CODE, NOT CHAINS-OF-THOUGHT: LOCALIZING GUIDELINE-ADHERENCE RISK BY COMPILING GUIDELINES INTO MEDCODE

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical large language models can cite guidelines fluently while producing unsupported dosages, staging criteria, or treatment recommendations. We argue that this failure is partly architectural: when the same model both interprets evidence and emits the clinical action, there is no machine-checkable contract between a recommendation and the source guideline. We present **MEDCODE**, a model-agnostic framework that compiles oncology guidelines into executable, citation-bound Python rules and restricts the LLM to structured patient-state extraction. The rule engine, rather than the decoder, maps typed evidence to a closed `ClinicalAction` space with bounded verbatim citations and explicit abstention when required evidence is missing. This design does not eliminate clinical error; residual failures can still arise from extraction or rule authoring. Instead, it localizes guideline-adherence risk to two auditable choke-points, enforced by six static compliance gates and localized rule diffs under guideline updates. We instantiate MEDCODE on oncology staging and treatment tasks using NCCN-derived rule scaffolds and real clinical cohorts. On TNM staging, deterministic execution substantially improves smaller and open-weight models: Qwen3.6-27B rises from 25.95 to 54.91 TNM accuracy on LungCURE-EN, and Gemma-4-E4B-it rises from 17.31 to 58.70 on EsophagusPKUCH. On treatment recommendation, MEDCODE improves coverage of physician-specified gold-plan items across all evaluated models, although this rubric measures coverage rather than harmful or extraneous recommendations. Finally, the same executable rules serve as a verifiable GRPO reward for MEDCODE-LM: distilling a Qwen3.6-27B teacher into a Qwen3.5-4B student raises TNM accuracy from 35.64 to 50.34, approaching the teacher's 52.67 while preserving auditable decision semantics.

1 INTRODUCTION

Modern clinical large language models (LLMs) achieve strong performance on USMLE-style benchmarks (Singhal et al., 2023) yet remain unsafe at the bedside, where the failure mode of interest is not multiple-choice accuracy but *guideline-faithful action*. A model that prescribes capecitabine $1{,}000\,\mathrm{mg/m^2}$ "per NCCN" when the cited regimen actually specifies $1{,}250\,\mathrm{mg/m^2}$ is producing an unsupported recommendation indistinguishable from a correct citation in fluent prose.

Three properties of the dominant LLM-end-to-end design make such guideline-deviation risks structurally hard to audit: (i) the decision path is a string of natural-language tokens with no machine-checkable contract to the source guideline; (ii) every recommendation is a fresh generation, so the same patient asked twice may receive two different dosages; and (iii) when a guideline edition is updated, the implied behavioral change is delocalized across model weights, not isolated in a diff. *CancerGUIDE* (Yang et al., 2025), the current state-of-the-art guideline-faithful LLM trained on NSCLC, places the LLM at the decision layer and reports a Treatment-Match Accuracy ceiling of 36.4% on real patient cases. We read this as consistent with an *architectural* gap: even a strong CoT decoder reasoning over guideline PDFs lacks the static contract needed to make the three properties above inspectable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

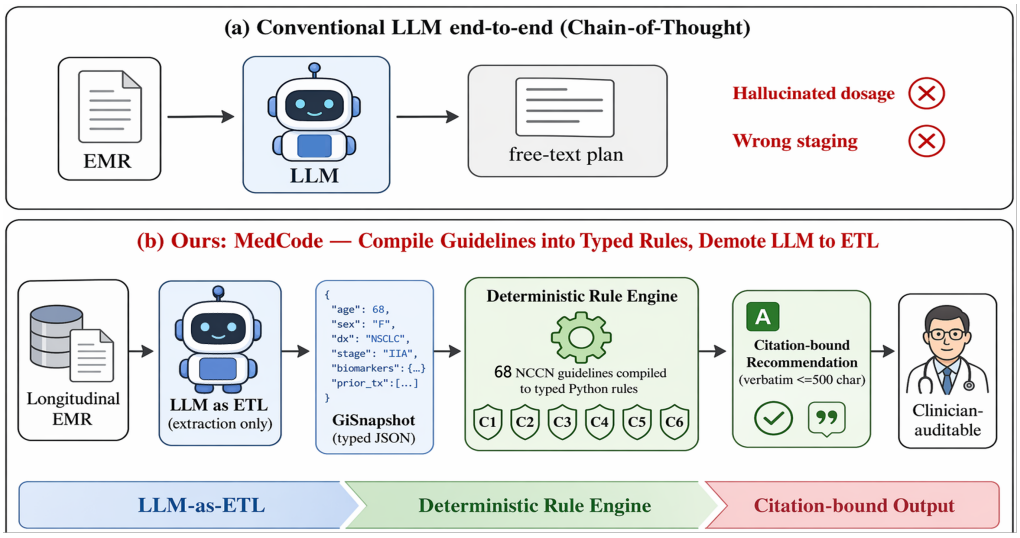

Figure 1: The MEDCODE pipeline. **(a)** The LLM-end-to-end design puts a decoder in the decision path, emitting a free-text plan whose dosages and staging may be unsupported or inconsistent with the cited guideline. **(b)** MEDCODE inverts this: guidelines are compiled into typed rules, the LLM is restricted to structured extraction, and a deterministic engine dispatches the recommendation.

Our thesis is to move the LLM out of the decision path entirely: compile the guideline into typed, registered Python rules that satisfy a static-compliance contract, and constrain the LLM to *extract-transform-load* (ETL) over the patient timeline. We do not claim that this makes clinical decision support error-free. Rather, it changes where errors can occur: instead of allowing a model to synthesize treatment decisions in free text, it forces the decision through executable rules and exposes the remaining failure modes as extraction errors or rule-authoring errors. We instantiate this as MEDCODE, an open scaffold with catalog entries for all 68 NCCN tumor-specific guidelines (National Comprehensive Cancer Network, 2026); this is source-document and schema coverage, not a claim that every guideline page or decision branch has been authored as an executable rule or evaluated. The contributions of this paper are:

1. **A typed, model-agnostic framework** (see *The MedCode Framework*) with six static gates, a closed `ClinicalAction` sum-type, and verbatim-citation enforcement that refuses to ship if the rule corpus violates any gate. Because the engine decides rather than the language model, it also exposes a first-class *abstention* behavior—returning no answer when the typed evidence is insufficient—that an LLM-end-to-end system cannot meaningfully emit.
2. **Externalizing guideline knowledge from weights into code** (see *Evaluation*): TNM staging and treatment-plan coverage are two instances of one mechanism—the guideline mapping is *executed* by the engine, not *recalled* from weights.
3. **A code-native training pipeline** (see *Toward a Medical Code Model*) that uses the rule engine itself as a verifiable GRPO reward (Shao et al., 2024): distilling a 27B teacher into a $6\times$ smaller, deployable student brings the small model close to a frontier proprietary model on the staging task, and the same rules serve as both the runtime safety substrate and the training corpus for a future code-pretrained medical foundation model.

2 RELATED WORK

Structured representation and knowledge grounding. The dominant strategy keeps the LLM generating clinical text but constrains it with external structure: converting longitudinal EHRs into structured form (Xu et al., 2024; Khatib et al., 2025; Wang et al., 2025), and grounding generation in knowledge graphs, from standard RAG (Lewis et al., 2020) to graph-aware RAG (Zhao et al., 2025a; Mavromatis & Karypis, 2025) to KG–LLM alignment (Jiang et al., 2024). *CliCARE* (Li

et al., 2026) is the high-water mark, aligning patient-specific temporal KGs to a guideline KG before the LLM writes its recommendation. In oncology, *CancerGUIDE* (Yang et al., 2025) imitates NCCN reasoning end-to-end on NSCLC, while *MedKP* (Wang et al., 2024a), *MedGUIDE* (Liu et al., 2025), *CDR-Agent* (Chen et al., 2025), and Gupta et al. (2024) couple retrieval with guideline-graph reasoning; *GuideBench* (Diao et al., 2025) and *Med-PRM* (Yun et al., 2025) push grounding into evaluation and process rewards. However deep the grounding, the LLM still emits the decision as free text with no machine-checkable contract to the source—CancerGUIDE's 36.4% Treatment-Match result is consistent with, though not by itself proof of, an architectural limitation in free-form decision generation.

Reliable evaluation of clinical generation. Because lexical metrics fail to capture clinical validity (Wang et al., 2023), evaluation has shifted to LLM-as-a-Judge (Zheng et al., 2023), complemented by hallucination benchmarks (Pal et al., 2023; Singhal et al., 2023) and the intrinsic-vs-extrinsic taxonomy we adopt (Ji et al., 2023). But an LLM judge is a post-hoc, itself-biased proxy: a survey (Gu et al., 2025) catalogs persistent failure modes—positional/verbosity/self-enhancement bias (Wang et al., 2024b) and vulnerability to trivial manipulation (Zheng et al., 2024; Zhao et al., 2025b)—and it offers no way to express that the system *should* stay silent: an always-decoding model has no formal abstention operator.

The shared premise, and our inversion. Both classes presuppose the LLM must decide and must speak, and so can only bolt grounding onto generation or a soft proxy onto evaluation. MEDCODE rejects the premise: the compiled decision path is deterministic and machine-checkable (answering the grounding class) and the rule engine itself replaces the soft judge as a verifiable reward, while a first-class *abstention* fires when the typed evidence is insufficient (answering the evaluation class). This connects to two further threads. *Computable-guideline DSLs*—GLIF, Arden Syntax (Hripcsak, 1994), EON, and PROforma (Sutton & Fox, 2003)—have decades of history (Peleg, 2013) but lack modern toolchains: no unit tests, no static gates, no LLM front-end mediating free-text EMRs; *Computable Clinical Guidelines: A Survey* (Staniek et al., 2025) identifies the citation-binding gap motivating gates C2/C5. And *code-for-reasoning* work shows code pretraining transfers to reasoning (Chen et al., 2021; Madaan et al., 2022) and that executable decision paths beat prose (Gao et al., 2023; Chen et al., 2023); we extend this from inference to training. Finally, selective prediction in healthcare (Abbasi-Yadkori et al., 2024; Mukherjee et al., 2024) formalizes the abstention our rule engine operationalizes.

3 THE MEDCODE FRAMEWORK

Figure 1 contrasts the generate-and-hope paradigm—where a chain-of-thought decoder can bury a $1{,}000$ vs. $1{,}250\,\mathrm{mg/m^2}$ dosing error in fluent prose—with MEDCODE, which routes the decision through a deterministic, citation-bound rule engine. MEDCODE rests on three design commitments, shown in Figure 1(b). (i) *Actions are a closed type*: every recommendation is an immutable record whose action is drawn from a fixed sum-type, so no free-text plan can enter the decision path. (ii) *Rules are statically verified code*: each guideline compiles to a flat-guard Python rule that must pass six compliance gates before shipping. (iii) *The corpus is a single source of truth*: all guidelines are materialized from one catalog, so a guideline update is a localized diff rather than a re-training. Together these make the decision path deterministic, citation-bound, and machine-checkable.

3.1 CLOSED SUM-TYPE OF CLINICAL ACTIONS

A `Recommendation` is an immutable record carrying (a) a `rule_id` traceable to a single source-of-truth rule, (b) an `action` drawn from a closed `Union` of frozen dataclasses (`ScreeningLowDoseCT`, `WorkupImaging`, `WorkupBiopsy`, `SurgeryAction`, `SystemicChemo`, `SystemicTargeted`, `SystemicImmunotherapy`, `Surveillance`, `Referral`, `Counseling`; visualized in the **SUPPLEMENTARY** material), (c) an NCCN strength category (`Category1`, `Category2A`, `Category2B`, `Category3`), (d) the `factors` observed on the patient that triggered the rule, and (e) a tuple of `SourceCitation` objects each bounded to ≤ 500 characters of verbatim quotation. Closing the action type forbids hallucinated free-text actions: a string like "consider best supportive care" cannot enter a position the contract types as `Counseling`.

Gate	Property checked
C1	No nested control flow inside `rule_*`; one observation, one decision.
C2	Every `verbatim_quote` is ≤ 500 characters.
C3	Every `rule_id` has 1–10 unit tests, positive and negative.
C4	Docstrings carry **Source/Factors/Action** triad.
C5	Each `verbatim_quote` is mirrored in a JSON IR.
C6	Catalog, `GuidelineId` enum, and filesystem layout are three-way consistent.

Table 1: The six static gates; the framework refuses to ship if any fails on the rule corpus.

3.2 Six static compliance gates

The gates form a fixed pre-ship checklist, analogous to the clerical checks an expense claim clears before reimbursement: they enforce that every rule is flat and readable (C1), cites a short, mirrored source (C2, C5), carries its own tests (C3) and a Source/Factors/Action docstring (C4), and lives in a registry consistent across catalog, enum, and filesystem (C6). They certify *structure and traceability*, not clinical correctness—which remains the responsibility of the authored tests and the human-checkable citation.

3.3 Rule shape

Each rule is a function decorated with `@register`, named with a stable `rule_id`, returning either a `Recommendation` or `None`. Its body is a *flat guard*: a sequence of `return None` short-circuits (wrong cancer, wrong stage, not treatment-naive, ...) followed by a single `Recommendation` construction that names the action, its NCCN strength category, the triggering factors, and the source citation. This is the shape gate C1 enforces—one observation, one decision, no nested control flow. A representative gastric treatment rule, shown exactly as it ships with its docstring (checked by C4) and its verbatim `SourceCitation` (bounded by C2/C5), appears in the **SUPPLEMENTARY** material (§"Full Rule Listing"). The shape is deliberately easier to audit than to express: a clinician reads the guards and the citation, not a control-flow graph.

3.4 Catalog of all 68 cancers

A single catalog is the source of truth for all guidelines: it lists each one as a `Guideline(slug, enum_name, enum_value, display, category, published_under, status)` record, and scaffolds all 68 NCCN tumor types with zero copyrighted content. The private licensed source corpus used for development contains the corresponding 68 NCCN tumor-specific PDF files, so source-document coverage is complete at the file level. The `GuidelineId` enum is materialized from the catalog at runtime, and gate C6 checks three-way consistency across catalog, enum, and filesystem, so adding or updating a guideline is a localized diff rather than a re-training. Institutions with NCCN licenses fill in the verbatim-quote fields locally; the framework verifies the strings as data, never asks the LLM to recall them.

4 System: LLM-as-ETL + Deterministic Engine

We instantiate the system on a gastrointestinal (GI) cancer cohort; the NSCLC staging track in the *Evaluation* section shares the same engine. The flow separates a deterministic data-preparation path from a single LLM call:

```
patient.csv -> read_timeline
            -> extract_drug_orders
            -> bucket_lines -> actual_line

narrative   -> LLM-as-ETL -> GiSnapshot
            -> rule_engine
            -> Recommendation* -> metrics
```

Coverage layer	What is covered in this submission
Source PDFs	68/68 NCCN tumor-specific guideline PDFs in the private licensed corpus.
Public scaffold	68/68 catalog entries and filesystem/enum slots; copyrighted text is not redistributed.
Authored rules	Runnable NCCN rules are reported for the NSCLC task family evaluated here; other catalog entries are scaffold placeholders until locally authored and reviewed.
Evaluation	Lung cancer TNM/treatment cohorts plus esophageal TNM; these benchmarks do not cover the full content of all NCCN guidelines.
Overall content coverage	Not claimed as a page-, node-, or branch-level proportion; measuring full-guideline content coverage is future audit work.

Table 2: Scope accounting for MEDCODE. "68/68" refers to source-document and scaffold coverage, not exhaustive executable-rule or evaluation coverage of every NCCN recommendation.

Dataset	Model	No guide	Prompt	MEDCODE	Δ no	Δ prompt
LungCURE-EN	GPT-5.5	53.60	52.34	52.64	−0.96	+0.30
LungCURE-EN	DeepSeek-V4-Pro	20.31	26.02	47.38	+27.07	+21.36
LungCURE-EN	Qwen3.6-27B	25.95	35.91	54.91	+28.96	+19.00
LungCURE-EN	Gemma-4-31B-it	38.27	39.84	56.36	+18.09	+16.52
EsophagusPKUCH	GPT-5.5	72.12	75.90	68.27	−3.85	−7.63
EsophagusPKUCH	Qwen3.6-27B	46.15	64.42	71.20	+25.05	+6.78
EsophagusPKUCH	Gemma-4-12B-it	13.46	59.62	67.31	+53.85	+7.69
EsophagusPKUCH	Gemma-4-E4B-it	17.31	34.62	58.70	+41.39	+24.08

Table 3: Summary TNM accuracy for representative models. The main paper reports joint TNM accuracy only; full per-model and per-axis T/N/M results are in the supplementary material.

The timeline and drug-order stages are pure, deterministic Python. The LLM is invoked *only* at the structured-extraction step, behind a client interface whose default implementation is a deterministic mock, so the entire pipeline—including its test suite—runs without an API key. This isolates the one non-deterministic component and makes the decision path reproducible. Two mechanics that keep this path auditable—how a new cancer family is added as a typed `GiSnapshot` plus a thin adapter (no engine rewrite), and how the regimen recognizer maps drug orders to canonical regimens while returning *unrecognized* rather than a spurious nearest match—are detailed in the **SUPPLEMENTARY** material.

5 EVALUATION

5.1 BENCHMARKS

The *LungCURE* dataset. *LungCURE* (Hao et al., 2026) is the first standardized multimodal benchmark for lung-cancer decision support: 1,000 real-world cases from >10 Chinese hospitals (2019–2025), each consolidating imaging, pathology, clinical, and genomic reports into one de-identified case document across the major histologies and stages I–IV. Its gold standard is set by board-certified thoracic oncologists under a two-stage protocol (TNM staging then treatment planning) against AJCC 8th edition. We evaluate on the released 1,000-case set unchanged.

The *LungFAHGMU* dataset. *LungFAHGMU* is a real-world lung-cancer cohort we curated in-house at the First Affiliated Hospital of Guangzhou Medical University (2021–2026), de-identified, retaining only patients with both an established TNM conclusion and the documentation that substantiates it. Unlike *LungCURE*'s consolidated case documents, it consists of raw clinical narratives, and its role is to test *generalization*: strong performance on both cohorts indicates a model is not overfit to *LungCURE*'s house style. The per-model TNM results mirror Table 3 and are deferred to

Method	① No guide	② Typed, no guide	③ w/ MEDCODE	Δ_{typed}	Δ_{engine}	Δ_{total}
GPT-5.5	72.12	66.35	68.27	-5.77	$+1.92$	-3.85
Qwen3.6-27B	46.15	45.88	71.20	-0.27	$+25.32$	$+25.05$
Gemma-4-12B-it	13.46	24.04	67.31	$+10.58$	$+43.27$	$+53.85$

Table 4: Per-model ablation on EsophagusPKUCH (TNM-Acc). Columns ①/③ reuse the No-guide and w/ MEDCODE TNM columns summarized in Table 3; ② is the added run—the LLM fills the same typed slots from its *internal knowledge, with no guideline in the prompt*, isolating how much of the mapping the model already holds (Δ_{typed}=②−①). Δ_{engine}=③−② bundles supplying the externalized guideline mapping and executing it deterministically; Table 3's Guideline-in-prompt vs. MEDCODE contrast isolates the pure-execution effect alone.

Method / model	GPT-5.5	GLM-5.2	DeepSeek-V4-Pro	Qwen3.5-397B-A17B	Gemma-4-31B-it	Qwen3.6-27B
Guideline in prompt	69.67±1.41	61.18±1.58	53.45±1.28	58.00±1.71	47.75±1.50	59.27±2.34
w/ MEDCODE	75.31±0.90	71.32±1.27	67.68±2.59	64.68±0.69	53.75±1.10	69.67±0.29
Δ rubric-score	+5.64	+10.14	+14.23	+6.68	+6.00	+10.40

Table 5: Treatment-plan item coverage (Rubric Score, 0–100) on the *LungFAHGMU* treatment subset (202 real cases), each model prompting the guideline directly (*Guideline in prompt*) vs. constrained by MEDCODE. Cells are mean±sd over three random seeds; the Δ row is the mean per-seed gain. The Rubric Score is defined in *Metrics*; it measures coverage of gold-plan items and is not a direct penalty for extraneous or harmful recommendations.

the **SUPPLEMENTARY** material (§"LungFAHGMU"). We additionally filter it to the most recent year and to guideline-conformant cases, yielding a 202-case subset for the Rubric-Score evaluation below.

The *EsophagusPKUCH* dataset. *EsophagusPKUCH* is a real-world esophageal-cancer TNM cohort from Peking University Cancer Hospital: 104 clinically staged cases built from raw clinical narratives, each with a T/N/M conclusion traceable to line-level evidence and passing an automated gold-versus-evidence consistency check. Spanning stages I–IVB (majority III–IV), it extends the TNM evaluation to a second anatomy under authentic documentation noise.

5.2 METRICS

TNM staging. Across all TNM benchmarks we adopt the original *LungCURE TNM Staging Accuracy* definition, under which a prediction is correct *iff* its T, N, and M components are all correct.

Treatment recommendation. We score treatment recommendations with the *Rubric Score* defined below. We use this score as a coverage metric for whether a recommendation includes physician-specified gold-plan items, not as a complete hallucination or safety metric. Beyond point accuracy, a property unique to our architecture is that *silence is a first-class output*: when the typed slots that a rule guards on are not supported by the record, the engine returns no recommendation rather than a fabricated one. An LLM-end-to-end system has no formal silence operator and must always emit text, so this abstention behavior has no analogue there; we illustrate it concretely on the TNM task with a worked case in the **SUPPLEMENTARY** material (§"Abstention on Insufficient Evidence"), where a deterministic staging aggregator abstains on an axis whose evidence is insufficient instead of guessing.

Rubric score. For each case, a physician-authored gold plan is decomposed into a checklist of discrete treatment items, each labeled as *key* (weight 2) or *minor* (weight 1). A judge model (GPT-5.4) determines, for every rubric item, whether the model's recommendation covers it. The per-case score is the weight-normalized coverage $s = \frac{2\,h_{\text{key}}+h_{\text{minor}}}{2\,N_{\text{key}}+N_{\text{minor}}}$, where N_{key} and N_{minor} are the numbers of key and minor items, and h_{key}, h_{minor} are the corresponding weighted hits (covered $= 1$, partially covered $= 0.5$, missed $= 0$). The reported *Rubric Score* is the arithmetic mean of s over all cases, averaged over three random seeds. This metric deliberately does *not* score action–citation entail-

ment, penalize unsupported extra treatments, or grade the clinical severity of wrong, conflicting, or harmful recommendations; therefore Table 5 supports a claim about improved gold-plan coverage, while the paper's hallucination-risk claim rests on the architectural contract and qualitative abstention analysis rather than on this coverage score alone.

5.3 Results on TNM Staging

Table 3 summarizes joint TNM accuracy for representative models under three conditions—No guide (parametric memory), Guideline in prompt, and MedCode (typed TNM extraction)—on the public *LungCURE* benchmark and the real-world *EsophagusPKUCH* cohort. Full per-model and per-axis T/N/M results are reported in the **SUPPLEMENTARY** material.

The effect is strongly *model-dependent*, and reading it honestly is central to our claim. On the small and open-weight models the gains are large and consistent: relative to No guide, MedCode lifts TNM-Acc by up to $+53.9$ (Gemma-4-12B) and $+37.5$ (Qwen3.5-9B) on esophagus and by $+29.0$ (Qwen3.6-27B) on lung, with the improvement concentrated on the T and N axes where free-text staging is most error-prone. The two strongest models are the exception: GPT-5.5 and Qwen3.5-397B are already near their ceiling under prompting and move slightly negative under MedCode (-0.96 and -8.91 on lung; -3.85 and -2.89 on esophagus). We read this mechanistically rather than as MedCode "hurting" capable models: frontier models have largely memorized the staging thresholds, so externalizing them yields little, whereas smaller models have not, so the compiled mapping is pure gain—the engine contributes exactly where the mapping knowledge is missing from weights. The practical consequence is that MedCode narrows the spread across model families: the weakest models are pulled up to within a few points of the strongest, so a small open-weight model routed through the engine reaches staging accuracy that otherwise requires a frontier model. Table 4 decomposes this—type-only extraction (Δ_{typed}) helps only smaller models, while the externalized, engine-executed mapping (Δ_{engine}) is where their gains concentrate.

5.4 Results on Treatment Recommendation

We next evaluate whether the MedCode representation improves coverage of physician-specified treatment-plan items across model families. Table 5 scores six proprietary and open-weight models on the *LungFAHGMU* treatment subset (the 202 real cases defined above) under the Rubric Score defined above, comparing each model prompting the guideline directly against the same model constrained by MedCode. Every model improves under MedCode—by $+5.6$ (GPT-5.5) to $+14.2$ (DeepSeek-V4-Pro) rubric points, with the largest gains on the smaller open-weight models—indicating that the typed, engine-dispatched decision path improves gold-item coverage across model families. Because the rubric is coverage-only, these results should not be read as measuring reductions in harmful or extraneous recommendations.

Case study: where the gain comes from. On the same stage-IIIB lung-adenocarcinoma case, direct prompting promotes pemetrexed+cisplatin to "first choice" and demotes weekly carboplatin+paclitaxel to fallback, even though the guideline lists these concurrent regimens as co-equal `category 2A` options. Routed through MedCode, the engine preserves this equivalence class and covers the gold item ($0.80 \rightarrow 1.00$), with strength tags and provenance fixed by the rule. This illustrates a coverage gain, not a general harmful-content metric; the full side-by-side example is in the **SUPPLEMENTARY** material.

6 Toward a Medical Code Model: MedCode-LM

The same code-as-knowledge representation that anchors runtime safety also serves as a training corpus. We post-train a Qwen3.5-4B base model into MedCode-LM along three pillars:

Pillar 1: ETL distillation. A LoRA SFT recipe maps free-text Chinese EMR timelines to `GiSnapshot` JSON. Training pairs are generated by running the deterministic harness with a frontier teacher LLM and keeping only those snapshots whose rule-engine output matches a clinician-reviewed reference. We additionally apply on-policy distillation (OPD) to align the student's inter-

Model / stage	T	N	M	TNM
Student (Qwen3.5-4B)				
base	65.05	67.82	83.04	35.64
+ ETL SFT	76.33	76.67	83.67	49.33
+ GRPO (rule rwd.)	76.33±0.34	75.89±0.84	83.45±0.39	50.34±0.58
Δ vs. base	+11.28	+8.07	+0.41	+14.70
References (w/ MEDCODE)				
GPT-5.5 (frontier)	79.93	74.25	83.61	53.85
Qwen3.6-27B teacher	74.33	81.00	85.33	52.67

Table 6: Distilling the Qwen3.6-27B teacher into a $6\times$ smaller Qwen3.5-4B student: TNM accuracy (%). All rows are measured on the same *LungCURE-EN* test split (n=300, seed 42); the GRPO row is mean±sd over seeds at the best checkpoint, step 100.

mediate extractions with the teacher's, reducing extraction drift that would otherwise propagate into wrong rule firings.

Pillar 2: Rubric-judge SFT. A second LoRA recipe trains the model to score recommendations along a five-axis clinical rubric (citation faithfulness, factor coverage, contraindication safety, abstention appropriateness, regimen specificity). The rubric model is then used as a learned-judge inside the GRPO loop.

Pillar 3: GRPO with verifiable reward. The decisive pillar (G=8, β=0.04) sets the reward signal to a rule-engine-replay verifier $r = 0.4\, r_{\text{citation}} + 0.4\, r_{\text{action}} + 0.2\, r_{\text{abstain}}$, where r_{citation} is the verbatim-quote overlap between the model's generated citation and the rule's authored `verbatim_quote` (gate C2 in training-time use); r_{action} is the `ClinicalAction` sum-type equality between the model's output and the deterministic engine's output on the same input; r_{abstain} is whether the model abstained when the engine abstained. Because the verifier is the engine itself, the reward is bit-deterministic and free of human labeling.

What this looks like at scale. The data flywheel closes on three axes: (a) every additional authored rule increases the size of the GRPO training set by exactly the rules' coverage on the cohort; (b) every additional cohort patient extends the ETL distillation set; (c) the trained MEDCODE-LM can in turn propose candidate rule bodies for clinician review, lowering the marginal cost of authoring rule $k{+}1$.

The extraction backbone of MEDCODE-LM is evaluated on the TNM-staging task, where the rule engine supplies a fully verifiable reward. Table 6 distills a Qwen3.6-27B *teacher* (run with MEDCODE) into a $6\times$ smaller Qwen3.5-4B *student* on the *LungCURE-EN* test split under identical toolcall-format extraction. The effect of distillation is large: ETL distillation lifts the student's TNM accuracy from 35.6% to 49.3%, and GRPO with the rule-engine reward adds a further point to 50.3%—a $+14.7$ gain over the base model that closes most of the gap to the 52.7% teacher, at a size small enough for on-device deployment. The gains concentrate on the T and N axes, while the already-strong M axis is essentially unchanged.

The crucial design property—and the answer to "why train at all if the rules already decide?"—is generalization to inputs the rule corpus does not yet cover. Outside the GRPO support set MEDCODE-LM has no verifiable reward to fall back on, but the inductive bias it absorbs during GRPO is precisely the disposition to abstain on inputs that fail the upstream guards rather than to confabulate.

7 COMPARISON WITH STATE-OF-THE-ART

Decision-layer grounding. *CancerGUIDE* (Yang et al., 2025) keeps the LLM at the decision layer and reports 36.4% Treatment-Match on its NSCLC set. Our 40.9% *LungFAHGMU* number is not head-to-head because the case sets differ; the point is architectural. MEDCODE confines the LLM to extraction, derives actions through deterministic citation-bound rules, exposes first-class abstention, and updates behavior by rule diffs rather than retraining.

Prompt-time grounding. *LungCURE*'s LCAgent (Hao et al., 2026) routes guideline text into expert agents at inference time. MEDCODE instead compiles guidelines ahead of inference: the LLM fills typed slots, and the engine dispatches actions.

8 DISCUSSION

Localized, not eliminated, risk. MEDCODE does not make clinical decision support error-free: wrong extraction triggers the wrong rule, and a wrong authored rule is reproduced consistently. Its benefit is localization: failures are assigned to typed extraction or rule authoring, where they can be tested, audited, and updated independently. The LLM-as-ETL step still carries hallucination risk, bounded by JSON-shape validation, closed enums, and measurable extraction accuracy.

Evaluation scope. The treatment Rubric Score measures coverage of physician-specified gold-plan items; it does not score action–citation entailment, unsupported extra actions, or severity of harmful contradictions. Thus the treatment table supports a coverage claim, while broader safety claims require rule review, extraction audits, abstention analysis, and future negative-error metrics.

Guideline and rule coverage. The 68-guideline claim in Table 2 refers to source PDFs and scaffold entries, not exhaustive authored rules or evaluation over every NCCN page. The experiments stress lung-cancer staging/treatment and esophageal TNM, so conclusions apply to these covered tasks. Rule bodies and citations are model-drafted but verified line-by-line by a licensed reviewer; the static gates catch structure, while clinical entailment remains a human-review responsibility. Further limitations are in the **SUPPLEMENTARY** material.

9 ETHICS STATEMENT

All patient data are de-identified records from a single center; no intervention was performed, and raw patient text is never redistributed. Public benchmarks are used under their released terms. NCCN content is copyrighted: the public scaffold ships zero copyrighted strings, and `verbatim_quote` slots must be populated locally by an operator holding a current NCCN license. Runtime uses one extraction call plus a deterministic engine; only optional MEDCODE-LM post-training incurs a one-time training cost.

10 CONCLUSION

We argue that the bottleneck for guideline-faithful clinical AI is architectural: the LLM-end-to-end design places free-form generation in a decision path no static check can audit. Compiling the guideline into typed, citation-bound Python rules—and demoting the LLM to extract-transform-load—makes the decision path machine-checkable, deterministic given a fixed extracted snapshot, and diff-localizable. This does not eliminate clinical error; it localizes residual risk to extraction and rule authoring, where failures can be measured, reviewed, and repaired. TNM staging and treatment-plan coverage are two instances of one mechanism: the guideline's evidence-to-decision mapping is executed by the engine rather than recalled from weights, so the gain concentrates on models that have not internalized that mapping (small and open-weight) and vanishes on those that already have (frontier). The same rules then serve as a verifiable GRPO reward, distilling a 27B teacher into a $6\times$ smaller, on-device student—closing a flywheel from runtime safety to a future medical code foundation model.

11 AI USE DISCLOSURE

Generative AI tools were used only to aid and polish writing, including grammar, clarity, and phrasing. They were not used to generate scientific claims, conduct experiments, analyze data, or make clinical judgments for this paper. The authors reviewed and take full responsibility for all manuscript content.

REFERENCES

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, et al. To believe or not to believe your LLM: Iterative abstention via uncertainty estimation. *arXiv preprint arXiv:2406.02543*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wei Chen, Hao Li, Tianyi Zhang, Qi Wu, and Bowen Zhou. CDR-Agent: Clinical decision reasoning over guideline knowledge graphs with multi-step llm agents. *Journal of the American Medical Informatics Association*, 2025.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. GuideBench: Benchmarking domain-oriented guideline following for LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2505.11368.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 10764–10799, 2023.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2025.

Bhumika Gupta, Pralaypati Ta, Keerthi Ram, and Mohanasankar Sivaprakasam. Comprehensive modeling and question answering of cancer clinical practice guidelines using LLMs. In *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2024. arXiv:2501.13984.

Fangyu Hao, Jiayu Yang, Yifan Zhu, Zijun Yu, Jiawei Li, Shihao Li, Qicen Wu, Yulin Liu, Zhonghong Ou, Yunlong Wang, Xu Zeng, Meina Song, Guanting Chen, Mengyang Sun, Haoran Luo, Yu Shi, and Yingyi Wang. LungCURE: Benchmarking multimodal real-world clinical reasoning for precision lung cancer diagnosis and treatment. *arXiv preprint arXiv:2604.06925*, 2026.

George Hripcsak. Writing Arden Syntax medical logic modules. *Computers in Biology and Medicine*, 24(5):331–363, 1994.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Zhouyu Jiang, Xinrun Han, Zhixing Liu, and Maosong Sun. Efficient knowledge infusion via KG-LLM alignment. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.

Hassan S. Al Khatib, Sanjay Mittal, Shahram Rahimi, Nina Marhamati, and Sepideh Bozorgzad. From patient consultations to graphs: Leveraging LLMs for patient journey knowledge graph construction. *arXiv preprint arXiv:2503.16533*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Dongchen Li, Jitao Liang, Wei Li, Xiaoyu Wang, Longbing Cao, and Kun Yu. CliCARE: Grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)*, 2026. arXiv:2507.22533.

Peng Liu, Lei Zhao, Hao Sun, and Xin Wang. MedGUIDE: A primary-care clinical-guideline assistant with retrieval-augmented reasoning. *npj Digital Medicine*, 8:112, 2025.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*, 2022.

Costas Mavromatis and George Karypis. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 16682–16699, 2025.

Sandipan Mukherjee, Kenneth Holstein, and Percy Liang. Selective prediction in clinical decision support: Calibration and coverage tradeoffs. *Proceedings of Machine Learning for Healthcare (MLHC)*, 2024.

National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology, version 2026. `https://www.nccn.org/guidelines`, 2026.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedHALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 314–334, 2023.

Mor Peleg. Computer-interpretable clinical guidelines: A methodological review. *Journal of Biomedical Informatics*, 46(4):744–763, 2013.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Karan Singhal, Tao Tu, Juraj Gottweis, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Michal Staniek, Stefan Riezler, and Stefan Wager. Computable clinical guidelines: A survey of representations, tooling, and evaluation. *arXiv preprint arXiv:2503.07291*, 2025.

David R. Sutton and John Fox. The syntax and semantics of the PROforma guideline modeling language. *Journal of the American Medical Informatics Association*, 10(5):433–443, 2003.

Hao Wang, Yan Zhang, Wei Liu, Jie Chen, and Min Yang. MedKP: Knowledge-augmented prompting for guideline-faithful clinical question answering in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024a.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron Wallace. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9871–9889, 2023.

Peiyi Wang, Lei Li, Liang Chen, and Bill Yuchen Lin. Large language models are inconsistent and biased evaluators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 461–472, 2024b.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, et al. ColaCare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM Web Conference 2025*, pp. 2250–2261, 2025.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D. Wang, Joyce Ho, and Carl Yang. RAM-EHR: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pp. 754–765, 2024.

Cheng Yang, Lei Wang, Yifan Chen, Xiang Liu, and Maosong Sun. CancerGUIDE: Guideline-adherent large language models for cancer treatment recommendation. *arXiv preprint arXiv:2508.04114*, 2025.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv:2506.11474.

Xuejiao Zhao, Siyan Liu, Su Yang Yang, et al. MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM Web Conference 2025*, pp. 4442–4457, 2025a.

Yulai Zhao, Haolin Liu, Dian Yu, S. Y. Kung, Haitao Mi, and Dong Yu. One token to fool LLM-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic LLM benchmarks: Null models achieve high win rates. *arXiv preprint arXiv:2410.07137*, 2024.