Dissecting Neuro-Symbolic Quality Assurance for Synthetic Oncology Data Generation

Laxmigayathri Challa, Yuhan Zhou, Ana Cleveland, Haihua Chen, *Member, IEEE*

***Abstract*—Synthetic clinical data generation with large language models addresses the scarcity that limits cancer staging research, but oncology hallucinations are categorically harmful: one clinically impossible staging assignment contaminates every downstream model trained on it. Neuro-symbolic pipelines validate during generation, yet the contribution of individual quality-assurance components remains unclear. We report three controlled studies isolating gate necessity, constraint attribution, and retrieval conditionality, holding generation protocol, diversity thresholds, and fine-tuning hyperparameters constant across adapter conditions. The symbolic gate enforces schema completeness, ontology coverage against the Systematized Nomenclature of Medicine, and staging-logic consistency under American Joint Committee on Cancer eighth-edition rules. Ungated, 29.9% of records contain schema failures and 20.1% contain clinically invalid staging. Schema validation is the load-bearing filter: within the fully gated corpus it rejects 148 of 512 records, ontology grounding a further 24, and staging-logic validation none—the only generator producing logic violations is already excluded on schema, making clinical-logic validation a generator-conditional safeguard rather than the dominant filter. Retrieval augmentation is strongly model-dependent: it improves gate compliance for one generator by 12.5 percentage points, has no measurable effect for a second, and collapses output in a third. Across gated configurations ontology density is largely unchanged, indicating that symbolic validation improves clinical validity rather than vocabulary richness. Symbolic gating therefore buys corpus validity but no commensurate gain on real lung-cancer notes in this study; retrieval should be evaluated per model, and ontology density should not be reported as a proxy for corpus quality.**

***Index Terms*—synthetic clinical data, neuro-symbolic AI, component attribution, hallucination detection, large language models.**

I. Introduction

Lung cancer remains the leading cause of cancer mortality worldwide, with an estimated 2.48 million new cases and 1.8 million deaths in 2022 [1]. Almost everything about a patient's prospects turns on how far the disease has spread when it is found: in the United States, five-year relative survival is 63.7% while the tumor is still localized and 8.9% once it has metastasized, yet only 22% of cases are caught at the localized stage [2]. Stage decides whether a patient is offered curative resection or systemic therapy, which trials they are eligible for, and how their outcome is judged against expected care.

Facts of this kind are recorded for every patient, but they are recorded as narrative (pathology reports, operative notes, oncology summaries) rather than as coded fields a query can reach [3], and recovering them is complex, time-consuming, and heavily manual [4]. Yet nearly every use we would make of them operates at a scale prose cannot serve: assembling trial cohorts, auditing care against guidelines, curating registries, and training the clinical models meant to support all three. Working directly from real records runs into protections that exist for good reason: patient data is governed under HIPAA and GDPR through slow, jurisdiction-specific sharing mechanisms [5], and a solution that works by loosening them is not a solution. Where records are shareable, the clinically meaningful variables are rarely structured, and recovering them retrospectively demands expertise no annotation budget supports. Where they are reliably structured, the structure is thin. Registries capture a fixed set of abstracted variables rather than the reasoning in the note; trial data is richer but covers a narrow slice, with only 7.1% of adults with cancer enrolling in treatment trials [6]. The scarcity is structural, and it will persist.

Synthetic generation offers a way through—clinical text carrying the properties of real records while describing no real patient—and large language models make it tractable, producing rich clinical narratives at scale [7], showing promise across precision-oncology tasks [8], and grounding in domain evidence through retrieval-augmented generation (RAG) [9]. The difficulty is knowing whether what comes out is correct, and for most clinical content nothing decides that but expert judgment. Cancer staging is a deliberate exception: under AJCC 8th Edition rules [10], T, N, and M values map to a stage group through a fixed lookup, so a generated record either satisfies the rules or contradicts them—machine-decidably, and without reference to any patient. This work therefore begins with TNM staging: clinically consequential in its own right, and structured enough that a generator can be taught it and a validator can check it.

Feasible is not trustworthy. LLMs hallucinate [11], and in oncology hallucination is categorical rather than gradient: a note assigning T2 N0 M0 to mediastinal lymph node involvement is not slightly off but clinically impossible, and any model trained on it inherits an error that propagates without warning. Model-collapse work [12] sharpens the concern, since generative systems trained on their own outputs degrade across generations. Corpus quality therefore cannot be repaired downstream; it has to be enforced at the point of generation. The natural response is a symbolic gate: deterministic validators that screen every record before it enters the corpus, enforcing JSON schema completeness, SNOMED CT ontology coverage, and AJCC clinical-logic consistency as

L. Challa, Y. Zhou, and A. Cleveland are with the Department of Information Science, University of North Texas, Denton, TX 76203 USA.

H. Chen is with The Anuradha and Vikas Sinha Department of Data Science, University of North Texas, Denton, TX 76203 USA.

Corresponding author: Haihua Chen (e-mail: Haihua.Chen@unt.edu).

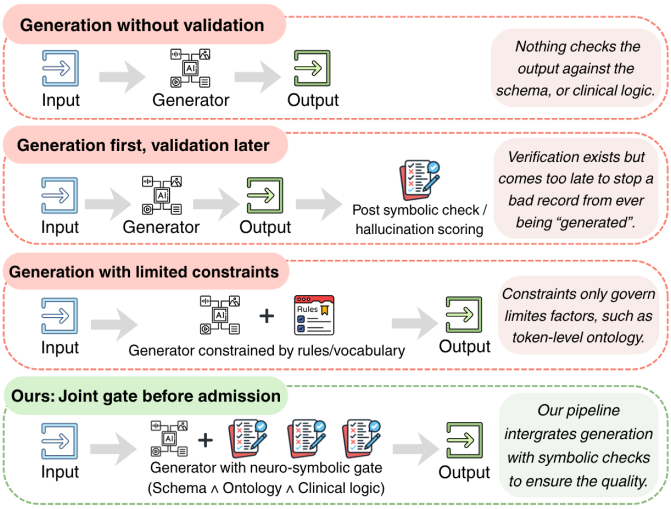

Fig. 1: Demonstration of different synthetic data generation methods.

hard admission constraints. Prior work [13] has shown that such a pipeline—neuro-symbolic generation coupled with a binary gate $G(x) = S \wedge O \wedge C$—improves downstream T-stage extraction on real clinical notes, recovering minority stages that zero-shot models cannot identify at all. That establishes the pipeline as a whole. It does not establish *which components produce the effect*, nor whether the quality signals used to justify them measure what they are assumed to measure.

The purpose of this study is to quantify the marginal contribution of each quality-assurance component—the gate as a whole, each of its three constraints (S, O, C), and retrieval augmentation—to two distinct outcomes: the clinical validity of the generated corpus, and its downstream utility on real oncology notes under a train-on-synthetic, test-on-real protocol. Each of these components carries a real cost in API calls, computational overhead, reduced yield, and architectural complexity, and a practitioner needs to know which of those costs are justified. A system study answers whether an architecture works; only a controlled decomposition answers which of its parts are load-bearing, and at what cost.

The research questions below are operationalized through controlled experiments in which a single pipeline component is manipulated while all others are held constant. Generation protocols, label diversity, entropy thresholds, QLoRA hyperparameters, and evaluation procedures are fixed across conditions, ensuring that observed differences are attributable to the manipulated variable alone. The research questions are:

1) **RQ1 (Gate Necessity):** What enters a synthetic clinical corpus when symbolic quality assurance is removed?
2) **RQ2 (Constraint Attribution):** Which symbolic constraint contributes most to corpus quality?
3) **RQ3 (Retrieval Conditionality):** When does retrieval augmentation improve generation quality?
   Section IV maps each question to the conditions that answer it.

The contributions of this study are as follows:

- **A component attribution methodology:** Five matched adapter conditions and a within-corpus decomposition procedure that isolate the gate, its three constraints, and retrieval as independently testable variables under fixed generation and fine-tuning protocols.
- **Evidence on what gating does and does not buy:** The failures an ungated corpus admits are structurally indistinguishable from valid records, establishing a class of contamination that post-hoc inspection cannot recover—and, separately, that removing it does not by itself improve transfer to real pathology reports.
- **A ranking of the symbolic constraints by marginal contribution:** We separate filtering work from standardization work, and identify clinical-logic validation as a generator-conditional safeguard rather than a high-volume filter.
- **A per-generator decision procedure for retrieval:** Gate compliance serves as the diagnostic for whether retrieval will help, do nothing, or cause output collapse in a given generator.

II. Related Work

Across synthetic clinical data generation, hallucination research, retrieval- augmented generation, and neuro-symbolic clinical NLP, validation is consistently applied *after* generation: as a benchmark score, a post-hoc filter, or a normalization pass over records already produced. None of these literatures treats the validators themselves as a variable to be measured, so a pipeline that couples generation with symbolic checks is evaluated only end-to-end—leaving the marginal contribution of any single quality-assurance component unknown. Table I situates the present work against the closest prior systems on this dimension.

A. Synthetic Clinical Data: From Tabular Simulation to LLM Generation

The earliest synthetic clinical data systems operated on structured EHR extracts. GAN-based approaches—MedGAN [14], medBGAN [15], and HealthGAN [16]—learn to approximate the marginal and joint statistics of tabular records and have been validated through TSTR protocols on structured data [17]. Synthea [18] takes a mechanistic route, producing structurally valid FHIR records by construction from probabilistic disease models. Both paradigms share a structural ceiling: their output is a row of feature values, and the unstructured clinical text in which a physician encodes staging rationale—tumor invasion depth, lymph node involvement, metastatic spread—is outside their generative scope entirely. Ontologies, where they appear at all, serve as vocabulary mappings applied to already-generated content rather than as constraints on what enters the corpus.

LLM-based generation removed the free-text barrier [19], [20]. Models conditioned on structured metadata produce linguistically plausible clinical notes at scale, and MedSyn [21] demonstrated that knowledge-graph-conditioned generation can improve downstream ICD-10 coding accuracy, with the largest gains on long-tail codes. Yet the dominant design pattern carries over from the GAN era unchanged: records are produced first, and quality checks follow. Ontology mapping

tools— MetaMap [22], cTAKES [23], and the BioPortal Annotator API [24]—are applied as normalization steps on records that already exist—so post-hoc validation cannot catch what it cannot see.

This work enforces ontology coverage and clinical-logic consistency as admission constraints *during* generation rather than as normalization applied afterward (Fig.1).

B. Hallucination in Clinical LLMs: Evaluation Versus Prevention

Hallucination in clinical LLMs has so far been treated chiefly as an evaluation problem rather than a prevention one. Med-HALT [25] introduced a benchmark taxonomy for medical hallucination types; large-scale clinical LLM evaluations [26] showed that high scores on general benchmarks do not transfer reliably to domain-specific clinical tasks requiring precise factual grounding; MedHallu [27] extended this to a comprehensive detection benchmark across clinical domains. Together these establish that hallucination is pervasive, measurable, and domain-specific—but measuring it after the fact is not the same as preventing it at the point of generation.

The clinical hallucination analysis of [28] applies clinician-in-the-loop evaluation as post-hoc analysis, correctly identifying that hallucination rates differ by task type and model family, but again after generation is complete. The closest prior approach to generation-time prevention is ontology-guided constrained decoding [29], which reduces hallucination during clinical summarization by restricting the model's output vocabulary to ontology-consistent tokens. Constrained decoding, however, operates on individual token predictions; it does not enforce higher-level semantic constraints such as AJCC staging logic, which requires cross-field consistency across an entire structured record rather than local token validity.

That gap reflects a distinction this paper treats as load-bearing: hallucination as a *gradient* failure—a claim that is somewhat inaccurate but recoverable—versus hallucination as a *categorical* failure, one that is clinically impossible and corrupts downstream learning silently. In oncology the latter dominates. Embedding AJCC rules as a binary gate is what converts this categorical failure from an invisible contaminant into a measurable engineering property, and the gate decomposition experiment in this paper quantifies how much of that filtering work the logic constraint performs relative to schema and ontology.

C. Neuro-Symbolic Approaches in Clinical NLP

The neuro-symbolic paradigm [30] combines the generative expressiveness of neural models with the precision of symbolic reasoning, and in clinical NLP it has been applied chiefly to classification and coding rather than generation. HybridCode v2 [31] uses neuro-symbolic verification for ICD-10 coding, achieving zero Type-I hallucination on MIMIC-III by layering symbolic rule checks over neural candidate generation. Ontology-guided decoding [29], introduced above as the closest prior approach to generation-time constraint enforcement, is the same pattern applied to summarization. In both cases the symbolic component operates as a post-generation filter or decoding constraint, not as a corpus admission gate, and neither addresses synthetic data *generation* at all.

The architectural distinction between a filter and a gate is not merely terminological. A post-generation filter incurs the full cost of generating every record before deciding which to discard. A corpus admission gate makes admission a binary property of each record independently: records that fail do not enter the corpus, and the gate's pass rates are directly interpretable as measurements of pipeline component contribution. This interpretability is what the component attribution study in this paper exploits—by varying which constraints are active while holding all other variables fixed, the marginal downstream contribution of each symbolic component becomes directly measurable rather than inferred.

D. Retrieval-Augmented Generation in Biomedicine

Retrieval-augmented generation (RAG) [9] grounds LLM outputs in verifiable retrieved evidence, reducing hallucination by anchoring generation to specific source documents. MedCPT [32] extends this to biomedicine through contrastive pre-training on PubMed citation pairs, enabling zero-shot biomedical retrieval without task-specific fine-tuning. Clinical LLM work has concentrated on question answering [26] and summarization, and where retrieval has been applied it has consistently improved factual accuracy—and it is this consistency that has hardened into an implicit assumption: that retrieval is universally beneficial, and more grounding context simply produces more accurate output.

To our knowledge, that assumption has not been tested in the structured generation setting used here, where the interaction between retrieved context and a model's format-compliance capability is non-trivial. A model with fragile instruction-following may fail to integrate retrieved content without format disintegration, reducing effective corpus yield even as the retrieved context enriches the individual records that do pass validation. The retrieval ablation in this paper is the first controlled test of the universality assumption in a structured clinical data generation context, and the results challenge it directly: retrieval improves one generator, has no measurable effect on a second, and causes systematic output collapse in a third.

E. TSTR Evaluation and TNM Staging Automation

The Train-on-Synthetic, Test-on-Real (TSTR) protocol [33] measures synthetic data utility through downstream task performance rather than distribution-similarity proxies. Prior clinical TSTR evaluations have focused on structured tabular EHR data [16], [17]; this paper applies TSTR to synthetic oncology note generation with explicit staging annotations, evaluated on TCGA-derived lung-cancer and cross-tumor clinical note cohorts.

Cancer staging extraction spans rule-based methods, classical machine learning, and deep learning in comparable numbers, with transformer-based approaches a smaller but growing share [3]. Reported performance on curated benchmarks is strong—macro-F1 above 0.90 for both a transformer-plus-rules staging pipeline [34] and a fine-tuned generative model

TABLE I: Capability comparison with representative prior work on synthetic clinical data generation, hallucination prevention, and neuro-symbolic medical NLP. Per-system detail is given in Section II.

Legend: ✓ present; (✓) present but not as a generation-time, record-level admission constraint; ✗ absent.

Work	System type	Free-text	Neuro-symbolic gate				Attribution
			Schema	Ontology	Clinical logic	Gen-time gate	
MedGAN (2017) [14]	Tabular EHR (GAN)	✗	✗	✗	✗	✗	✗
Synthea (2018) [18]	FHIR record sim. (rule-based)	✗	(✓)[a]	✗	✗	✗	✗
MedSyn (2024) [21]	LLM text (KG-conditioned)	✓	✗	(✓)[b]	✗	✗	✗
Kasthurirathne et al. (2023) [20]	LLM text	✓	✗	✗	✗	✗	✗
Mehenni & Zouaq (2024) [29]	Constrained decoding	✓	✗	✓	✗	(✓)[c]	✗
Asgari et al. (2025) [28]	Hallucination eval	✓	✗	NA	✗	✗	✗
Our work (2026)	**Neuro-symbolic LLM + TSTR**	✓	✓	✓	✓	✓	✓

[a] FHIR records are structurally valid *by construction*, not validated as a generation-time admission gate.
[b] Knowledge-graph sampling conditions the generation prompt; it biases output toward valid content but rejects nothing.
[c] Constraints act at generation time but on token-level vocabulary, not as a record-level admission gate.

applied to the same task [35]—but performance on real-world oncology notes remains substantially lower, a gap that reflects the data rather than the model: curated benchmarks provide clean, expert-annotated records, while clinical dictation is noisy, abbreviated, and inconsistently formatted. Synthetic data generation targeted at the structured oncology domain is therefore a direct intervention on the most limiting constraint in TNM staging automation. Whether the symbolic quality-assurance components of such a pipeline are individually necessary to realize that intervention, and which contribute most, is the question the remainder of this paper answers.

III. Clinical Foundation

A. TNM Staging and AJCC 8th Edition

The TNM classification system codifies cancer extent along three independent dimensions. T (T0–T4) describes the size and local invasiveness of the primary tumor; N (N0–N3) describes regional lymph node involvement; M (M0/M1) describes the presence or absence of distant metastasis. Combined under AJCC 8th Edition rules [10], these three values determine a clinical stage group (I–IV) through a deterministic lookup: the mapping is fixed, not probabilistic, and violations are logically impossible rather than merely unlikely.

This determinism is consequential for synthetic data generation. A generated record assigning M1 staging without a corresponding Stage IV designation, or Stage I with N2 lymph node involvement, is not an imprecise approximation of a valid record—it is a logical contradiction. Such contradictions pass structural formatting checks, carry normal ontology density, and satisfy JSON schema requirements. They are invisible to every quality metric except an explicit clinical-logic validator. Identifying and quantifying how many generated records contain such contradictions—and how much downstream harm they cause when admitted to a training corpus—is one of the central questions this paper answers.

For non-small cell lung cancer, the 32-cell grid $\{\text{T1, T2, T3, T4}\} \times \{\text{N0, N1, N2, N3}\} \times \{\text{M0, M1}\}$ covers the main staging categories. Within this grid, T3 and T4 carry particular clinical weight: they mark the point at which local tumor extent begins to determine resectability and the choice between surgery and definitive chemoradiation, and they are the categories a staging model most often gets wrong. A model that cannot distinguish T3 from T4 is least reliable exactly where the treatment decision is most consequential. T3 and T4 are also minority classes in real cohorts, which makes macro-F1 rather than aggregate accuracy the informative measure of staging performance on the real-world oncology note cohorts evaluated here.

B. Clinical Ontologies as Validity Anchors

The Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) provides a standardized, hierarchically organized vocabulary of more than 360,000 clinical concepts [36], covering anatomy, findings, procedures, substances, and observable entities. Its relevance to synthetic data quality is not as a retrieval resource but as a *validity anchor*: a generated clinical record whose concepts cannot be mapped to any SNOMED CT term is producing fabricated terminology—terminology that has no grounding in any recognized clinical vocabulary. This is detectable without access to the patient's actual record, making SNOMED CT coverage checkable on the generated text alone

We operationalize this through the BioPortal Annotator API [24], which maps free text to SNOMED CT concepts and returns matched terms with concept IDs. Length-normalized SNOMED density—terms per 100 words—provides a cross-model vocabulary richness metric used throughout the ablation analyses. A key question this paper addresses is whether enforcing SNOMED CT coverage as a corpus admission constraint produces records with measurably richer clinical vocabulary, or whether the gate targets validity rather than richness. The gate decomposition experiment answers this directly.

IV. Methodology

This section describes the neuro-symbolic generation pipeline and the experimental design through which its components are isolated. The pipeline couples LLM generation with a symbolic gate $G(x) = S \wedge O \wedge C$ applied to every

candidate record before admission to the training corpus; in the retrieval condition, MedCPT-retrieved PubMed context is prepended to the generation prompt before the gate is applied. Figure 2 presents the full pipeline (top) together with the five matched conditions through which its components are dissected (bottom).

A. Problem Formulation

We formulate synthetic clinical data generation as a constrained generation problem. Let x denote a candidate structured patient record produced by a generator model $\mathcal{M}$ from a prompt p: $x \sim \mathcal{M}(\theta, p)$, where θ denotes model parameters. For x to be admitted to the training corpus, it must satisfy a constraint set $\mathcal{C} = \{c_{\text{schema}}, c_{\text{ontology}}, c_{\text{logic}}\}$, evaluated through a binary symbolic gate:

$$G(x) = \begin{cases} 1 & \text{if } x \text{ satisfies all } c \in \mathcal{C}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The gate decomposes as $G(x) = S \wedge O \wedge C$, where S denotes schema validity (JSON field completeness enforced by the *rigid.v3* schema), O denotes ontology coverage (at least one SNOMED CT concept confirmed via the BioPortal Annotator API [24]), and C denotes AJCC clinical-logic consistency, implemented over the most commonly violated rules (M1 staging implies Stage IV; stage group is non-empty when T, N, and M values are all present); Section V-C reports the effect of expanding this rule set. The gate is applied per record during corpus construction: records for which $G(x) = 0$ are discarded and do not enter the fine-tuning corpus under any condition.

Pipeline quality is assessed along three dimensions: structural validity (JSON schema compliance rate), label diversity (Shannon entropy per TNM dimension, with floors $H_T, H_N \geq 1.109$ and $H_M \geq 0.554$), and downstream utility (T-, N-, and M-stage extraction accuracy and macro-F1 under TSTR).

B. Generation Setup

All ablation conditions share the generation setup described below.

TNM grid: As discussed in Section III-A, records are seeded from a 32-cell TNM grid $\{\text{T1, T2, T3, T4}\} \times \{\text{N0, N1, N2, N3}\} \times \{\text{M0, M1}\}$, assigned round-robin across generation runs, with each cell randomizing patient age (45–80 years), sex, and histology subtype.

Label diversity: Two Shannon-entropy checkpoints, applied post-generation and pre-training, abort the pipeline if any TNM dimension approaches single-class concentration. Floors and checkpoint placement are given in Appendix A.

Schema: The *rigid.v3* schema encodes eight non-optional clinical domains per record: TNM staging (AJCC 8th Edition), histology (ICD-O-3 and SNOMED CT codes), molecular drivers, demographics, imaging findings, treatment modalities, health equity factors, and free-text narrative. All fields are mandatory; empty strings trigger $G(x) = 0$.

Generator models: Three models are evaluated across all ablation conditions. **GPT-4o** [37] is accessed via the OpenAI API with `response_format={type: json_object}`, which enforces JSON output at the API level. **Llama-3.3-70B-Instruct** [38] is loaded locally in 4-bit NF4 quantization on an NVIDIA H200 GPU. **ClinicalCamel-70B** [39] is loaded under the same quantization. It does not ship with a chat template, so we attach a Llama-style `<|start_header_id|>` fallback template at model load time; this is the only model-specific modification applied in any condition. All three models receive identical prompts per ablation condition, and identical decoding parameters throughout (temperature 0.7, top-p 0.9, top-k 50, sampling enabled, 1,024 maximum new tokens).

Retrieval: In Ablation 3 (RAG condition only), each generation call is preceded by a MedCPT Query Encoder [32] lookup over a FAISS [40] flat inner-product index built from ≈2,000 PubMed lung-cancer abstracts encoded with the MedCPT Article Encoder. The query is constructed from the record's T, N, and M seed values together with its histology subtype; the top $k = 2$ retrieved abstracts are prepended to the generation prompt as grounding context. In all other conditions—including the no-RAG arm of Ablation 3—no retrieval context is added. [1]

The component attribution study rests on a single design principle: one variable is manipulated per experiment while all others are held fixed. Table II summarizes the five adapter conditions and their mapping to the three ablation experiments, corresponding to the dissection panel of Figure 2. Each adapter is trained on a condition-specific corpus generated under the corresponding study configuration.

Research questions and how they are operationalized

RQ	Question	Contrast	Answered in
1	Gate necessity: what enters a corpus when symbolic quality assurance is removed?	A vs. D	§IV-C1, §V-B
2	Constraint attribution: which symbolic constraint does the filtering work?	B, C, D; per-record S, O, C flags	§IV-C2, §V-C
3	Retrieval conditionality: when does retrieval augmentation improve quality?	D vs. E	§IV-C3, §V-D

Generation protocol, TNM seeding grid, entropy floors, QLoRA hyperparameters, and evaluation procedure are fixed across all conditions, so each contrast varies one component alone.

C. Ablation Design

1) Ablation 1: Gate Necessity (RQ1): Two matched corpora are generated from the same TNM grid using the same models and hyperparameters. The *ungated* corpus (Adapter A) admits every record for which JSON parses, regardless of schema completeness, SNOMED CT coverage, or AJCC logic. The *gated* corpus (Adapter D) admits only records passing the full gate $G(x) = S \wedge O \wedge C$. Both corpora must pass all three entropy floors before training. Gate component flags (S, O, C) are logged for every record in the ungated corpus whether or not they are used for admission.

[1] Generation calls fall back to a keyword retriever when the MedCPT index returns no qualifying match: 64 of 448 calls (32 Llama-3.3-70B and 32 ClinicalCamel-70B); all GPT-4o calls use MedCPT.

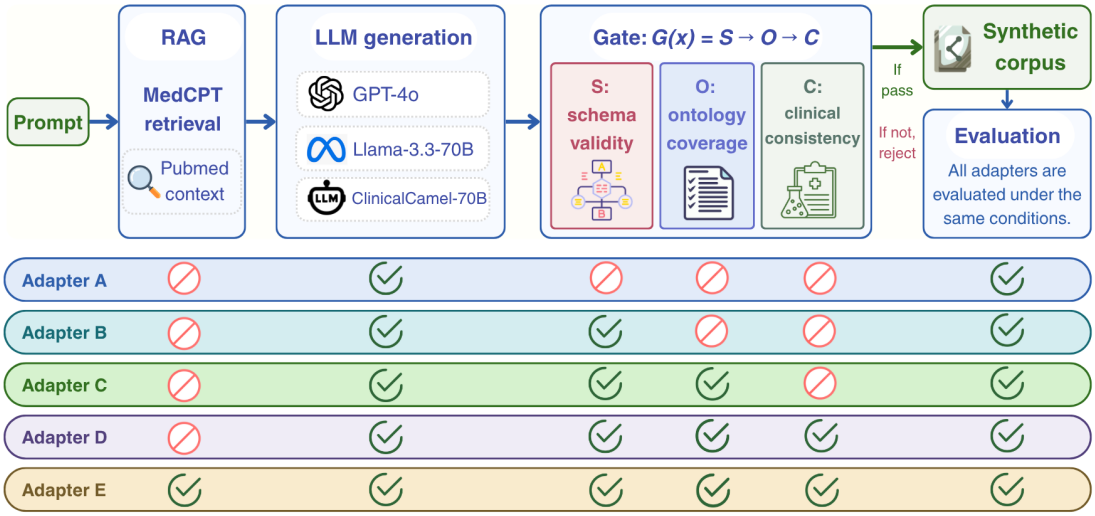

Fig. 2: Neuro-symbolic generation pipeline and its component dissection. *Top (target pipeline).* The full framework couples LLM generation with a neuro-symbolic gate $G(x) = S \wedge O \wedge C$—schema validity ($S$), SNOMED CT ontology coverage (O), and AJCC clinical-logic consistency (C)—and, in the retrieval condition, prepends MedCPT-retrieved PubMed context to the generation prompt as grounding before the gate is applied. Records that pass the gate enter the synthetic corpus as structured *rigid.v3* fields plus a free-text clinical narrative. *Bottom (dissection of components).* Five matched conditions (Table II) remove or add one component at a time: ungated baseline (Adapter A); schema only (Adapter B); schema+ontology (Adapter C); full gate, no retrieval (Adapter D); and full gate+RAG (Adapter E). The three within-gate conditions (B, C, D) isolate the marginal contribution of each constraint (RQ 2); A versus D isolates gate necessity (RQ 1); and D versus E isolates retrieval conditionality (RQ 3). The three generators (GPT-4o, Llama-3.3-70B, ClinicalCamel-70B), the 32-cell TNM seeding grid, the entropy floors, the QLoRA hyperparameters, and the TSTR evaluation protocol are held fixed across every condition (Sections IV-B–IV-D), so that any measured difference is attributable to the manipulated component alone.

TABLE II: Adapter conditions and ablation mapping. Adapter D is the shared full-gate, no-RAG baseline across all three ablations. Per-model run counts differ by construction: GPT-4o contributes 128 generations in every condition, 64 API calls having returned no output; Llama-3.3-70B and ClinicalCamel-70B contribute 192 each (no-RAG) and 160 each (RAG). Totals are 512 per no-RAG condition and 448 for the RAG condition. All adapters use identical QLoRA configuration (Appendix C).

Adapter	Gate condition	RAG	Ablation
A	None (ungated)	No	1
B	S only	No	2
C	$S \wedge O$	No	2
D	$S \wedge O \wedge C$	No	1, 2, 3
E	$S \wedge O \wedge C$	Yes	3

2) Ablation 2: Constraint Attribution (RQ2): Three corpora are generated under progressively stricter gates:

1) **Schema only (S):** Admits records passing JSON completeness; SNOMED CT and AJCC logic are computed and logged but not used for admission. Trains Adapter B.
2) **Schema + Ontology ($S \wedge O$):** Adds SNOMED CT coverage as an admission requirement; AJCC logic is computed and logged but not enforced. Trains Adapter C.
3) **Full gate ($S \wedge O \wedge C$):** All three constraints enforced simultaneously. Trains Adapter D.

Although three corpora are generated to train Adapters B–D, the marginal contribution of each constraint is computed *within a single corpus* from the per-record gate flags (S, O, C logged for every generated record), so that the attribution reflects the constraints themselves rather than sampling variation between independently generated corpora. SNOMED density is computed across all three conditions on the admitted-record basis.

3) Ablation 3: Retrieval Conditionality (RQ3): Two matched corpora, identical in every respect except retrieval context. Both pass the full gate $G(x) = S \wedge O \wedge C$ and the same entropy floors. The *no-RAG* corpus (Adapter D) uses schema plus TNM seed description only. The *RAG* corpus (Adapter E) prepends MedCPT-retrieved PubMed abstracts to each generation prompt via the FAISS index described in Section IV-B. Both corpora are generated from the same TNM grid with the same three models.

D. Evaluation Protocol

All five adapters are evaluated under identical TSTR conditions at three levels of distributional distance from the training distribution:

1) **Synthetic held-out** ($n = 62$): in-distribution evaluation.
2) **TCGA Lung** ($n = 737$): real lung-cancer notes.
3) **TCGA Cross-Tumor** ($n = 3{,}161$): out-of-distribution oncology evaluation.

For the evaluation, accuracy is reported against the majority-class baseline, with macro-F1 as the metric sensitive to minority-class recovery. Per-axis scoring is restricted to notes carrying a valid gold label for that axis, and per-axis macro-F1 is reported in Appendix B.

The base model for all adapters is `meta-llama/Meta-Llama-3-8B-Instruct`; full QLoRA hyperparameters are given in Appendix C.

TABLE III: Corpus quality summary by ablation condition. Yield = admitted / generated. Schema, Onto., and Logic pass rates are computed over all generated records. All entropy floors ($H_T, H_N \geq 1.109$; $H_M \geq 0.554$) pass in every condition. SNOMED density is length-normalized (terms per 100 words) over admitted records. Unique = distinct SNOMED concept IDs across all admitted records.

Adapter	Condition	Gen.	Yield	Schema	Onto.	Logic	SNOMED/100w	Unique
A	Ungated	512	100.0%	70.1%	100.0%	79.9%	28.39	213
B	Schema only	512	70.5%	70.5%	100.0%	78.1%	29.71	192
C	Schema + Onto.	512	70.3%	70.3%	100.0%	77.0%	29.86	201
D	Full $G(x)$, No-RAG	512	66.4%	71.1%	95.3%	77.1%	29.83	198
E	Full $G(x)$ + RAG	448	69.2%	70.3%	83.3%	71.0%	29.52	191

V. RESULTS

Findings by research question

RQ	Finding	Evidence
1	Ungated generation admits 29.9% schema failures and 20.1% AJCC logic violations; the full gate rejects one record in three, and the cost falls almost entirely on a single generator.	Fig. 3, Tables III, IV
2	Schema is the load-bearing filter (148 of 512 rejected); ontology removes 24 more; clinical logic removes none, making it a generator-conditional safeguard rather than a high-volume filter.	Fig. 4, Table V
3	Retrieval is model-dependent: +12.5 pp gate compliance for one generator, no measurable effect for a second, and output collapse in a third.	Figs. 5, 6

Across gated configurations SNOMED density is flat (29.71–29.86 terms/100w), and corpus-quality gains do not produce commensurate improvement on real lung-cancer notes (§V-E).

A. Corpus Quality Across Ablation Conditions

Table III reports corpus-level quality metrics for all five adapter conditions: generation yield, per-constraint pass rates (schema, ontology, AJCC logic), length-normalized SNOMED CT density, and the number of distinct SNOMED concepts in each admitted corpus. Three cross-cutting observations follow.

- **Label diversity is preserved across all conditions:** Every corpus passes all three entropy floors, with measured diversity at the theoretical maxima (Appendix A). No entropy intervention is triggered at any phase across any adapter condition.
- **SNOMED density is insensitive to gate strictness:** Corpora B, C, and D show near-identical density (29.71, 29.86, and 29.83 terms per 100 words respectively). Enforcing progressively stricter gate constraints does not produce records with more ontological content; the gate filters for clinical validity, not vocabulary richness.
- **Model compliance is strongly bimodal:** GPT-4o and Llama-3.3-70B achieve near-perfect schema and logic compliance under gated conditions. ClinicalCamel-70B maintains 20–23% schema compliance regardless of gate configuration or retrieval context—a property of the model, not the gate. These effects are reported per-model in each ablation below.

B. RQ1 (Gate Necessity) — Gated vs. Ungated Generation

Figure 3 presents gate failure attribution and per-component pass rates across all five conditions. The ungated corpus (Adapter A) admitted all 512 generated records but silently included substantial clinical noise: 29.9% of records failed schema validation and 20.1% violated AJCC clinical logic (Fig. 3a, column A). Neither failure is detectable through standard formatting inspection—they manifest as missing JSON fields and logically impossible staging combinations respectively. The full-gate corpus (Adapter D) admitted 340 of 512 generated records (66.4% yield), with every admitted record satisfying all three constraints simultaneously. The gate therefore eliminates one in three generated records and removes both detected categories of clinical noise from the training corpus.

TABLE IV: Ablation 1 gate metrics by model. Schema, Onto., and Logic are pass rates over all generated records. Yield for the ungated condition is 100% by construction (no gate applied); for the full-gate condition it equals the fraction of records satisfying all three constraints simultaneously.

Model	Condition	Schema	Onto.	Logic	Yield
GPT-4o	Ungated	100.0%	100.0%	100.0%	100.0%
GPT-4o	Full gate	100.0%	100.0%	100.0%	100.0%
Llama-3.3-70B	Ungated	100.0%	100.0%	100.0%	100.0%
Llama-3.3-70B	Full gate	100.0%	87.5%	100.0%	87.5%
ClinicalCamel	Ungated	20.3%	100.0%	46.4%	100.0%
ClinicalCamel	Full gate	22.9%	100.0%	39.1%	22.9%

The gate cost is asymmetric across generator models (Table IV). GPT-4o achieves 100% gate compliance under the full gate (128/128 records admitted) with no yield loss, reflecting API-level JSON enforcement (`response_format={type: json_object}`) that provides structural guarantees unavailable to locally-served models. Llama-3.3-70B achieves 100% schema and logic compliance in both conditions; under the full gate its yield drops to 87.5% (168/192 records), with all 24 rejected records failing the ontology constraint specifically. ClinicalCamel-70B exhibits near-identical schema compliance ungated (20.3%) and under the full gate (22.9%), confirming that its compliance limitation is a model-level property orthogonal to gate configuration.

Over admitted records, mean SNOMED density rises only slightly, from 28.39 (ungated) to 29.83 terms per 100 words under the full gate (+5.1%). This modest increase reflects the exclusion of low-density ClinicalCamel records—which dominate the ungated corpus failures—rather than any vocabulary enrichment effect of the gate itself.

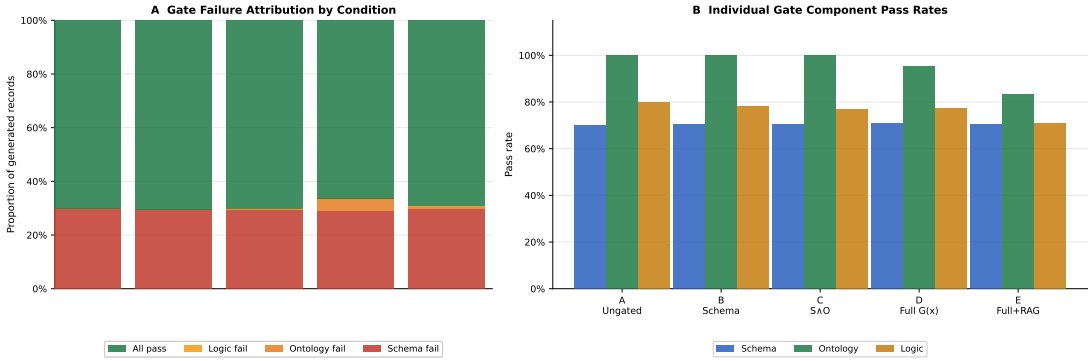

Fig. 3: Gate vs. No-Gate. *(a)* Stacked gate failure attribution across all five conditions. Column A (Ungated) contains 29.9% schema failures and 20.1% logic failures that would silently enter fine-tuning without the gate. Column D (Full $G(x)$) shows the residual failure structure after all three constraints are applied; admitted records pass all three simultaneously. *(b)* Individual gate component pass rates. Ontology coverage reaches 100% for conditions A, B, and C by design (the schema mandates histology codes); the full gate (D, E) applies all three constraints independently, and the residual ontology failures at D are the 24 Llama records identified in Fig. 4.

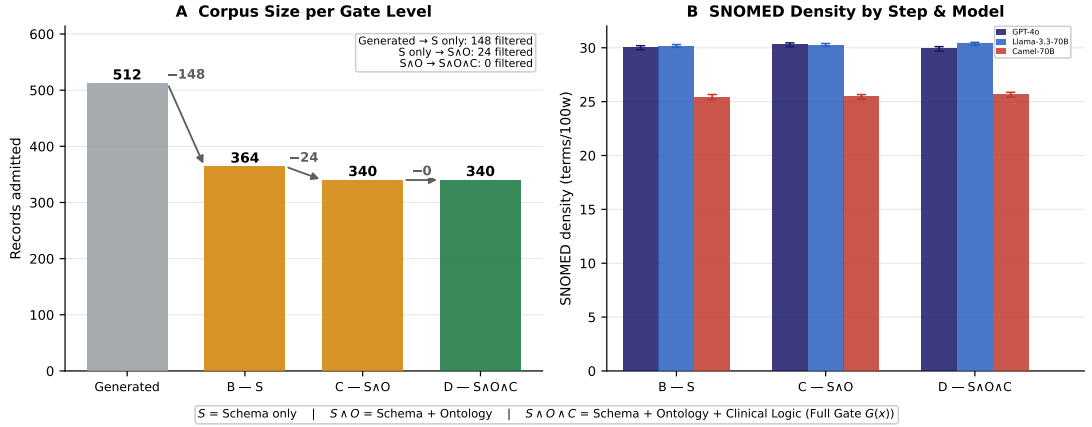

Fig. 4: Gate decomposition (within-corpus). *(a)* Gate funnel within the full-gate corpus: of 512 generated records, schema admits 364; ontology grounding then removes 24, leaving 340; and AJCC clinical-logic validation removes none. The marginal rejections—schema 148 (all ClinicalCamel), ontology 24 (all Llama), logic 0—identify schema as the load-bearing filter. *(b)* SNOMED density by model across gate levels; flat for GPT-4o and Llama (≈30 terms/100w), lower for ClinicalCamel (≈25), confirming the gate filters validity, not vocabulary.

C. RQ2 (Constraint Attribution) — Marginal Contribution of Gate Components

Figure 4 shows the within-corpus gate funnel, the marginal rejection attributable to each constraint, and SNOMED density across gate levels. Computed within the full-gate corpus (512 generated records), schema validation is the load-bearing filter: it admits 364 records and rejects 148, overwhelmingly from ClinicalCamel-70B. Of the 364 schema-passing records, the ontology constraint removes a further 24—all from Llama-3.3-70B—and the AJCC clinical-logic constraint removes none (Fig. 4a).

The zero marginal contribution of the logic constraint is robust. Applying a substantially fuller AJCC consistency check (M1⇒Stage IV, M0⇏Stage IV, Stage I with N≥1 or T≥3, N3 below Stage IIIB) to every schema-passing record with machine-parseable staging surfaces no clinically impossible assignments. The compliant generators (GPT-4o, Llama-3.3-70B) do not produce AJCC contradictions, and the generator that does (ClinicalCamel-70B) is removed by the schema constraint before logic is evaluated. The clinical-logic gate is therefore a *generator-conditional* safeguard: the class of error it is designed to catch—for example, M1 designation without a Stage IV assignment, or Stage I co-occurring with N2/N3 involvement—does occur in principle, but no such record reaches the logic check in this study because schema filtering removes the only non-compliant generator first.

SNOMED density is flat across the three gate levels on the admitted-record basis: 29.71, 29.86, and 29.83 terms per 100 words for B, C, and D respectively (Fig. 4b). Enforcing the ontology constraint does not enrich the vocabulary of admitted records; it standardizes a small fraction whose generated concepts fall outside the controlled vocabulary. The gate operates on clinical validity, not vocabulary richness.

By model, GPT-4o and Llama-3.3-70B maintain 100% schema and logic compliance at every gate level (Table V). ClinicalCamel-70B's logic pass rate over all generated records is low and roughly constant across gate levels (41.7%, 38.5%, 39.1%), but its logic-violating records also fail schema, so they are removed by the schema constraint rather than contributing marginal logic-gate rejections.

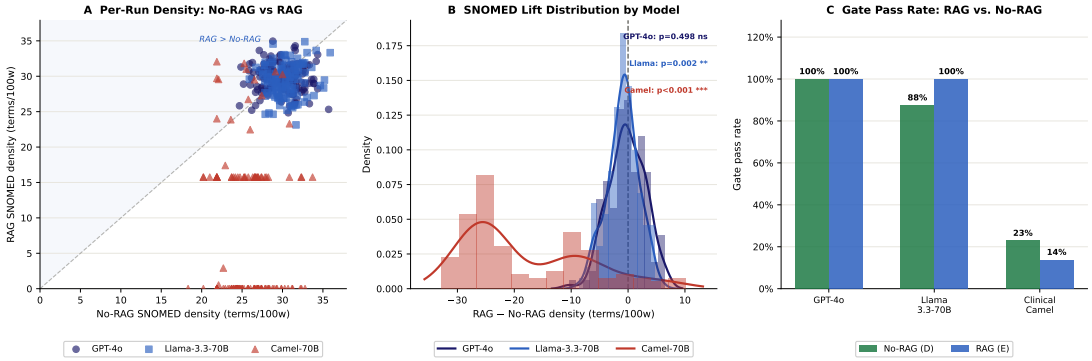

Fig. 5: RAG vs. No-RAG. *(a)* Per-run SNOMED density scatter. GPT-4o and Llama points cluster symmetrically around the diagonal; ClinicalCamel triangles collapse to the y=0 and y=15.75 floor under RAG. *(b)* Distribution of per-run density delta (RAG − No-RAG) by model. GPT-4o centered near zero (two-sided $p = 0.498$, ns); Llama a small but significant negative shift ($p = 0.002$); ClinicalCamel strongly negative ($p < 0.001$), reflecting retrieval collapse. *(c)* Gate pass rate by model. RAG lifts Llama gate compliance from 87.5% to 100%; ClinicalCamel falls from 22.9% to 13.8%.

TABLE V: Ablation 2 gate metrics by model across all three gate levels. GPT-4o and Llama-3.3-70B maintain near-perfect compliance at every level; ClinicalCamel-70B's compliance (21–23%) is unchanged by which constraints are active; its logic violations co-occur with schema failures, so they are removed by the schema constraint rather than contributing marginal logic-gate rejections.

Model	Gate level	Schema	Onto.	Logic	Yield
GPT-4o	S	100.0%	100.0%	100.0%	100.0%
GPT-4o	$S \wedge O$	100.0%	100.0%	100.0%	100.0%
GPT-4o	$S \wedge O \wedge C$	100.0%	100.0%	100.0%	100.0%
Llama-3.3-70B	S	100.0%	100.0%	100.0%	100.0%
Llama-3.3-70B	$S \wedge O$	100.0%	100.0%	100.0%	100.0%
Llama-3.3-70B	$S \wedge O \wedge C$	100.0%	87.5%	100.0%	87.5%
ClinicalCamel	S	21.4%	100.0%	41.7%	21.4%
ClinicalCamel	$S \wedge O$	20.8%	100.0%	38.5%	20.8%
ClinicalCamel	$S \wedge O \wedge C$	22.9%	100.0%	39.1%	22.9%

D. RQ3 (Retrieval Conditionality) — Retrieval-Augmented vs. Non-Augmented Generation

Figure 5 presents the RAG ablation results. The retrieval intervention exhibits strongly model-dependent behavior, motivating examination of generator-specific effects before evaluating downstream utility.

Overall, the RAG corpus (Adapter E) shows a gate pass rate of 69.2% versus 66.4% for No-RAG (Adapter D, +2.8 pp). On the admitted-record basis, mean SNOMED density (29.52 vs 29.83) and unique concept coverage (191 vs 198) are essentially unchanged: retrieval does not enrich the corpus a model trains on. The aggregate effect of retrieval is instead dominated by a single generator, examined below. (Measured over *all* generation attempts—including rejected and collapsed runs—mean density falls from 27.05 to 23.18 terms per 100 words; this all-generation drop is driven entirely by the ClinicalCamel collapse, not by any change to admitted records.)

GPT-4o shows a mean per-run density delta of −0.20 terms per 100 words (Mann-Whitney, two-sided, $p = 0.498$, ns), with a gate pass rate of 100% in both conditions. RAG adds no measurable signal for a model already operating at ceiling on every metric.

Llama-3.3-70B exhibits the most consequential RAG effect: gate pass rate increases from 87.5% (No-RAG) to 100.0% (RAG), a +12.5 pp lift (Fig. 5c). The 24 records that Llama fails in the No-RAG condition all fail the ontology constraint; retrieved PubMed context provides sufficient SNOMED CT grounding to bring these records into compliance. The mean per-run density delta is −0.92 terms per 100 words (Mann-Whitney, two-sided, $p = 0.002$): retrieval slightly *lowers* per-run density even as it lifts gate compliance, confirming that RAG helps Llama pass the gate by admitting previously-failing records rather than by enriching the vocabulary of already-admitted ones.

ClinicalCamel-70B is the source of the aggregate density drop. In the No-RAG condition ClinicalCamel produces a mean SNOMED density of 25.64 terms per 100 words and a gate pass rate of 22.9%, consistent with its behavior across all other conditions. Under RAG, 75 of 160 runs (46.9%) produce a SNOMED density of exactly 0.0, and 113 of 160 runs (70.6%) produce a density at or below 15.75 (Mann-Whitney $p < 0.001$). Gate pass rate falls from 22.9% to 13.8%; 138 of 160 RAG runs fail the gate entirely. This is a retrieval collapse: injecting retrieved context into ClinicalCamel's already-fragile generation format causes systematic output disintegration beginning around run 85, where density transitions from mixed outcomes to near-universal zero output. Figure 6 documents this collapse at the run level.

E. Downstream Utility: Train-on-Synthetic, Test-on-Real Evaluation

Figure 7 summarizes downstream T/N/M staging performance across all five adapter conditions. Each adapter is fine-tuned on its own admitted synthetic corpus—1,883 admitted records in total across conditions A–E—and evaluated under the train-on-synthetic, test-on-real (TSTR) protocol on three benchmarks of increasing distributional distance: a held-out synthetic test set ($n = 62$), TCGA Lung ($n = 737$), and TCGA Cross-Tumor ($n = 3{,}161$). Each adapter produces a T-, N-, and M-stage prediction for every note—3,960 evaluation notes per adapter, 19,800 note-level predictions across the five conditions—with per-axis scoring restricted to notes carrying a valid gold label for that axis.

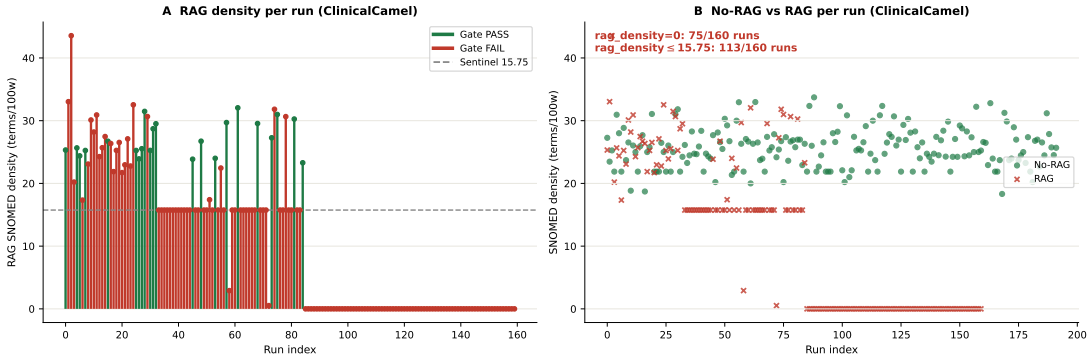

Fig. 6: ClinicalCamel-70B RAG Collapse. *(a)* RAG SNOMED density per run colored by gate outcome (green = PASS, red = FAIL). Runs 0–84 show mixed outcomes with densities around and below 15.75; from run 85 onward, density collapses to exactly 0.0 in nearly all runs, corresponding to complete output failure. *(b)* No-RAG (circles) versus RAG (crosses) density paired by run. No-RAG density is stable throughout (20–33 terms/100w); RAG density disintegrates in the second half of the run sequence. 75/160 runs produce zero SNOMED output and 113/160 produce density ≤ 15.75 under RAG.

The following patterns were observed:

First, all adapter conditions perform well in-distribution. Across the held-out synthetic benchmark, accuracy and macro-F1 remain consistently high regardless of gate configuration. Differences between adapters are modest relative to the real-world benchmarks, indicating that every synthetic corpus contains sufficient signal to support learning within the distribution from which it was generated.

Second, strong synthetic performance does not translate to TCGA Lung. Despite substantial differences in corpus quality between the five generation conditions, no adapter exceeds the majority-class baseline on any axis of the lung pathology benchmark. Improvements in schema validity, ontology compliance, and clinical-logic consistency produce cleaner synthetic corpora, but these gains do not yield commensurate improvements in lung-cancer generalization. The dominant effect is therefore the transition from synthetic generation outputs to real pathology reports rather than differences among gate configurations.

Third, the largest between-adapter differences appear on the most heterogeneous benchmark. On TCGA Cross-Tumor every adapter clears the T-stage baseline decisively (0.59–0.62 against a baseline of 0.35), the one benchmark where the synthetic corpora transfer. Adapters C and D additionally record the strongest N-stage performance, a separation not visible on the synthetic benchmark and only weakly apparent on TCGA Lung. With a single training seed per condition this gap is not separated from seed variance; we report it as an observation consistent with ontology grounding carrying portable signal across disease contexts, not as an established effect.

A consistent pattern across all conditions is the discrepancy between accuracy and macro-F1 for M-stage prediction. While M-stage accuracies appear high, macro-F1 remains substantially lower, reflecting the strong class imbalance of the evaluation sets and the predominance of M0 cases. Consequently, M-stage accuracy should not be interpreted as evidence of superior metastatic classification performance.

VI. Discussion

A. A Ranked Answer to the Component Attribution Question

The three ablations provide direct answers to RQ1, RQ2, and RQ3 while also allowing the pipeline components to be ranked by their marginal contribution to corpus quality.

Schema validation is the load-bearing constraint. Within a single corpus it rejects 148 of 512 generated records—almost all from the one generator (ClinicalCamel-70B) whose structured-output compliance is low—and the records it removes are also the ones that carry AJCC logic violations. Ontology grounding contributes a small, generator-specific filtering increment (24 Llama records) and otherwise standardizes vocabulary. AJCC clinical-logic validation contributes no marginal filtering under the conditions of this study, because the only logic-violating generator is already excluded by schema; it functions as a generator-conditional safeguard that would become load-bearing for any generator emitting schema-valid but clinically impossible records.

This ranking has a direct practical implication. A practitioner building a similar pipeline should implement schema validation first as a minimum viable quality floor, retain clinical-logic validation as a generator-conditional safeguard, treat ontology grounding as a normalization step rather than a filter, and evaluate retrieval augmentation per-generator before enabling it. The ordering is not arbitrary—it reflects the marginal contribution of each component to corpus clinical validity, measured independently under controlled conditions.

B. The Gate Filters Validity, Not Vocabulary

A consistent finding across all three ablations is that SNOMED CT density is insensitive to gate strictness. Corpora B, C, and D show densities of 29.71, 29.86, and 29.83 terms per 100 words—essentially flat, with any residual variation attributable to model mix rather than gate configuration. This directly falsifies the hypothesis that the gate enriches vocabulary by excluding low-density records.

The practical implication is that SNOMED density is not a sufficient proxy for corpus quality. A record can carry 35 SNOMED terms per 100 words while assigning M1 staging

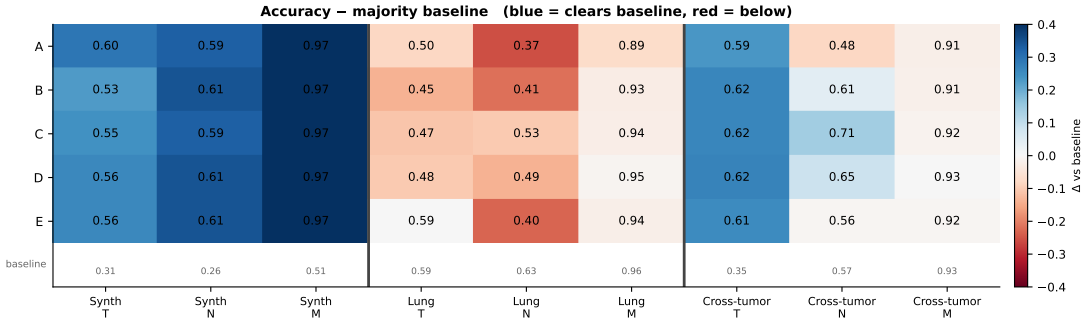

Fig. 7: Downstream TSTR by axis and cohort. Cell color is accuracy minus the majority-class baseline (blue clears the baseline, red falls below); the printed value is raw accuracy, and the bottom row gives the baseline per column. In-distribution (Synthetic) every adapter clears the baseline on all three axes; on real TCGA accuracy collapses toward the majority class, especially lung N-stage. On Cross-Tumor, ontology-aware adapters C and D most decisively clear the N-stage baseline. Per-axis macro-F1 is reported in Table VII.

without a Stage IV label—ontologically rich and clinically invalid simultaneously. Ontological richness and clinical correctness are independently evaluable properties, and conflating them leads to a false sense of corpus quality. Pipelines that report only vocabulary coverage metrics as quality indicators are measuring the wrong thing: a high SNOMED density confirms that admitted records use recognized clinical terminology, but says nothing about whether the clinical reasoning those terms encode is internally consistent.

It also separates the two constraints: O does standardization, not filtering, while C is the mechanism that would catch logical errors. The ontology gate is not redundant, but its contribution is qualitatively different from the logic gate's.

C. RAG as Gate Compliance Mechanism, Not Vocabulary Enrichment

The RAG ablation resolves a question left open by the vocabulary enrichment finding from prior work in this space: does SNOMED density enrichment from retrieval translate into downstream training signal, or does the gate already enforce sufficient ontological grounding to make retrieval redundant? The answer is model-conditional, and the model-dependency is sharp enough to constitute a design warning rather than a nuanced finding.

For Llama-3.3-70B, RAG functions as a gate compliance mechanism rather than a vocabulary enrichment tool. The 24 records that fail the ontology constraint in the No-RAG condition are resolved when retrieved PubMed context provides sufficient SNOMED grounding, lifting gate pass rate from 87.5% to 100%. Crucially, density does not increase on already-admitted records: RAG admits records that would have been excluded, not richer versions of records that would have been admitted anyway. The effect is yield improvement, not quality improvement of individual records.

For GPT-4o the gate is already satisfied at 100%, so retrieval has no failures to resolve and contributes nothing measurable.

For ClinicalCamel-70B, RAG is actively harmful. Injecting retrieved context into a generation format that is already fragile causes systematic output disintegration. The transition is abrupt: the first 85 runs show mixed outcomes with densities around and below 15.75; from run 85 onward, 75 consecutive runs produce zero SNOMED output. This is not a gradual degradation—it is a collapse, and its abruptness suggests a context-length or prompt-template threshold rather than a gradual interference effect.

The practical lesson is that retrieval augmentation should be treated as a per-generator decision rather than a pipeline default. The gate provides a natural diagnostic: if a generator's No-RAG gate compliance is already at ceiling, RAG adds no value and may introduce failure modes. If compliance is below ceiling, RAG may resolve specific constraint failures—but only if the generator can integrate retrieved context without format disintegration, which must be verified empirically before deployment.

D. Domain Pre-training and Format Compliance Are Orthogonal

ClinicalCamel-70B was included to test whether domain-adaptive clinical pre-training confers an advantage for structured oncology generation. It does not: schema compliance holds at 20–23% across every gate configuration and retrieval condition—roughly one valid record in five runs regardless of what constraints are active.

This finding challenges an assumption that is common in clinical NLP: that a model pre-trained on clinical text will produce better-structured clinical output than a general-purpose model. The evidence here suggests that domain knowledge and structured format compliance are largely orthogonal capabilities. ClinicalCamel-70B demonstrably encodes clinical knowledge— its SNOMED density on the records it does produce is comparable to Llama-3.3-70B—but its instruction-following capability and JSON format compliance are insufficient for structured generation tasks regardless of that knowledge.

The broader implication is that generator selection should be evaluated primarily on format compliance and instruction-following fidelity, not on clinical pre-training provenance. The gate enforces the same quality floor regardless of generator, but generator choice determines yield: domain pre-training is neither necessary nor sufficient for the compliance that matters here.

E. Toward Downstream Transferability

The downstream results locate the next problem precisely. The gate secures the validity of admitted records, but on real pathology reports it is the synthetic-to-real gap, not the gate configuration, that dominates T-stage performance. The generated corpus, though clinically valid, is narrow: the fixed 32-cell TNM seeding grid and templated prompts were chosen for controlled, reproducible ablation, but they bound the lexical and structural variety of the reports a model trains on, and a clean but narrow corpus transfers imperfectly to the full distribution of real pathology narratives. A second limitation concerns yield rather than validity—the records the gate discards are errors the generator still produces, rejected at a direct cost in corpus size without any change to generator behavior. A third observation is where transfer did succeed: ontology-aware filtering cleared the majority-class baseline on cross-tumor nodal staging, suggesting that ontological grounding carries portable signal across disease contexts. The central limitation is that each condition was trained once. The downstream comparisons therefore establish that a single gated run did not outperform a single ungated run on real lung notes; they do not establish that gating cannot improve transfer. Distinguishing the two requires replication across seeds, which at roughly one minute of fine-tuning per adapter is inexpensive and is the first thing we would add.

F. Future Directions

Three directions follow; The first is to generate richer notes rather than to filter harder: prompting from real pathology exemplars and institution-specific reporting styles, covering TNM combinations beyond the 32-cell grid used here, and generating a patient's record over time rather than a single snapshot, so that a model sees how stage is revised across a course of treatment. The second is to move the ontology and AJCC checks into decoding, so that the generator produces fewer invalid records instead of discarding them afterward. This raises yield, not validity: the corpus would be no more valid than the gate already makes it. The third is to widen the subject domain to tumor types whose AJCC rules are built differently—bladder, colorectal, head and neck—and to widen C from the commonly violated rules implemented here to the full staging logic. Generator-specific robustness assessment was beyond the scope of this study.

VII. Conclusion

This paper measured which components of a neuro-symbolic synthetic clinical data pipeline are load-bearing, and by how much, through five matched adapter conditions and three controlled ablations.

Four findings follow. Schema validation does most of the filtering, and because the only logic-violating generator fails schema first, it also removes the impossible-staging records before any later check reaches them. Clinical-logic validation removes nothing marginal here; it is a generator-conditional safeguard, not a high-volume filter. Ontology grounding standardizes rather than filters: SNOMED density is flat across gate levels. Retrieval augmentation is model-conditional, resolving ontology failures for one generator, adding nothing for a second, and collapsing the output of a third.

These results support a general design claim. Formalized clinical rule systems—staging manuals, coding taxonomies, drug-interaction databases—can be encoded as generation-time admission constraints rather than applied as post-hoc evaluation labels, which makes hallucination detection binary, auditable, and reproducible at the cost of reduced yield.

They also locate the remaining leverage. Across real TCGA pathology reports, gated and ungated adapters were largely indistinguishable on T-stage. The gate is where validity is secured. Downstream utility is won on the generation side: producing more clinical language, more varied in vocabulary and structure, and longitudinal rather than snapshot—records that follow a patient across a course of treatment, where stage is revised as evidence accumulates. Such a corpus still needs the gate. It also needs to be far wider than the one measured here.

Data and Code Availability

Code, generation logs with per-record gate flags, and adapter checkpoints are available at https://github.com/LGChalla/Synthetic-Data-Ablation-Oncology. Evaluation cohorts derive from The Cancer Genome Atlas, accessed under the NIH Genomic Data Sharing Policy; all generated records are synthetic and no protected health information was used.

Ethical Considerations

The *rigid.v3* schema mandates a `health_equity_factors` field, so demographic and equity attributes are model-generated rather than sampled from a reference population, and the gate does not constrain them: it validates staging logic and ontology coverage, not whether generated associations between demographic attributes and stage, histology, or treatment intent reproduce biases in the generators' pre-training data. We did not audit for such associations, and corpora produced this way should not be used to study disparities without one. This is the construct problem the paper documents for SNOMED density—a constraint certifies only what it is defined over—and extending admission constraints to demographic plausibility follows directly from the mechanism described here.

Appendix A
Entropy Floor Derivation and Label Diversity

Entropy floors are set at 80% of maximum entropy per TNM dimension. For T and N ($k = 4$): $H_{\max} = \ln 4 = 1.386$ nats, floor $= 1.109$. For M ($k = 2$): $H_{\max} = \ln 2 = 0.693$ nats, floor $= 0.554$. The floors are checked twice—after generation, on the admitted corpus before the train/test split, and before training, on the training partition after stratified splitting—with fine-tuning aborted if any dimension is single-class. No entropy intervention was triggered at either checkpoint in any of the five adapter conditions.

TABLE VI: Label diversity (Shannon entropy, nats) of the admitted corpus per TNM axis. All conditions sit at the theoretical maxima ($\ln 4 = 1.386$ for T,N; $\ln 2 = 0.693$ for M), well above the 80% floors, because the 32-cell seed grid balances labels by construction.

Adapter	H_T	H_N	H_M
A–E (all)	1.39	1.39	0.69
Floor (0.8×max)	1.109	1.109	0.554

APPENDIX B
PER-AXIS MACRO-F1 SCORES

Table VII reports per-axis macro-F1 for all five adapters under the TSTR protocol, complementing the accuracy heatmap in Fig. 7.

TABLE VII: Per-axis macro-F1 for all five adapters under TSTR, across the three evaluation cohorts. Macro-F1 weights each class equally, making it sensitive to minority-class recovery in a way aggregate accuracy is not—particularly for M-stage, where the M0 majority inflates accuracy while macro-F1 stays low.

	Synthetic			TCGA Lung			Cross-Tumor		
Adapter	T	N	M	T	N	M	T	N	M
A	0.56	0.60	0.97	0.41	0.33	0.53	0.56	0.53	0.67
B	0.50	0.61	0.97	0.40	0.33	0.59	0.58	0.56	0.66
C	0.48	0.59	0.97	0.41	0.41	0.59	0.59	0.64	0.67
D	0.50	0.62	0.97	0.42	0.39	0.58	0.59	0.60	0.66
E	0.53	0.61	0.97	0.44	0.33	0.58	0.57	0.55	0.69

APPENDIX C
QLORA CONFIGURATION

All five adapters (A–E) are fine-tuned from the same base model using identical QLoRA hyperparameters [41], [42] (Table VIII), ensuring performance differences across adapter conditions are attributable to corpus quality rather than training procedure. Training seed 42 is applied identically across all adapters, so differences between adapters are not attributable to seed selection. A single seed per condition does not, however, bound seed-to-seed variance itself.

APPENDIX D
PROMPT DETAILS

This appendix documents the two prompt families used in the pipeline. The structured-generation prompt conditions each generator on a single TNM seed drawn from the 32-cell grid (Section IV-B) and instructs it to return a complete record conforming to the *rigid.v3* schema; in the RAG condition (Adapter E), the top-k MedCPT-retrieved PubMed abstracts are prepended to this template as grounding context preceding the instruction. The extraction prompt elicits TNM staging from real clinical notes under the TSTR protocol.

Structured Generation Prompt (*rigid.v3*)

```
You are a clinical oncology data generator.
Generate one synthetic non-small-cell lung cancer
patient record as a single JSON object that
conforms exactly to the rigid.v3 schema. The
record must be internally consistent with AJCC 8th
Edition staging. Seed: T={T}, N={N}, M={M}.
Required fields: tnm_staging, histology (ICD-O-3
and SNOMED CT codes), molecular_drivers,
demographics, imaging_findings,
treatment_modalities, health_equity_factors, and
narrative. Return only the JSON object with no
commentary.
```

TABLE VIII: QLoRA fine-tuning configuration, identical across all five adapters (A–E).

Setting	Value
Base model and Quantization	
Base model	`meta-llama/Meta-Llama-3-8B-Instruct`
quantization	4-bit NF4, double quantization
Compute dtype	bfloat16
LoRA	
Rank (r)	16
Scaling (α)	32
Dropout	0.05
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`
Trainable params	13,631,488 / 8,043,892,736 (0.1695%)
Optimization	
Epochs	3
Learning rate	2×10^{-4}, cosine decay
Batch size	2 per device
Grad. accum.	4 (effective batch size 8)
Warmup steps	10
Optimizer	`paged_adamw_8bit`
Precision	fp16 mixed
Max seq. length	1,024 tokens
Random seed	42 (all adapters)
Hardware and runtime	
GPU	Single NVIDIA H200
Training time	≈60 s per adapter; <5 min total

TSTR Extraction Prompt

```
You are a clinical data extractor. Read the
clinical note and extract the TNM staging. Return
a strictly formatted JSON object with keys 'T',
'N', and 'M'. Always use the full prefixed format
(e.g., T2, N0, M0). If a value is not found, use
'Unknown'.
```

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] National Cancer Institute, "SEER cancer stat facts: Lung and bronchus cancer," Surveillance, Epidemiology, and End Results Program, Bethesda, MD, 2024, survival and stage distribution from SEER 22 (excluding IL/MA), 2014–2020. [Online]. Available: https://seer.cancer.gov/statfacts/html/lungb.html

[3] I. Hands and R. Kavuluru, "A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications," *Artificial Intelligence Review*, vol. 58, no. 10, Art. no. 314, 2025.

[4] M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen, "Challenges and opportunities beyond structured data in analysis of electronic health records," *WIREs Computational Statistics*, vol. 13, no. 6, Art. no. e1549, 2021.

[5] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature Medicine*, vol. 25, no. 1, pp. 37–43, 2019.

[6] J. M. Unger, L. N. Shulman, M. A. Facktor, H. Nelson, and M. E. Fleury, "National estimates of the participation of patients with cancer in clinical research studies based on Commission on Cancer accreditation data," *Journal of Clinical Oncology*, vol. 42, no. 18, pp. 2139–2148, 2024.

[7] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, 2021.

[8] S. Liang, J. Zhang, X. Liu, Y. Huang, J. Shao, X. Liu, W. Li, G. Wang, C. Wang *et al.*, "The potential of large language models to advance precision oncology," *eBioMedicine*, vol. 115, p. 105695, 2025.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[10] American Joint Committee on Cancer, *AJCC Cancer Staging Manual*, 8th ed. Springer, 2017.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[12] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, 2024.

[13] Author names withheld, "Neuro-symbolic synthetic clinical data generation for lung cancer TNM staging," 2026, manuscript under review. Author identities withheld to preserve anonymity of a concurrent submission and will be restored in the camera-ready version.

[14] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Proceedings of the 2nd Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 68. PMLR, 2017, pp. 286–305.

[15] M. K. Baowaly, C.-C. Lin, C.-L. Liu, and K.-T. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019.

[16] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data," *Neurocomputing*, vol. 416, pp. 244–255, 2020.

[17] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digital Health*, vol. 2, no. 1, p. e0000082, 2023.

[18] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 230–238, 2018.

[19] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of LLMs help clinical text mining?" *arXiv preprint arXiv:2303.04360*, 2023.

[20] S. N. Kasthurirathne, B. Mamlin, and P. Biondich, "Improving the ability of US healthcare systems to address the covid-19 pandemic using synthetic patient data," *Journal of the American Medical Informatics Association*, vol. 30, no. 1, pp. 118–127, 2023.

[21] G. Kumichev, P. Blinov, Y. Kuzkina, V. Goncharov, G. Zubkova, N. Zenovkin, A. Goncharov, and A. Savchenko, "MedSyn: LLM-based synthetic medical text generation framework," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2024, pp. 215–230.

[22] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[23] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[24] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: Enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W541–W545, 2011.

[25] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical domain hallucination test for large language models," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023, pp. 314–334.

[26] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, P. Mansfield, G. S. Corrado, Y. Matias, V. Natarajan *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[27] S. Pandit, J. Xu, J. Hong, Z. Wang, T. Chen, K. Xu, and Y. Ding, "MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025, pp. 2858–2873.

[28] E. Asgari, S. Khalil, N. Montaña-Brown, M. Dubois, J. Balloch, J. A. Yeung, and D. Pimenta, "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," *npj Digital Medicine*, vol. 8, no. 1, p. 274, 2025.

[29] G. Mehenni and A. Zouaq, "Ontology-constrained generation of domain-specific clinical summaries," in *International Conference on Knowledge Engineering and Knowledge Management*. Springer, 2024, pp. 382–398.

[30] A. d. Garcez and L. C. Lamb, "Neural-symbolic learning and reasoning: A survey and interpretation," *Neurocomputing*, vol. 479, pp. 132–145, 2022.

[31] Y. Yu, "Hybrid-code: A privacy-preserving, redundant multi-agent framework for reliable local clinical coding," *arXiv preprint arXiv:2512.23743v2*, 2026.

[32] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu, "MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, p. btad651, 2023.

[33] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," *arXiv preprint arXiv:1706.02633*, 2017.

[34] D. Hu, H. Zhang, S. Li, Y. Wang, N. Wu, and X. Lu, "Automatic extraction of lung cancer staging information from computed tomography reports: Deep learning approach," *JMIR Medical Informatics*, vol. 9, no. 7, Art. no. e27955, 2021.

[35] S. Kim, S. Jang, B. Kim, L. Sunwoo, S. Kim *et al.*, "Automated pathologic TN classification prediction and rationale generation from lung cancer surgical pathology reports using a large language model fine-tuned with chain-of-thought: Algorithm development and validation study," *JMIR Medical Informatics*, vol. 12, Art. no. e67056, 2024.

[36] SNOMED International, "What is SNOMED CT?" London, U.K., 2026. [Online]. Available: https://www.snomed.org/what-is-snomed-ct. Accessed: Aug. 21, 2026.

[37] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[38] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[39] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, "Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding," *arXiv preprint arXiv:2305.12031*, 2023.

[40] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss library," *arXiv preprint arXiv:2401.08281*, 2024.

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[42] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 10 088–10 115.