	498	1877	1276	**396**	**184**	68
+IPO	.585	.605	.624	.645	.704	.737	.555	.616	.237	.341	.743	.839	.563	.681	.781	.793	.599	.639	.602	.652	478	1242	1743	527	504	57
+ORPO	.481	.511	.651	.670	.746	.735	.612	.662	.229	.271	.776	.845	.561	.734	.772	.790	.593	.642	.608	.647	480	1614	**1056**	434	210	66

Table 27: Per-field and overall performance of base and fine-tuned variants on the 500-case test set. S = chrF; M = MoverScore. Overall columns report aggregate chrF and MoverScore point estimates under the same evaluation protocol as the main benchmark. Miss.I/S = missing item/subfield counts. Hall.I/S = hallucinated item/subfield counts. Misp. = misplacement flags. Bold = best reported value per family. Base models are reproduced from the main leaderboard for comparison. The DPO rows are fully reported to align the appendix with the four-objective fine-tuning setup described in §6.2.

while leaving attributes such as dose, route, frequency, follow-up aim, or referral purpose unspecified.

ORPO gives a better balance between hallucination reduction and subfield completeness. Llama-ORPO reduces hallucinated items from 1,791 to 641 while keeping missing subfields below the base model. Qwen-ORPO reduces hallucinated items from 1,043 to 434 and keeps missing subfields close to the base level. However, ORPO still underperforms SFT in overall schema fidelity. These results motivate reporting item-level hallucination, subfield omission, and misplacement counts in addition to aggregate similarity metrics.

H.6 Conservative Field Under Fine-tuning

The `conservative` field remains difficult after fine-tuning. The full per-field breakdown shows that conservative-field MoverScore reaches 0.294 for Llama-DPO, 0.293 for Llama-SFT, 0.364 for Llama-IPO, and 0.268 for Llama-ORPO, compared with 0.123 for the Llama base model. For Qwen, the corresponding values are 0.290 for DPO, 0.289 for SFT, 0.341 for IPO, and 0.271 for ORPO, compared with 0.166 for the Qwen base model. Although all fine-tuned variants improve over the corresponding base models, conservative remains near the bottom of the field-level distribution after fine-tuning.

This result is consistent with the main error analysis. The hardest cases require distinguishing clinician-directed conservative management, such as splinting or procedural intervention, from patient-directed lifestyle or behavioral advice, such as rest, diet, or activity modification. The persistence of this error after task-specific supervision suggests that this boundary is a schema-grounding problem rather than only a data-size or hallucination problem.

I Extraction Prompt

The following guidelines are passed verbatim as part of the user message to every evaluated model. The required JSON shape template is appended immediately after the guidelines and before the clinical dialogue.

Listing 1: Extraction guidelines passed to all models.

```
## Task
Given a doctor-patient clinical conversation,
     extract the current
treatment/management plan into the following
     categories: investigations,
monitoring, treatment, follow_up,
     patient_education, referral,
when_to_seek_medical_care.

## Core Rules
1. Dialogue-only: Extract only what is explicitly
      stated. Do NOT infer.
2. Exact wording: Copy phrasing exactly,
     including numerals and timing.
3. Concise but faithful: Remove only filler;
     never paraphrase.
4. Multiple slots: If one item belongs to
     multiple slots, add to all.
5. Deduplicate: Include repeated content once,
     keeping most detailed.
6. Split combined advice into separate items
     unless interdependent.

## Medication Rules
- Do not infer missing fields (especially dosage
     form or drug class).
- condition_treated: fill only if dialogue
     explicitly links drug to
  condition.

## Conservative Treatment
- IV fluids and oral rehydration -> Conservative,
      not Medical.
- conservative_method is a list of strings.

## Complementary Therapies, External Equipment,
     Prevention
- Plain lists of STRINGS only. NEVER use
     medication objects or dicts.
- CORRECT:   "Prevention": ["Flu vaccine", "Avoid
      prolonged sun exposure"]
- INCORRECT: "Prevention": [{"name": "Flu vaccine
     ", "dose": null, ...}]
```

```
## Investigations / Labs
- Use the most specific test name stated.
- Exclude already-done or currently-awaited
    investigations.
- Monitoring parameters (e.g., blood glucose, BP)
     -> Monitoring.

## Referral
- Only for specialist care, consultation,
    counseling, or programs.
- Routing to a lab/test -> Investigations, not
    Referral.

## Follow-up
- FOLLOW_UP is a list of follow-up objects.
- Append any conditions/triggers into
    scheduled_follow_up_time text.

## Patient Education
- Only if dialogue explicitly mentions materials:
     pamphlets, handouts,
  videos, written instructions. "Discuss/talk
      about" does NOT qualify.

## When to Seek Care
- List specific symptoms or situations explicitly
     stated.

## Exclude
- Future-dependent actions ("once we get results,
     we'll refer you").
- Generic instructions ("complete your course", "
    take your medication").

## Output
- Null for any field not explicitly stated.
- Empty array [] if a list category has no items.
- Return ONLY the JSON object -- no commentary,
    no markdown fences.
```

The required JSON shape template appended after the guidelines is as follows:

Listing 2: Required JSON shape template appended to all prompts.

```
{
  "INVESTIGATIONS": {
    "Labs": [],
    "Imaging": [],
    "Nerve_Muscle_Conduction_Studies": [],
    "Tissue_Sampling": []
  },
  "MONITORING": [],
  "TREATMENT": {
    "Prevention": [],
    "Conservative": [],
    "Medical": [],
    "Surgical": [],
    "COMPLEMENTARY_THERAPIES": [],
    "EXTERNAL_EQUIPMENT": []
  },
  "FOLLOW_UP": [],
  "PATIENT_EDUCATION": null,
  "REFERRAL": [],
  "WHEN_TO_SEEK_MEDICAL_CARE": null
}
```

J Additional Experimental Parameters

This appendix reports implementation details not already specified in the main text or reproducibility appendix. All fine-tuning experiments use LoRA with rank $r = 16$, scaling $\alpha = 32$, dropout 0.05, bfloat16 precision, gradient checkpointing, cosine learning-rate scheduling, warmup ratio 0.1, weight decay 0.01, gradient clipping at norm 1.0, maximum sequence length 4096, random seed 42, and three training epochs. LoRA adapters are applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`.

Objective	LR	Batch	Objective-specific setting
SFT	2×10^{-5}	8	Cross-entropy on assistant JSON output
DPO	5×10^{-7}	8	$\beta = 0.1$; frozen reference model
IPO	5×10^{-7}	8	$\beta = 0.1$
ORPO	8×10^{-6}	8	$\lambda = 0.1$

Table 28: Objective-specific parameters for the fine-tuning experiments.

For preference optimization, we construct 2,181 chosen–rejected pairs from the CAPE training split. Chosen responses are the physician-authored reference treatment plans. Rejected responses are model-generated outputs whose chrF scores fall below the 20th percentile on the training split. DPO, IPO, and ORPO are initialized from the corresponding base instruction-tuned checkpoints rather than from the SFT checkpoints.

K Reproducibility

K.1 Reproducibility of model rankings

Proprietary and API-accessed models were evaluated using the exact API identifiers listed in Table 6. Because provider-hosted general-purpose models are updated by their providers without stable version guarantees, results may not be exactly reproducible with later API calls to the same endpoint string. Open-source models evaluated via HuggingFace are pinned to their published checkpoint identifiers, making those results more reproducible. Rankings among frontier and API-accessed models should therefore be interpreted as tied to the specific identifiers and access conditions reported in Table 6, rather than as timeless properties of the model families. CAPE provides a reusable test set to enable future re-evaluation as models evolve.

K.2 Data release

CAPE will be released as an open-source research resource to support reproducible evaluation. The release will include the benchmark data, fixed train/test split, physician-authored reference plans, test-set review metadata, JSON schema, controlled vocabularies, annotation guidelines, prompt templates, parsing and normalization code, metric implementations, structured-error counting scripts, and raw model outputs used in the paper. For open-weight models, we report exact checkpoint

identifiers; for provider-hosted models, we report endpoint identifiers, access route, and evaluation date, and release the generated outputs to preserve score reproducibility despite possible API drift. All reported aggregate results are computed on the fixed 500-case test set using the same field-level aggregation protocol, with confidence intervals and paired comparisons estimated by bootstrap over test cases. The CAPE dataset, schema, scoring code, and leaderboard will be versioned so that future submissions can be compared under a fixed and auditable protocol.

L Annotator Compensation

The CAPE annotations were produced by medically trained annotators: ten resident physicians with outpatient documentation experience, supervised by two attending clinicians. Annotators completed training and qualification before production annotation, and the test set was further reviewed for quality control. Each annotator was compensated at a fixed rate of USD 8 per hour for annotation-related work. This rate was applied uniformly across annotators and covered training, qualification, production annotation, and review tasks.

M Compute Infrastructure

Local open-source model inference and fine-tuning experiments were run on a workstation/server equipped with 4× NVIDIA A100 80GB GPUs, 1 TB system RAM, and 2× AMD EPYC-class CPUs, running Ubuntu Linux with CUDA-enabled PyTorch. This configuration was sufficient for inference over the evaluated 7B–90B open-weight models using tensor parallelism or quantized loading where needed, and for LoRA fine-tuning of the 7B–8B models. Closed-source and provider-hosted models were accessed through their respective APIs and therefore were not run on the local hardware.