

# Embeddings to Diagnosis: Latent Fragility under Agentic Perturbations in Clinical LLMs

Raj Krishnan Vijayaraj  
raj.vijayaraj@alumni.utoronto.ca  
Independent Researcher  
Toronto, Ontario, Canada

## Abstract

LLMs for clinical decision support often fail under small but clinically meaningful input shifts such as masking a symptom or negating a finding, despite high performance on static benchmarks. These reasoning failures frequently go undetected by standard NLP metrics, which are insensitive to latent representation shifts that drive diagnosis instability. We propose a geometry-aware evaluation framework, LAPD (Latent Agentic Perturbation Diagnostics), which systematically probes the latent robustness of clinical LLMs under structured adversarial edits. Within this framework, we introduce Latent Diagnosis Flip Rate (LDFR), a model-agnostic diagnostic signal that captures representational instability when embeddings cross decision boundaries in PCA-reduced latent space. Clinical notes are generated using a structured prompting pipeline grounded in diagnostic reasoning, then perturbed along four axes—masking, negation, synonym replacement, and numeric variation to simulate common ambiguities and omissions. We compute LDFR across both foundation and clinical LLMs, finding that latent fragility emerges even under minimal surface-level changes. Finally, we validate our findings on 90 real clinical notes from the DiReCT benchmark (MIMIC-IV), confirming the generalizability of LDFR beyond synthetic settings. Our results reveal a persistent gap between surface robustness and semantic stability, underscoring the importance of geometry-aware auditing in safety-critical clinical AI.

## CCS Concepts

• **Computing methodologies** → **Language models**; *Machine learning algorithms*; *Robustness*; *Neural networks*.

## Keywords

LLM Evaluation, Diagnostic Reasoning, Latent Space Analysis, Model Robustness, Perturbation Analysis, Clinical NLP, Explainable AI, AI Safety

## ACM Reference Format:

Raj Krishnan Vijayaraj. 2025. Embeddings to Diagnosis: Latent Fragility under Agentic Perturbations in Clinical LLMs. In *Proceedings of Agentic & GenAI Evaluation Workshop at KDD 2025 (Agentic & GenAI Evaluation at*

*KDD 2025*). ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

## 1 Introduction

LLMs like GPT-4[16] achieve near-human accuracy on medical benchmarks such as CMExam and MedQA[6], but these static metrics do not assess stability under clinically realistic variation. In high-stakes environments, robustness is very critical, like accuracy. While benchmarks provide a point estimate of correctness, they fail to measure whether models produce consistent diagnoses when inputs vary in natural but subtle ways.

Unlike output-level agreement metrics, LDFR reveals latent representation shifts that signal unstable diagnostic reasoning. Common metrics such as F1, EM, BERTScore[29], BLEU[17], and ROUGE[11] focus on surface similarity, not deeper clinical reasoning. They overlook shifts introduced by negation (e.g., “denies chest pain”), masking (e.g., privacy redactions), or synonym substitution (“heart attack” vs “myocardial infarction”). As a result, models may score highly yet remain brittle to changes that are common in clinical documentation.

We refer to this vulnerability as **diagnostic fragility**: when small, clinically grounded edits to input cause significant changes in the model’s output diagnosis. MedFuzz[15] showed that even high-performing LLMs exhibit such volatility, but prior work has focused on whether the output label changes without examining how the internal representation shifts.

We argue that robustness also involves stability in the model’s internal reasoning. To capture this, we go beyond output agreement and introduce a metric that measures latent instability: the **Latent Diagnosis Flip Rate (LDFR)**. LDFR quantifies how often a structured perturbation causes the latent embedding to cross a diagnostic decision boundary in PCA-reduced space. These flips reveal underlying model sensitivity even when surface-level similarity remains high.

This motivates our framework: **LAPD (Latent Agentic Perturbation Diagnostics)**. LAPD uses synthetic clinical notes generated from an LLM-based pipeline and applies structured perturbations masking, negation, synonym replacement, and numerical edits at controlled thresholds. These edits are not semantically neutral, but deliberately designed as stressors to probe stability in diagnostic reasoning, as outlined in Figure 1. To quantify this instability, we introduce the **Latent Diagnosis Flip Rate (LDFR)**, which measures how often perturbations cause embeddings to cross diagnostic decision boundaries in latent space.

In our analysis, we find that entity masking and negation consistently trigger large changes in LDFR, even when surface metrics like BERTScore remain above 0.9. Latent decision flips tend to occur along high-variance PCA axes, indicating that a small number of

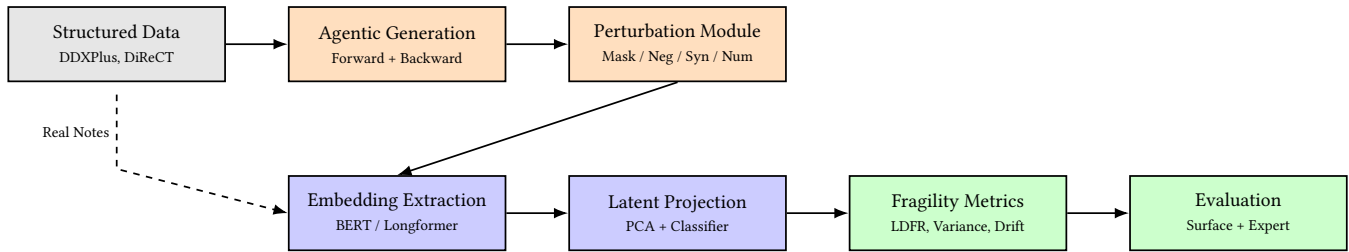
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*Agentic & GenAI Evaluation at KDD 2025, Toronto, Canada*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>



**Figure 1: Overview of the LAPD evaluation pipeline. Synthetic or real clinical notes are processed through structured perturbation and latent embedding projection. The resulting representations are analyzed for fragility using geometry-aware metrics (e.g., LDFR) and surface-level clinical agreement.**

directions in the embedding space explain most of the instability. Importantly, these patterns generalize from synthetic notes to real clinical records from MIMIC-IV (DiReCT), confirming the broader applicability of our findings.

Our main contributions are:

- We present **LAPD**, a geometry-aware evaluation framework that probes the latent robustness of clinical LLMs using structured, clinically grounded perturbations.
- We introduce the **Latent Diagnosis Flip Rate (LDFR)**, a model-agnostic metric that captures semantic instability through boundary crossings in PCA-reduced embedding space.
- We show that LDFR exposes diagnostic fragility overlooked by surface metrics, and generalizes to real clinical notes from DiReCT supporting its use in auditing clinical LLMs.

## 2 Related Work

Robustness evaluation of clinical LLMs has primarily focused on surface-level stability under textual perturbations. Techniques such as MedFuzz [15] and PerturbScore [8] assess performance degradation under attribute-level fuzzing or paraphrase variations, but rely on output fidelity metrics like BERTScore or NER-F1. These approaches do not capture whether the underlying diagnostic reasoning process remains intact, nor whether latent representations remain within stable decision regions. Moreover, these evaluations are typically conducted on pre-curated datasets, rather than generated in an interactive, agent-driven pipeline.

Early representation methods such as Med2Vec [2] and transformer-based EHR embeddings [27] laid the groundwork for patient-level understanding, but they were not designed to capture robustness under naturalistic variation.

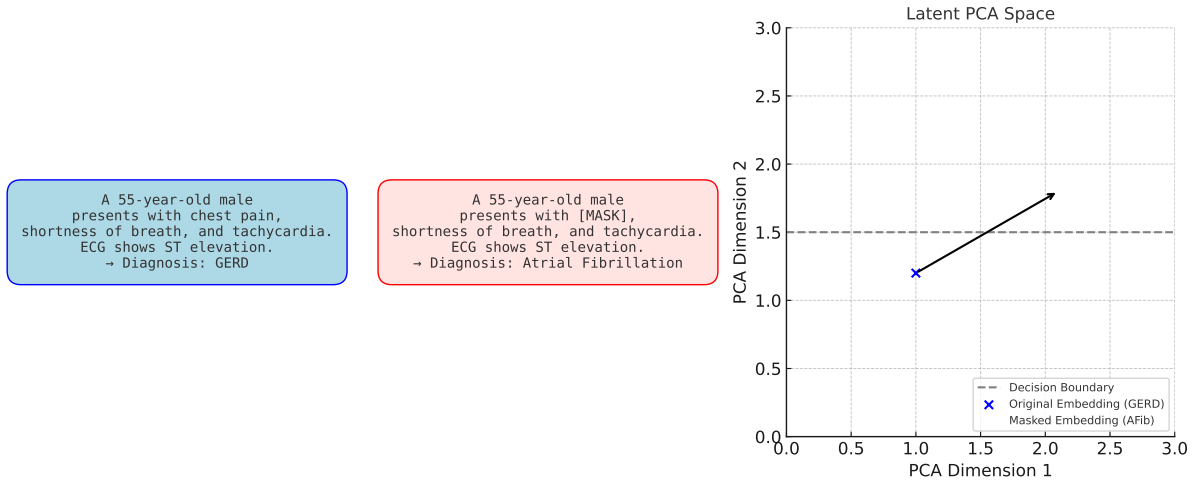
Recent work has explored geometry-aware metrics to analyze latent drift and embedding robustness. Studies have used PCA variance [28], Procrustes distance [5], and low-rank subspace methods to characterize representation shifts. Manifold-based techniques [20] and vocabulary unification strategies [7] have further aimed to align clinical semantics, though their robustness implications remain underexplored. The work by Liu et al. [12] extends latent probing to adversarial reasoning stability by testing logical consistency properties such as transitivity and negation invariance across LLM outputs. However, their framework focuses on abstract reasoning tasks and does not ground latent shifts in specific clinical

diagnostic decisions, as our LDFR metric does. Our proposed Latent Diagnosis Flip Rate (LDFR) fills this gap by detecting whether controlled, clinically plausible perturbations cause embeddings to cross decision boundaries tied to actual diagnostic outcomes. Complementary efforts in probing semantic generalization [18] similarly explore latent-space reasoning, though not in diagnosis-specific settings.

In broader evaluation research, tools like ROSCOE [4], ReCEval [19], and CheckList [21] test reasoning consistency and behavioral coverage, while factuality frameworks such as FactScore and FEVER assess correctness in generated content. TextFlint [25] and TextAttack [14] provide adversarial input generation pipelines, but lack mechanisms to evaluate whether such perturbations lead to meaningful shifts in diagnostic decision-making. These methods are valuable for general-purpose LLM auditing, but are not designed to evaluate diagnostic integrity or label alignment in latent space. Unlike these, LDFR focuses on *reasoning stability* rather than hallucination detection or surface fluency.

A growing body of work now leverages agent-based frameworks not just for generation, but also for evaluation. **Agent-as-a-Judge framework** [30] proposes a process in which agents autonomously assess the reasoning accuracy of other agents’ output, creating a scalable and automated audit loop. This paradigm inspires our own design, where the same LLM agents generate reasoning-grounded clinical notes and perform structured adversarial perturbations for robustness assessment. [26] introduces a latent chain probing strategy to assess LLM outputs using internal embedding dynamics rather than surface labels. Their findings support our core claim that latent shifts rather than just output deviations can serve as a diagnostic signal of reasoning instability. Our framework integrates these ideas by applying geometric probes to detect clinically-relevant diagnosis flips in a label-aware latent space. Related retrieval-augmented models like Med-PaLM [22] and clinical entity linking approaches [13] enhance factual grounding in LLMs, but remain largely unevaluated under perturbation.

Finally, the need for robust diagnostic evaluation arises across clinical NLP benchmarks. While tasks like MedQA [6], CMExam, and MedNLI focus on accuracy or entailment under static conditions, recent studies [3, 9] show that model representations in biomedical contexts are often unstable. Transformer-based clinical documentation systems [1, 10, 24] exhibit high performance on static tasks, but their internal consistency under adversarial conditions remains underexamined. LAPD unifies these strands by



**Figure 2: Overview of our framework (LAPD) through a constructed example illustrating latent diagnostic fragility.** The left panel shows a synthetic clinical note generated by our prompting pipeline and diagnosed as GERD. The center panel applies masking to a key symptom (“chest pain”), resulting in a surface-similar variant that elicits a different diagnosis: Atrial Fibrillation. In the right panel, PCA-projected latent embeddings show that this perturbation crosses a diagnostic decision boundary. While this example is illustrative and not drawn from the evaluation set, it reflects a broader trend observed in our results: entity masking can cause latent representation shifts that lead to diagnosis flips, even under minor surface edits.

examining whether reasoning remains geometrically stable under adversarial perturbations.

### 3 Agentic Generation and Perturbation Setup

#### 3.1 Agentic Simulation of Diagnostic Notes

The LAPD pipeline is shown in Figure 1. Agentic note generation (left) begins with structured prompting grounded in DDXPlus, producing diagnosis-grounded narratives. Perturbation modules (middle) apply perturbation edits along four semantic axes. These perturbed notes are passed through frozen clinical encoders (right), and their latent representations are evaluated via PCA-based classification, variance analysis, and latent flip detection (LDFR).

We generate structured clinical notes using a reasoning-augmented agentic pipeline grounded in the DDXPlus dataset [23]. Each note is produced through three chained prompts: (1) forward reasoning (symptoms to diagnosis), (2) backward justification (diagnosis to rationale), and (3) narrative construction. This agentic simulation creates interpretable, diagnosis-grounded synthetic notes while avoiding direct reuse of patient records. These controlled inputs enable systematic robustness evaluation. Full prompting templates are available in Appendix A.2.

#### 3.2 Clinically Grounded Perturbation Strategies

We apply four types of clinically motivated perturbations to test robustness:

- **Entity Masking (Omission):** Replaces medically salient entities with [MASK].
- **Negation (Omission):** Reverses polarity of symptoms, e.g., “has chest pain” → “no chest pain.”

- **Synonym Replacement (Substitution):** Substitutes phrases with equivalent clinical terms.
- **Numerical Perturbation (Distortion):** Alters vitals/labs by  $\pm 5\%$ – $15\%$ .

Each method is applied at 0%–100% intensity by proportion of entities perturbed. For example, at 50% threshold, half of all extractable clinical entities are modified. These edits are *not* meant to preserve meaning, but to stress-test the LLM’s internal diagnostic consistency.

#### 3.3 Evaluation on Real Clinical Notes

To test robustness beyond synthetic data, we apply the same pipeline to 90 notes from the DiReCT benchmark based on MIMIC-IV. These notes average 870 tokens and include complex entity structures. We use Clinical-Longformer to encode full-length notes without truncation. This setting approximates real-world usage and complements our controlled synthetic evaluations.

### 4 Latent Robustness Evaluation Framework

#### 4.1 Surface-Level Evaluation Baselines

We compute the following standard metrics to quantify textual fidelity:

- **ROUGE-L, BERTScore (F1):** Lexical and semantic similarity.
- **NER F1, Jaccard Index:** Entity preservation.
- **Centroid Drift:** Euclidean distance between original and perturbed note embeddings.

These serve as baselines. However, they fail to capture semantic instability where diagnosis changes despite surface similarity.

## 4.2 Latent Representation and PCA Probing

We embed notes using frozen ClinicalBERT (synthetic) or Clinical-Longformer (real). PCA is applied to unperturbed embeddings to retain 90% variance, yielding 30–40 principal components. A logistic regression classifier is trained on these projections to predict original diagnoses. This classifier models latent decision boundaries. It is used solely for probing, not as a substitute model.

To ensure classifier robustness, we validate performance via 10-fold cross-validation. Hyperparameters are reported in Appendix A.6.

## 4.3 Latent Diagnosis Flip Rate (LDFR)

We define LDFR as the proportion of perturbed samples whose latent representations cross the diagnostic boundary:

$$\text{LDFR}(t) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[d_0^{(i)} \neq d_t^{(i)}]$$

where  $d_0^{(i)}$  and  $d_t^{(i)}$  are classifier predictions for unperturbed and perturbed notes, respectively. We also compute **DFR**, the LLM’s own diagnostic flip rate, to assess alignment with boundary-crossing behaviour.

## 4.4 Complementary Latent Metrics

We further interpret latent robustness through:

- **Centroid Displacement:** Mean embedding shift across perturbations.
- **Per-Dimension Variance:** Encoder output dimensional variance reveals which directions drive instability.

## 4.5 Clinician Evaluation of Diagnostic Realism

Two clinicians evaluated five synthetic notes on structure, diagnostic clarity, and reasoning depth (0–3 scale). Scores ranged from 2.0–2.7. Qualitative feedback noted missing vitals, incomplete differentials, and unrealistic symptom presentation. This affirms the use of controlled perturbations to surface latent fragility. Reviewers were independent and unaffiliated with the project.

# 5 Experimental Setup

We now describe the experimental setup used to instantiate the evaluation framework described in Section 4. This includes the choice of language models, dataset sources, perturbation statistics, and embedding details.

## 5.1 Model Selection

We evaluate two categories of models:

### General-Purpose Foundation Models:

gpt-3.5-turbo-0125, gpt-4o-mini-2024-07-18, llama-3.1-8b-instant, and open-mistral-7b are included to assess fragility at the foundational level before domain-specific adaptation. This helps reveal vulnerabilities that may persist across deployment contexts.

### Clinical-Specific Models:

medgemma-27b is tested to benchmark robustness among medical-domain fine-tuned models. These help evaluate if clinical training enhances perturbation resistance.

All models are accessed through inference APIs or local deployment, with temperature fixed at 0 or 0.01 and max token length of 512.

## 5.2 Dataset Sources

### Agentic Synthetic Notes:

Derived from DDXPlus dialogues across 7 diagnoses: Pneumonia, Pulmonary Embolism, Atrial Fibrillation, Tuberculosis, GERD, Asthma, COPD. We generate 100 notes per diagnosis, each averaging 410 tokens (SD = 56) and containing 8 extractable clinical entities.

### Real Clinical Notes:

We evaluate 90 discharge summaries from the DiReCT benchmark (MIMIC-IV). These average 870 tokens (SD = 142) and contain 13.2 named entities per note. This enables testing on longer, messier, and more realistic inputs.

## 5.3 Perturbation Procedure and Matrix

We apply four perturbation types (defined in Section 4.2) masking, negation, synonym replacement, and numerical edits, at 5 intensity thresholds (0%, 25%, 50%, 75%, 100%).

### Perturbation Matrix Statistics:

- **Synthetic notes:** 100 unperturbed notes  $\times$  4 perturbation types  $\times$  5 thresholds = 2,000 perturbed samples.
- **Real notes:** 90 notes  $\times$  4 types  $\times$  5 thresholds = 1,800 additional samples.

We extract entities using a clinical NER pipeline, and perturbations are applied proportionally to threshold.

## 5.4 Embedding and Classifier Configuration

Synthetic notes are encoded using ClinicalBERT; real notes are encoded with Clinical-Longformer to avoid truncation. All embeddings are frozen. PCA is applied to unperturbed samples, retaining 90% variance (typically 30–40 components). A logistic classifier is trained in this space to estimate latent boundaries.

Classifier performance is validated via 5-fold cross-validation, and hyperparameter details are in Appendix A.6.

# 6 Results

To assess the robustness of diagnostic reasoning in LLMs, we evaluate model behavior under structured perturbations across both surface-level and latent metrics. Our goal is to understand not only whether predictions change, but also how internal representations shift even when the text appears similar on the surface. We begin by comparing traditional NLP metrics with our latent diagnostic flip rate (LDFR) to uncover hidden fragility that surface scores may overlook.

## 6.1 Surface vs Latent Metrics

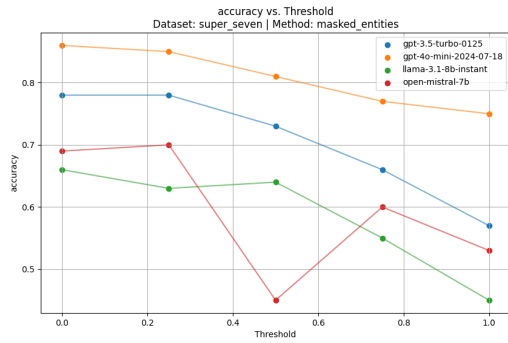
Appendix A.5 shows that surface metrics such as ROUGE-L and BERTScore remain high even when the model’s diagnosis changes. For example, at 25% masking, BERTScore stays above 0.89 while diagnosis accuracy drops significantly. At the same time, LDFR decreases, indicating that latent agreement breaks down even when surface similarity remains high. This highlights a key limitation

**Table 1: Boundary crossings in latent space predict real diagnostic flips. This table reports Pearson and Spearman correlations between PCA-based classifier predictions and LLM outputs under increasing perturbation thresholds. Masked entities cause the sharpest drop in correlation, suggesting nonlinear semantic drift, while other perturbations (e.g., numerical) preserve alignment, highlighting varying degrees of latent instability.**

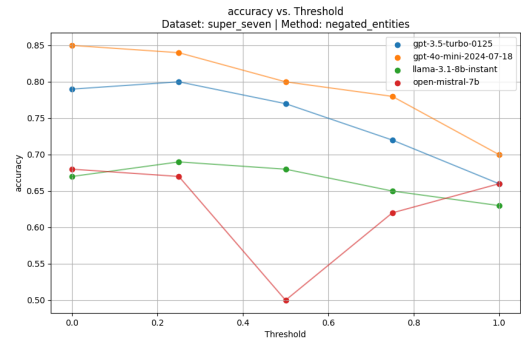
Method	Threshold	Pearson	Spearman	LDFR
Masked Entities	0.00	0.944	0.944	0.9125
	0.25	0.493	0.493	0.5525
	0.50	0.380	0.380	0.4900
	0.75	0.322	0.322	0.4075
	1.00	0.317	0.317	0.3525
Negated Entities	0.00	0.944	0.944	0.9125
	0.25	0.848	0.848	0.8525
	0.50	0.744	0.744	0.8000
	0.75	0.709	0.709	0.7375
	1.00	0.702	0.702	0.7050
Synonym Replacement	0.00	0.944	0.944	0.9125
	0.25	0.845	0.845	0.8750
	0.50	0.793	0.793	0.8400
	0.75	0.728	0.728	0.7950
	1.00	0.692	0.692	0.7325
Numerical Perturbations	0.00	0.944	0.944	0.9125
	0.25	0.894	0.894	0.8925
	0.50	0.853	0.853	0.8750
	0.75	0.900	0.900	0.8950
	1.00	0.880	0.880	0.8925

**Table 2: LDFR alignment with LLM predictions on real clinical notes (DiReCT).**

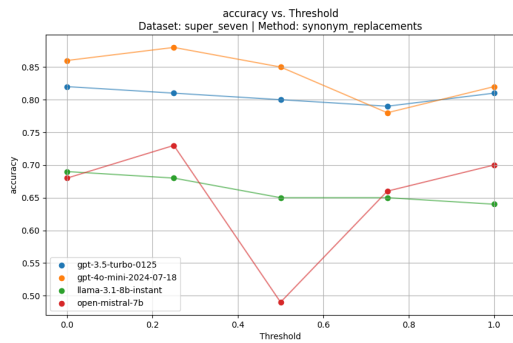
Method	Threshold	Pearson	Spearman	LDFR
Masked Entities	0.00	0.5456	0.5456	0.6923
	0.25	0.2640	0.2640	0.5187
	0.50	0.3001	0.3001	0.5143
	0.75	0.3377	0.3377	0.4681
	1.00	0.2601	0.2601	0.4198
Negated Entities	0.00	0.5456	0.5456	0.6945
	0.25	0.1621	0.1621	0.4000
	0.50	0.2991	0.2991	0.4110
	0.75	0.2058	0.2058	0.3956
	1.00	0.2130	0.2130	0.4198
Synonym Replacement	0.00	0.5456	0.5456	0.6989
	0.25	0.5270	0.5270	0.6879
	0.50	0.4586	0.4586	0.6352
	0.75	0.3406	0.3406	0.6022
	1.00	0.4032	0.4032	0.6022
Numerical Perturbations	0.00	0.5590	0.5590	0.6923
	0.25	0.5270	0.5270	0.6857
	0.50	0.5079	0.5079	0.6769
	0.75	0.5079	0.5079	0.6989
	1.00	0.4823	0.4823	0.6813



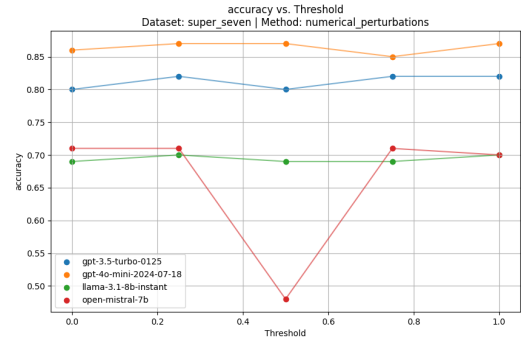
(a) Masked Entities: Most disruptive perturbation. Steep accuracy drop across all models suggests LLMs heavily rely on explicitly stated entities for diagnosis.



(b) Negated Entities: Moderate performance drop. Indicates partial sensitivity to polarity shifts like symptom presence vs. absence.



(c) Synonym Replacements: Model-dependent effects. Some models show resilience while others misinterpret lexical variants.



(d) Numerical Perturbations: Least disruptive. Accuracy remains mostly stable, suggesting LLMs underutilize quantitative signals in clinical text.

**Figure 3: LLMs show inconsistent resilience to clinically realistic perturbations. This figure illustrates how diagnostic accuracy degrades as we increase perturbation intensity (0–100%) across four types. Masked entities cause the sharpest performance drop, while numerical edits have minimal effect, implying a greater LLM reliance on qualitative than quantitative reasoning.**

of surface-level evaluation in clinical reasoning: it misses internal inconsistencies that lead to different predictions.

Figure 3 confirms this pattern across perturbation types. Masking causes the sharpest accuracy drop, while numerical changes have little effect, suggesting models may over-rely on surface cues and underuse numerical information. To better understand how this latent disagreement arises, we now examine the effect of specific perturbation types.

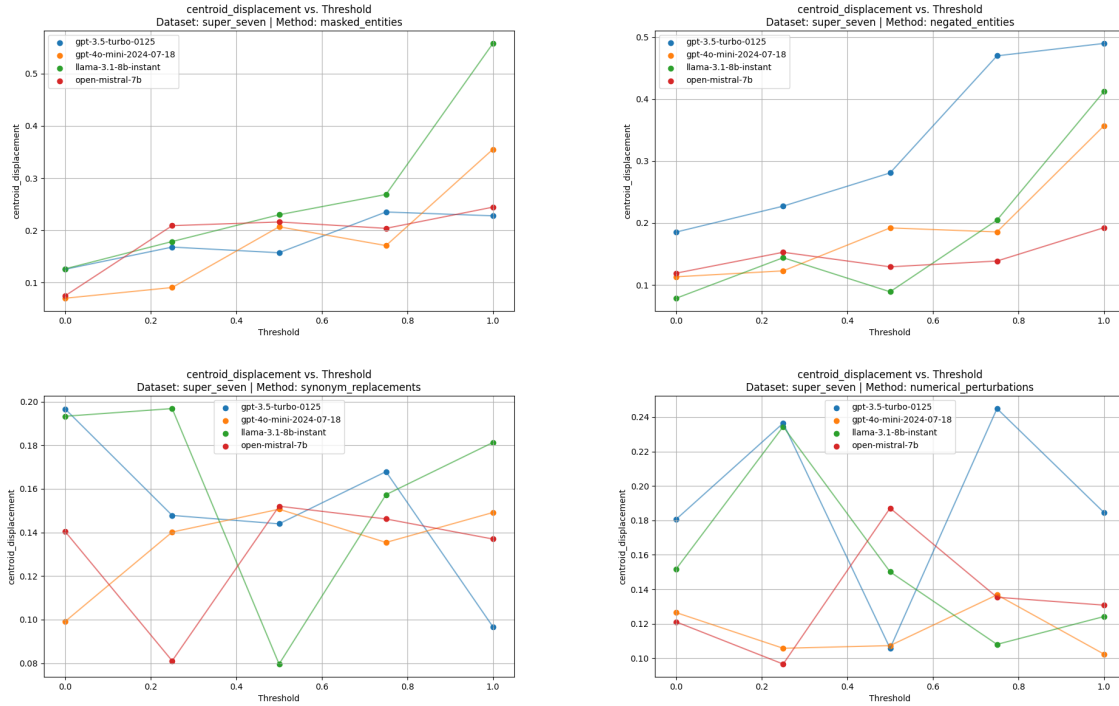
## 6.2 Fragility by Perturbation Type

Masked entities lead to the largest drop in both accuracy and LDFR as perturbation increases. This indicates strong dependence on explicitly stated symptoms. Negation causes a smaller but consistent drop in LDFR and accuracy, suggesting that polarity cues impact internal reasoning. Synonym and numeric changes maintain high ROUGE and BERTScore values, but still result in latent shifts, especially at higher perturbation levels. Table 1 summarizes these

trends, showing how latent alignment degrades differently depending on the type of edit. Having established that perturbation types induce varied latent instability, we next ask: do different models handle this stress differently?

## 6.3 Fragility by Model

Model behavior under perturbation varies. GPT-4o and GPT-3.5-turbo show lower LDFR and smaller accuracy drops for synonym and numeric edits. In contrast, LLaMA-3.1 and Mistral are more fragile under masking and negation, with LDFR values exceeding 0.5 in some cases. These trends are visualized in Figure 3, showing how different models react to increasing perturbation severity in accuracy. Notably, Mistral-7B, being a smaller and earlier-generation model, demonstrates poorer robustness, suggesting that certain diagnostic reasoning capabilities may emerge only at larger model scales.



**Figure 4: Global embedding shifts do not directly predict diagnostic failures.** This figure shows how perturbations move embeddings away from their original centroid, measured by Euclidean distance. While masking and negation induce smooth, monotonic drift, these displacements are weakly correlated with diagnosis changes, suggesting that even small shifts can cross latent decision boundaries and lead to fragility.

#### 6.4 Latent Dimensional Collapse by Perturbation Type

To complement global centroid drift, we analyze how variance re-distributes across PCA dimensions under increasing perturbation (Figure 5). Under **masked entities**, up to 20–30% of total variance is captured by a single dimension (e.g., dimension 752 in GPT-3.5 and LLaMA), indicating a bottleneck where embeddings compress into narrow subspaces. Negation and synonym replacements show model-specific trends: GPT-4o and Mistral maintain smoother variance distributions, while GPT-3.5 and LLaMA concentrate variance more tightly. Numerical perturbations produce flat variance curves across all models, reflecting stable and diffuse latent encodings.

These structured shifts align with our LDFR findings: as variance concentrates in fewer dimensions, LDFR increases, especially under masking and negation. Importantly, these effects occur within the 90% variance PCA subspace (Appendix A.9), preserving interpretability and classifier stability. In contrast, global centroid displacement (Figure 4) increases smoothly but does not predict diagnostic changes. Together, these results suggest that perturbations expose low-dimensional instability patterns invisible to surface-level metrics.

#### 6.5 Real vs. Synthetic Notes

A key concern is whether our findings on synthetic notes hold for real clinical documentation. To test this, real clinical notes from the

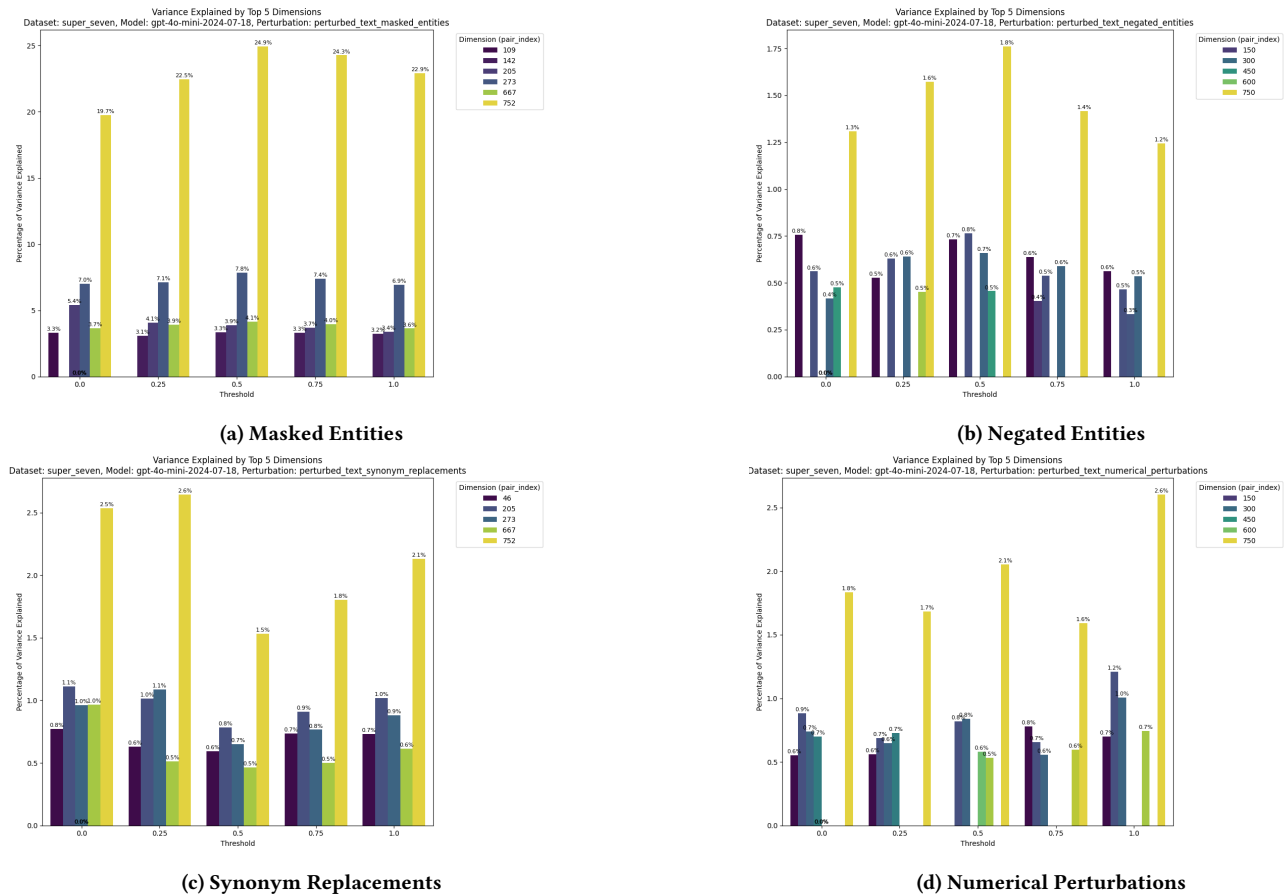
**Table 3: Average tokens and entities per chunk for real and synthetic notes.**

Type	Avg. Tokens	Avg. Entities
DiReCT	869.76	83.51
synthetic_open-mistral-7b	369.97	30.16
synthetic_gpt-4o-mini-2024-07-18	533.28	27.49
synthetic_gpt-3.5-turbo-0125	328.36	31.98
synthetic_llama-3.1-8b-instant	554.54	22.10

DiReCT dataset show similar patterns of latent fragility as synthetic notes. As shown in Table 2, correlations between LDFR and model predictions consistently drop as perturbations increase, confirming that our metric captures instability beyond synthetic settings.

Masked entity edits cause the largest decline in both cases, though the drop is more severe in synthetic notes. Notably, negation leads to sharper degradation in real notes, likely due to the complexity of natural language context. Synonym and numerical changes have smaller effects, with both note types showing similar trends. These results suggest that our framework generalizes across data sources and remains effective even on longer, noisier clinical text.





**Figure 5: Perturbations induce structured shifts in GPT-4O-mini’s latent space.** This figure shows how variance is distributed across the top 5 PCA dimensions under each perturbation type. Masked entities cause variance to concentrate sharply along a few axes, indicating semantic bottlenecks, whereas numerical perturbations yield flatter, more distributed profiles. Such variance patterns reveal how fragility aligns with low-rank latent distortions.

**Table 4: Clinical expert evaluation of synthetic notes.** Scores range from 0–3, where higher values indicate better note quality and diagnostic reasoning. Variability between reviewers highlights subjectivity in evaluating generated clinical text.

Note ID	Note Quality (R1)	Note Quality (R2)	Reasoning (R1)	Reasoning (R2)
Note_0	2	3	0	2
Note_1	2	3	2	3
Note_2	1	3	2	3
Note_3	3	3	3	3
Note_4	2	2	3	2

## 6.6 Human Expert Validation

Two clinicians scored five synthetic notes (0–3 scale). Mean scores were 2.0–2.7 across structure and diagnostic reasoning. Feedback highlighted missing vitals, incomplete differentials, and mismatches (e.g., hematemesis without GERD context). Table 4 summarizes scores. These qualitative insights affirm the need for latent audits even when surface fluency appears high.

## 6.7 Diagnostic Auditing Potential

LDFR identifies inputs where model outputs are unstable under light perturbations, flagging cases for downstream trust calibration. Unlike ROUGE or entity overlap, LDFR reveals semantic volatility invisible to surface metrics, offering a potential tool for auditing clinical assistant reliability. Taken together, our results show that LDFR captures subtle yet consequential breakdowns in diagnostic stability, providing a promising direction for future auditing tools in clinical AI systems.



## 7 Conclusion

We present a framework to uncover latent diagnostic instability in clinical language models by applying structured perturbations to synthetic clinical notes and analyzing semantic boundary shifts in latent space. Under entity masking, surface similarity remains high across thresholds, yet diagnosis flips become increasingly frequent highlighting fragility not captured by metrics like BERTScore. LDFR drops from 91.3% to 55% as perturbations increase, showing that the latent classifier no longer follows the LLM’s diagnosis reliably. This suggests that small changes in input can cause significant shifts in latent space, exposing both semantic drift and structural fragility. PCA-based boundary flips and per-axis variance shifts reveal latent vulnerabilities in diagnostic reasoning. Of all tested perturbations, entity masking proved most effective at exposing instability, inducing sharp transitions in low-dimensional latent space and misalignments that surface metrics fail to detect. Together, our findings underscore the need for geometry-aware evaluation to ensure safe, interpretable deployment of clinical LLMs in real-world settings. Our experiments with real notes from Di-ReCT and medical-specific models like MedGemma confirm that LDFR-based fragility patterns persist across synthetic and real clinical notes. Together, these findings support LDFR as a diagnostic tool for internal reasoning instability and a complement to traditional surface metrics in clinical LLM evaluation. Our central contribution, LDFR, offers a geometry-aware diagnostic signal that is not captured by surface-level metrics. Its consistent degradation under structured perturbations across both synthetic and real clinical notes highlights its utility for model auditing and evaluation beyond controlled benchmarks.

## 8 Limitations and Future Directions

Our framework offers a geometric perspective on diagnostic robustness but has limitations. Synthetic notes, while controllable, may not reflect real-world clinical variability, and using a fixed BERT embedding space may misalign with model-specific representations. While we include limited expert evaluation for unperturbed synthetic notes, we did not assess whether clinician reviewers agree with the diagnosis flips induced by perturbations. Future work should involve expert adjudication of perturbed samples to validate whether semantic edits, such as masking or negation, truly warrant diagnostic change. We selected a 90% PCA variance threshold based on the elbow plot (Appendix A.9), which showed that most diagnostic signal is captured within 30–45 components. This dimensionality preserves structure while ensuring that the LDFR classifier remains stable and interpretable across both synthetic and real note embeddings.

Future work will explore non-linear boundaries via manifold learning, expand perturbation analysis, and align latent shifts with expert judgments. Exploration of diagnostic manifolds as a foundation for clinical robustness is another direction illuminated.

## References

- [1] Yu-Wen Chen and Julia Hirschberg. 2024. Exploring Robustness in Doctor-Patient Conversation Summarization: An Analysis of Out-of-Domain SOAP Notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Danielle Bitterman (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1–9. doi:10.18653/v1/2024.clinicalnlp-1.1
- [2] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Joint Summits on Translational Science Proceedings* 2016 (2016), 41–50. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001761/> eCollection 2016.
- [3] Sherry J. H. Feng, Edmund M-K. Lai, and Weihua Li. 2024. Geometry of Textual Data Augmentation: Insights from Large Language Models. *Electronics* 13, 18 (2024). doi:10.3390/electronics13183781
- [4] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. arXiv:2212.07919 [cs.CL] <https://arxiv.org/abs/2212.07919>
- [5] Gokul Gowri, Xiao-Kang Lun, Allon M. Klein, and Peng Yin. 2024. Approximating mutual information of high-dimensional variables using learned representations. arXiv:2409.02732 [q-bio.QM] <https://arxiv.org/abs/2409.02732>
- [6] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv:2009.13081 [cs.CL] <https://arxiv.org/abs/2009.13081>
- [7] Robert Johnson, Uri Gottlieb, Gal Shaham, Liron Eisen, Jonathan Waxman, Shay Devons-Sherro, Charles R. Ginder, Peter Hong, Razi Sayeed, Barbra Y. Reis, Ran D. Balicer, Noa Dagan, and Marinka Zitnik. 2024. Unified Clinical Vocabulary Embeddings for Advancing Precision Medicine. *medRxiv* (10 December 2024). doi:10.1101/2024.12.03.24318322 Preprint.
- [8] Linyang Li, Ke Ren, Yunfan Shao, Pengyu Wang, and Xipeng Qiu. 2023. PerturbScore: Connecting Discrete and Continuous Perturbations in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6638–6648. doi:10.18653/v1/2023.findings-emnlp.442
- [9] Xin Li and Anand Sarwate. 2025. Unraveling the Localized Latents: Learning Stratified Manifold Structures in LLM Embedding Space with Sparse Mixture-of-Experts. arXiv:2502.13577 [cs.LG] <https://arxiv.org/abs/2502.13577>
- [10] Yizhan Li, Sifan Wu, Christopher W. Smith, Thomas Lo, and Bang Liu. 2024. Improving Clinical Note Generation from Complex Doctor-Patient Conversation. *ArXiv abs/2408.14568* (2024). <https://api.semanticscholar.org/CorpusID:271962802>
- [11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://aclanthology.org/W04-1013/>
- [12] Yinlong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulic, and Nigel Collier. 2025. Measuring, Evaluating and Improving Logical Consistency in Large Language Models. <https://openreview.net/forum?id=kjgi5ykK3t>
- [13] Isabel Lopez, Arjun Swaminathan, Karan Vedula, et al. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine* 8 (2025), 45. doi:10.1038/s41746-024-01377-1 Published: January 19, 2025.
- [14] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. arXiv:2005.05909 [cs.CL] <https://arxiv.org/abs/2005.05909>
- [15] Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering. arXiv:2406.06573 [cs.CL] <https://arxiv.org/abs/2406.06573>
- [16] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Justin Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming

- Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] <https://arxiv.org/abs/2303.08774>
- [17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135
- [18] Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2022. On the Usefulness of Embeddings, Clusters and Strings for Text Generator Evaluation. <https://api.semanticscholar.org/CorpusID:253735234>
- [19] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10066–10086. doi:10.18653/v1/2023.emnlp-main.622
- [20] Clément Péalat, Guillaume Bouleux, Vincent Cheutet, Maxime Maignan, Luc Provoost, Sylvie Pillet, and Olivier Mory. 2023. Using a manifold-based approach to extract clinical codes associated with winter respiratory viruses at an emergency department. *Expert Systems with Applications* 230 (2023), 120620. doi:10.1016/j.eswa.2023.120620
- [21] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4902–4912. doi:10.18653/v1/2020.acl-main.442
- [22] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (August 2023), 172–180. doi:10.1038/s41586-023-06291-2
- [23] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. arXiv:2205.09148 [cs.CL] <https://arxiv.org/abs/2205.09148>
- [24] Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. DiReCT: Diagnostic Reasoning for Clinical Notes via Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=F7rAX6yiS2>
- [25] Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, Heng Ji, Jong C. Park, and Rui Xia (Eds.). Association for Computational Linguistics, Online, 347–355. doi:10.18653/v1/2021.acl-demo.41
- [26] Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. 2025. Latent Space Chain-of-Embedding Enables Output-free LLM Self-Evaluation. arXiv:2410.13640 [cs.CL] <https://arxiv.org/abs/2410.13640>
- [27] Shang Xian, Maria E. Grabowska, Iftikhar J. Kullo, Yuan Luo, Jordan W. Smoller, Wei-Qi Wei, Gail Jarvik, Sean Mooney, and David Crosslin. 2024. Language-model-based patient embedding using electronic health records facilitates phenotyping, disease forecasting, and progression analysis. *Research Square* (23 September 2024). doi:10.21203/rs.3.rs-4708839/v1 Preprint.
- [28] Zi Yin and Yuanyuan Shen. 2018. On the Dimensionality of Word Embedding. arXiv:1812.04224 [cs.LG] <https://arxiv.org/abs/1812.04224>
- [29] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] <https://arxiv.org/abs/1904.09675>
- [30] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. 2024. Agent-as-a-Judge: Evaluate Agents with Agents. arXiv:2410.10934 [cs.AI] <https://arxiv.org/abs/2410.10934>

## A Appendix

### A.1 Code and reproduction

All prompting templates and code to reproduce our results are available at: [https://github.com/unni12345/geometric\\_diagnostics](https://github.com/unni12345/geometric_diagnostics)

### A.2 The Prompts

This prompt illustrated in Fig.6 is designed to transform a structured patient–physician dialogue into a synthetic clinical note. The input includes socio-demographic details and question–answer pairs labelled by symptom or antecedent type. The prompt provides a well-defined clinical template with standard sections (e.g., Chief Complaint, History of Present Illness, Physical Exam), guiding the LLM to produce notes that are realistic, interpretable, and diagnostic-ready. This structure ensures that downstream perturbations can be applied in a controlled manner without compromising the core semantic structure.

#### A.2.1 Clinical Note Prompt.

**A.2.2 Forward Reasoning Prompt.** This prompt illustrated in Fig.7 initiates forward diagnostic reasoning: from clinical note to inferred diagnosis. The model is instructed to extract relevant observations and formulate logical deductions in a step-by-step manner. This chain-of-thought generation encourages explicit reasoning and provides transparency into the model’s inference process. It is a crucial step for ensuring diagnostic traceability in the synthetic pipeline.

### A.3 Backward Reasoning Prompt

This prompt illustrated in Fig.8 facilitates backward reasoning: from an already inferred diagnosis to the supporting clinical evidence. The model must align symptoms, findings, and history in a way that supports the diagnosis logically. This step helps validate the consistency of the model’s decision-making and serves as a diagnostic sanity check, reinforcing causal alignment between data and prediction.

**A.3.1 Aggregator Prompt.** This prompt is illustrated in Fig.9 compares the forward and backward reasoning chains and compels

---

Patient Details:  
Age: 44  
Sex: F

Initial evidence provided: [['question': 'Do you have pain  
↪ somewhere, related to your reason for consulting?',  
↪ 'is\_antecedent': False, 'answer': True]]

Question & Answer Section:  
For the question "Do you have pain somewhere, related to your  
↪ reason for consulting?",  
the patient replied 'True', which is categorized as a symptom.  
For the question "Characterize your pain:", the patient  
↪ replied 'sickening',  
which is categorized as a symptom.  
For the question "Do you feel pain somewhere?", the patient  
↪ provided 'upper chest',  
indicating a symptom.

Clinical Report Template:  
Based on the information provided, generate a comprehensive  
↪ clinical report with the  
following sections:  
- Patient Details  
- Chief Complaint (extracted from the initial evidence)  
- History of Present Illness: Describe onset, duration,  
↪ severity, and relevant history.  
- Past Medical History: Summarize any significant conditions.  
- Medications and Allergies: List current medications and any  
↪ known allergies.  
- Physical Examination: Include vitals, general appearance,  
↪ and pertinent exam findings.

---

**Figure 6: Prompt for Clinical Note Generation.**

---

You are a diagnostic reasoning assistant. Read the following  
↪ clinical note and generate  
a step-by-step reasoning process that extracts key  
↪ observations and logical deductions  
leading to a diagnosis.

Clinical Note:  
<-----Clinical Note----->

Provide the list of observations and deductions. Be concise.

---

**Figure 7: Prompt for Forward Reasoning Generation.**

---

You are a diagnostic reasoning assistant. Given the following  
↪ clinical note and  
the final diagnosis '<diagnosis\_placeholder>', generate a  
↪ backward reasoning chain  
explaining how each observation supports the diagnosis.

Clinical Note:  
<-----Clinical Note----->

Final Diagnosis: <diagnosis\_placeholder>

Provide a concise list of observations and logical deductions.

---

**Figure 8: Prompt for Backward Reasoning Generation.**

the model to produce a final, consolidated rationale. The goal is to reconcile both perspectives into a coherent diagnostic explanation, reducing contradictions and highlighting mutually reinforcing observations. This step ensures logical integrity in the final output

and simulates a clinician’s process of reviewing both presentation and diagnostic hypothesis.

#### A.4 Note-to-Dialogue Comparison via Surface-Level Metrics

For completeness, we report automatic evaluation scores comparing the generated clinical notes to their source DDXPlus dialogues using standard surface-level metrics. These include BERTScore, ROUGE-L, and biomedical NER-based Jaccard and F1 scores, summarized in Table 5.

The absolute values of these metrics are relatively low (e.g., BERTScore ranges between 0.54–0.58), which is expected given the nature of the comparison: the source DDXPlus inputs are unstructured, sparse, and conversational, while the generated notes are formal and structured. As a result, high token or span level overlap is neither expected nor necessarily desirable.

We used bert-large-uncased for BERTScore and the d4data/biomedical-ner-all model for entity-level scoring. Both are pre-trained general-purpose tools and not specifically optimized for dialogue-to-note mapping. Furthermore, BERTScore captures token-level similarity and does not reflect higher-level clinical discourse structure or reasoning coherence.

These scores are therefore included for completeness and relative comparison across models, but they do not directly evaluate the quality of diagnostic reasoning or factual consistency. In fact, high surface-level similarity could indicate shallow copying, whereas meaningful abstraction or reasoning may naturally result in lower overlap—motivating the need for robustness evaluation beyond traditional text metrics.

**Table 5: Surface-level comparison between generated clinical notes and their source DDXPlus dialogues. Metrics include BERTScore, ROUGE-L, and biomedical NER-based Jaccard and F1 scores. The best-performing model per metric is bolded. While Open-Mistral-7B ranks highest on BERTScore and ROUGE-L, GPT-3.5-turbo yields the best entity-level scores—indicating that surface similarity and clinical entity preservation do not always align.**

Model	BERTScore	ROUGE-L	NER Jaccard	NER F1
GPT-3.5-turbo-0125	0.546	0.221	<b>0.171</b>	<b>0.283</b>
GPT-4o-mini-2024-07-18	0.559	0.219	0.141	0.243
LLaMA-3.1-8b-instant	0.575	0.240	0.161	0.270
Open-Mistral-7B	<b>0.580</b>	<b>0.244</b>	0.164	0.276

#### A.5 Supplementary Metric Plots

We have included full degradation plots for all NLP evaluation metrics—NER F1, ROUGE-L, and NER Jaccard—under each perturbation type. These figures provide a comprehensive view of how different perturbations affect semantic fidelity and entity preservation across clinical LLMs.

A.5.1 NER F1 vs. Threshold.

A.5.2 ROUGE-L vs. Threshold.

A.5.3 NER Jaccard vs. Threshold.

You are a diagnostic reasoning aggregator. Compare the following two reasoning chains and generate the final, consolidated rationale behind the diagnosis:

Forward Reasoning:

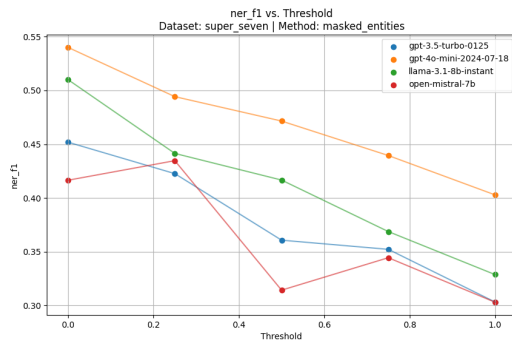
<-----Forward Reasoning Chain----->

Backward Reasoning:

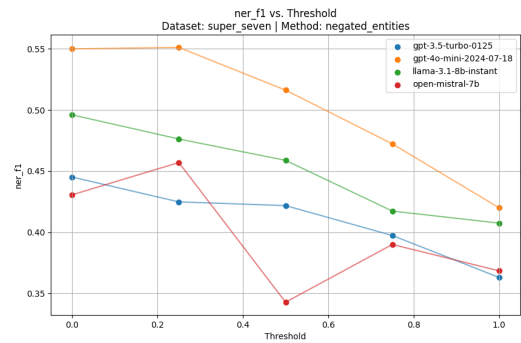
<-----Backward Reasoning Chain----->

Ensure that both chains are consistent in supporting the given diagnosis. Provide a concise list of key observations and deductions.

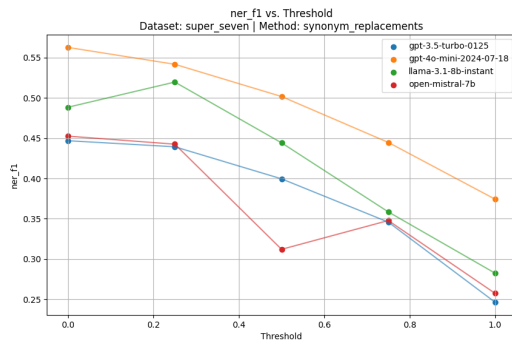
Figure 9: Prompt for Aggregation and Verification of Reasoning Chains.



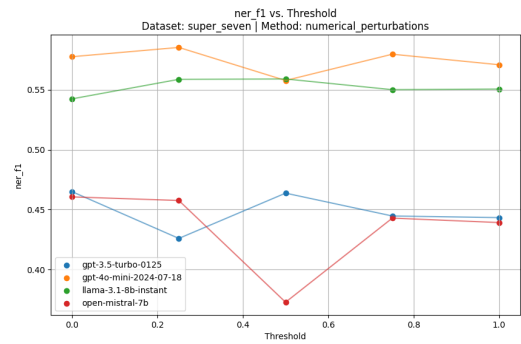
(a) Masked Entities



(b) Negated Entities



(c) Synonym Replacements



(d) Numerical Perturbations

Figure 10: Entity-level recognition performance degrades under perturbation, particularly with masking and synonym edits. This figure shows how NER F1 scores drop as perturbation intensity increases across four types. Masked and synonym-modified inputs reduce clinical entity recall most significantly.

## A.6 Hyperparameters for Geometric Analysis

### A.6.1 Embedding Extraction.

- **Embedding Model:** ClinicalBERT (pretrained, frozen)
- **Tokenization:** BertTokenizer (max length = 512 tokens)
- **Embedding Representation:** Mean pooling over last hidden layer
- **Embedding Dimension:** 768

### A.6.2 PCA Reduction.

- **Input:** Embeddings of unperturbed clinical notes (N = 100)

- **Normalization:** Zero-centered with unit variance per dimension
- **Components Retained:** 90% explained variance (typically 30-40 components)
- **Library Used:** scikit-learn PCA (randomized SVD solver)

### A.6.3 Latent-Space Classifier.

- **Classifier:** Logistic Regression (1-vs-rest)
- **Training Set:** PCA-reduced embeddings of original (unperturbed) notes
- **Labels:** Ground truth diagnoses (7-class classification)

- **Regularization:** L2, inverse regularization strength  $C = 10.0$
- **Solver:** lgfb
- **Max Iteration:** 1000
- **Cross-validation:** 10-fold

## A.7 Illustrated Example

The following example illustrates the flow of clinical data for a case with GERD as diagnosis. We illustrate the case of Masked Perturbation at 100%. The following entities will be masked during the perturbation.

We illustrate the Clinical Note with Fig 13, Forward Diagnosis for this case with Fig 14, backward diagnosis with Fig 15 and final diagnostic reasoning with Fig 16.

The masked entities perturbation case is illustrated with Fig 17, forward reasoning with Fig 18, backward reasoning with Fig 19, final reasoning with Fig 20.

**chronic sinusitis, persistent cough, family history of asthma, 26-year-old male, hospitalizations, respiratory distress, asthma attacks, bronchodilator, wheezing, common allergies**

---

"Patient Details:  
Age: 57  
Sex: Female

Chief Complaint:  
The patient reports experiencing symptoms that worsen after  
↪ eating.

History of Present Illness:  
The patient describes her pain as haunting, sensitive, a knife  
↪ stroke, sickening, and scary. The pain is located in the  
↪ lower chest, upper chest, hypochondrium (left), and  
↪ epigastric regions. The intensity of the pain is unknown,  
↪ and it radiates to the upper chest. The pain's onset and  
↪ duration are unknown, but it is alleviated while sitting  
↪ up. The patient also reports a burning sensation that  
↪ starts in the stomach and goes up into the throat,  
↪ associated with a bitter taste in the mouth.

Past Medical History:  
The patient has a hiatal hernia and asthma, for which she has  
↪ used a bronchodilator in the past. She is currently  
↪ pregnant.

Medications and Allergies:  
No current medications were reported, but the patient has a  
↪ history of using a bronchodilator for asthma. She has no  
↪ known allergies.

Physical Examination:  
Vitals: Not provided in the data  
General Appearance: Not provided in the data  
Pertinent Exam Findings: The patient reports pain in various  
↪ chest regions, and a burning sensation with a bitter taste  
↪ in the mouth. The pain is alleviated while sitting up. No  
↪ further physical examination findings were provided in the  
↪ data.

Additional Notes:  
The patient also reports being significantly overweight  
↪ compared to people of the same height, drinking alcohol  
↪ excessively, and smoking cigarettes. She has not traveled  
↪ out of the country in the last 4 weeks. These factors may  
↪ be relevant to the patient's current symptoms and should  
↪ be considered in the diagnostic and management process."

---

Figure 13: Clinical Note for diagnosis GERD

---

"1. The patient is a 57-year-old female.  
2. She reports symptoms worsening after eating.  
3. The pain is described as haunting, sensitive, a knife  
↪ stroke, sickening, and scary.  
4. The pain is located in the lower chest, upper chest,  
↪ hypochondrium (left), and epigastric regions.  
5. The pain radiates to the upper chest.  
6. The intensity and duration of the pain are unknown.  
7. The pain is alleviated while sitting up.  
8. A burning sensation is reported, starting in the stomach  
↪ and going up into the throat.  
9. The patient has a bitter taste in the mouth.  
10. She has a history of hiatal hernia and asthma.  
11. She has used a bronchodilator for asthma in the past.  
12. She is currently pregnant.  
13. No current medications were reported.  
14. She has no known allergies.  
15. The patient reports being significantly overweight.  
16. She drinks alcohol excessively.  
17. She smokes cigarettes.  
18. No further physical examination findings were provided.  
19. The patient has not traveled out of the country in the  
↪ last 4 weeks.  
20. These factors (overweight, alcohol consumption, smoking)  
↪ may be relevant to the patient's current symptoms."

---

Figure 14: Forward reasoning for GERD

---

"1. Patient reports symptoms worsening after eating.  
↪ (Observation)  
2. Patient describes pain as haunting, sensitive, and  
↪ radiating to the upper chest. (Observation)  
3. Patient reports a burning sensation from stomach to throat  
↪ with a bitter taste. (Observation)  
4. Patient has a hiatal hernia. (Past Medical History)  
5. Symptoms of GERD (gastroesophageal reflux disease) include  
↪ heartburn, regurgitation, and pain that worsens after  
↪ eating. (Knowledge of GERD)  
6. GERD symptoms can be exacerbated by hiatal hernia.  
↪ (Knowledge of GERD and hiatal hernia)  
7. The patient's symptoms, including pain, burning sensation,  
↪ and bitter taste, are consistent with GERD. (Logical  
↪ deduction)  
8. The patient's history of hiatal hernia, along with her  
↪ symptoms, increases the likelihood of GERD. (Logical  
↪ deduction)  
9. The patient's symptoms are not typical of other conditions,  
↪ such as heart disease or gallbladder disease, given her  
↪ age and lack of other risk factors. (Logical deduction)  
10. Therefore, the most likely diagnosis is GERD. (Final  
↪ deduction)"

---

Figure 15: Backward reasoning for GERD

## A.8 Per-Dimension Variance Plots and Interpretation Synthetic Notes

To support our analysis in Section 6.4, we present detailed variance decomposition plots for each model under all perturbation types. Each figure shows the top 5 latent dimensions with the highest explained variance across perturbation thresholds (0%, 25%, 50%, 75%, 100%). Illustrated through Fig 22, Fig 21 and Fig 23.

### Key Insights

- Dimensions such as 273 and 752 consistently exhibit high sensitivity to masked and negated entity perturbations.

"The forward and backward reasoning chains are consistent in  
 ↳ supporting the diagnosis of GERD (gastroesophageal reflux  
 ↳ disease).

Observations:

1. Symptoms worsen after eating.
2. Pain described as haunting, sensitive, and radiating to the  
 ↳ upper chest.
3. Burning sensation from stomach to throat with a bitter  
 ↳ taste.

Logical Deductions:

1. Patient's symptoms are consistent with GERD.
2. Patient's history of hiatal hernia increases the likelihood  
 ↳ of GERD.
3. Patient's symptoms are not typical of other conditions,  
 ↳ such as heart disease or gallbladder disease.
4. The patient's lifestyle factors (overweight, alcohol  
 ↳ consumption, smoking) may be relevant to her symptoms.
5. The patient's symptoms are alleviated while sitting up,  
 ↳ which is a common symptom of GERD.
6. The patient's use of a bronchodilator for asthma does not  
 ↳ directly contribute to the diagnosis of GERD, but it does  
 ↳ not rule it out either.
7. The patient's current pregnancy may not directly affect the  
 ↳ diagnosis of"

**Figure 16: Aggregated reasoning for GERD**

"Patient Details:

Age: 57  
 Sex: [MASK]

Chief Complaint:

The patient reports experiencing [MASK] that worsen after  
 ↳ eating.

History of Present Illness:

The patient describes her [MASK] as haunting, sensitive, a  
 ↳ knife stroke, sickening, and scary. The [MASK] is located  
 ↳ in the [MASK], [MASK], hypochondrium (left), and [MASK].  
 ↳ The intensity of the [MASK] is unknown, and it radiates to  
 ↳ the [MASK]. The [MASK]'s onset and duration are unknown,  
 ↳ but it is alleviated while sitting up. The patient also  
 ↳ reports a [MASK] that starts in the [MASK] and goes up  
 ↳ into the [MASK], associated with a [MASK] in the mouth.

Past Medical History:

The patient has a [MASK] and [MASK], for which she has used a  
 ↳ [MASK] in the past. She is currently [MASK].

Medications and Allergies:

No current medications were reported, but the patient has a  
 ↳ history of using a [MASK] for [MASK]. She has no known  
 ↳ allergies.

Physical Examination:

Vitals: Not provided in the data

General Appearance: Not provided in the data

Pertinent Exam Findings: The patient reports [MASK] in various  
 ↳ chest regions, and a [MASK] with a [MASK] in the mouth.  
 ↳ The [MASK] is alleviated while sitting up. No further  
 ↳ physical examination findings were provided in the data.

Additional Notes:

The patient also reports being significantly [MASK] compared  
 ↳ to people of the same height, drinking [MASK] excessively,  
 ↳ and [MASK]. She has not traveled out of the country in the  
 ↳ last 4 weeks. These factors may be relevant to the  
 ↳ patient's current [MASK] and should be considered in the  
 ↳ diagnostic and management process."

**Figure 17: Masked Clinical Note for diagnosis GERD**

- "1. Patient is a 57-year-old [MASK] with a chief complaint of  
 ↳ unexplained discomfort worsening after eating.
2. The discomfort is described as haunting, sensitive, a knife  
 ↳ stroke, sickening, and scary.
3. The discomfort is located in the epigastric region, left  
 ↳ hypochondrium, and umbilical region.
4. The intensity and duration of the discomfort are unknown,  
 ↳ but it radiates to the back.
5. The discomfort is alleviated while sitting up.
6. The patient also reports a burning sensation starting in  
 ↳ the throat and going up into the mouth, associated with a  
 ↳ sensation in the mouth.
7. The patient has a history of hypertension and diabetes, and  
 ↳ has used a diuretic in the past.
8. No current medications were reported, but the patient has a  
 ↳ history of using a diuretic for hypertension.
9. The patient has no known allergies.
10. The patient reports feeling significantly underweight,  
 ↳ drinking alcohol excessively, and having a poor appetite.
11. No further physical examination findings were provided.
12. The patient has not traveled out of the country in the  
 ↳ last 4 weeks.
13. These factors may be relevant to the patient's current  
 ↳ symptoms and should be considered in the diagnostic and  
 ↳ management process."

**Figure 18: Forward reasoning for Masked Perturbation of 100% GERD**

- "1. The patient reports symptoms of heartburn that worsen  
 ↳ after eating.
2. The heartburn is described as haunting, sensitive, and  
 ↳ radiating to the left hypochondrium and back.
3. The patient has a history of GERD and has used a proton  
 ↳ pump inhibitor (PPI) in the past.
4. The patient's symptoms are alleviated while sitting up,  
 ↳ which is a common symptom of GERD.
5. The patient's excessive alcohol consumption and obesity may  
 ↳ contribute to the worsening of GERD symptoms.
6. No physical examination findings suggest an alternative  
 ↳ diagnosis, such as a cardiac issue.

Logical deductions:

1. The patient's symptoms of heartburn, worsened by eating and  
 ↳ alleviated by sitting up, are consistent with GERD.
2. The patient's history of GERD and use of a PPI support the  
 ↳ diagnosis of GERD.
3. The patient's obesity and excessive alcohol consumption may  
 ↳ exacerbate GERD symptoms.
4. The absence of physical examination findings suggesting an  
 ↳ alternative diagnosis supports the diagnosis of GERD."

**Figure 19: Backward reasoning for Masked Perturbation of 100% GERD**

- GPT-3.5 and LLaMA display stronger activation along single dimensions, while GPT-4o exhibits more distributed variance across axes.
- Numerical perturbations exhibit the flattest distribution, indicating resilience in quantitative embeddings.

**A.8.1 Composite Visualization.** To reduce visual clutter, we aggregate per-model variance breakdowns into grouped figures by perturbation type. Full-resolution figures per model are available upon request.

---

"The forward and backward reasoning chains are consistent in  
 ↳ supporting the diagnosis of GERD.

Observations:

1. Patient's discomfort worsens after eating and is alleviated  
 ↳ while sitting up.
2. Discomfort is described as haunting, sensitive, and  
 ↳ radiates to the left hypochondrium and back.
3. Patient has a history of hypertension, diabetes, and past  
 ↳ use of a diuretic.
4. Patient reports excessive alcohol consumption, poor  
 ↳ appetite, and feeling significantly underweight.

Logical deductions:

1. Symptoms are consistent with GERD.
2. Patient's history of GERD and use of a PPI support the  
 ↳ diagnosis.
3. Patient's lifestyle factors may exacerbate GERD symptoms.
4. Absence of physical examination findings suggesting an  
 ↳ alternative diagnosis supports the diagnosis of GERD."

---

**Figure 20: Final diagnostic reasoning for Masked Perturbation of 100% GERD**

## A.9 Elbow Plot of Latent Embedding Dimensionality

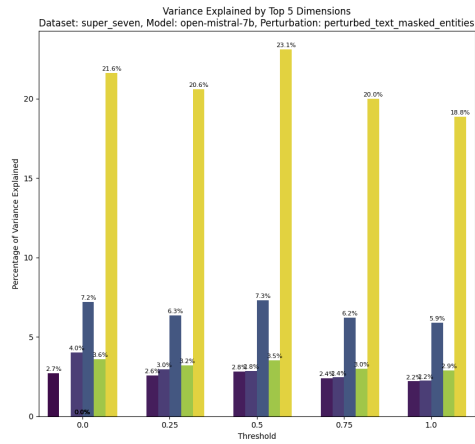
To select an appropriate dimensionality for LDFR evaluation, we performed Principal Component Analysis (PCA) on note embeddings from both synthetic and real clinical notes. The goal was to retain sufficient variance for meaningful diagnostic structure while avoiding overfitting or noise amplification in the latent classifier.

Figure 24 shows the cumulative explained variance as a function of the number of principal components.

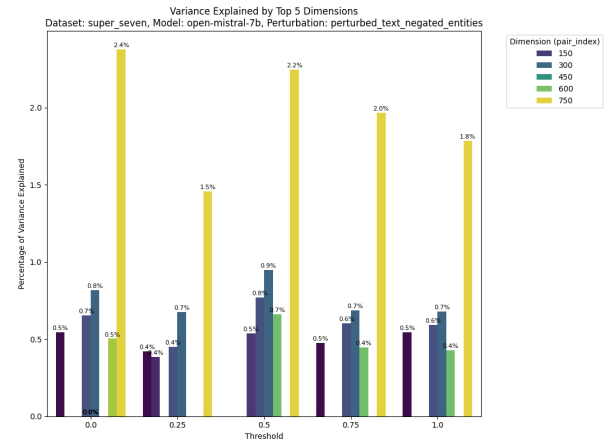
For synthetic notes, 90% of variance is typically captured within 30–35 components. In contrast, real clinical notes require approximately 45 components to reach the same threshold, consistent with their longer length and richer entity density.

We therefore fixed the PCA threshold at 90% explained variance throughout our experiments. This ensures that the latent classifier operates in a stable, low-dimensional subspace that is both interpretable and aligned with clinical signal, as supported by the elbow point in both cases.

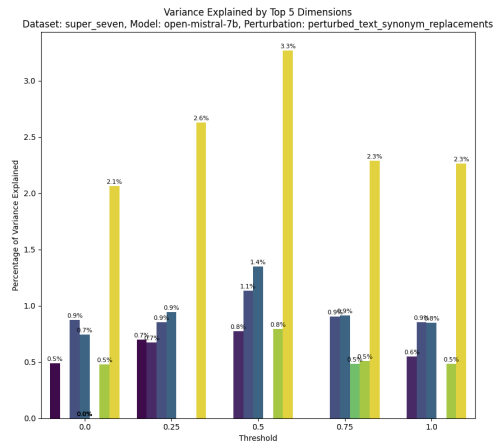




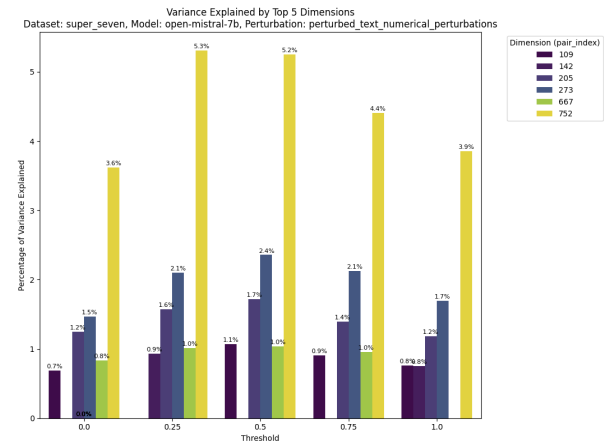
(a) Masked Entities



(b) Negated Entities



(c) Synonym Replacements



(d) Numerical Perturbations

Figure 21: Variance explained by top 5 latent dimensions for open-mistral-7b across perturbation types.

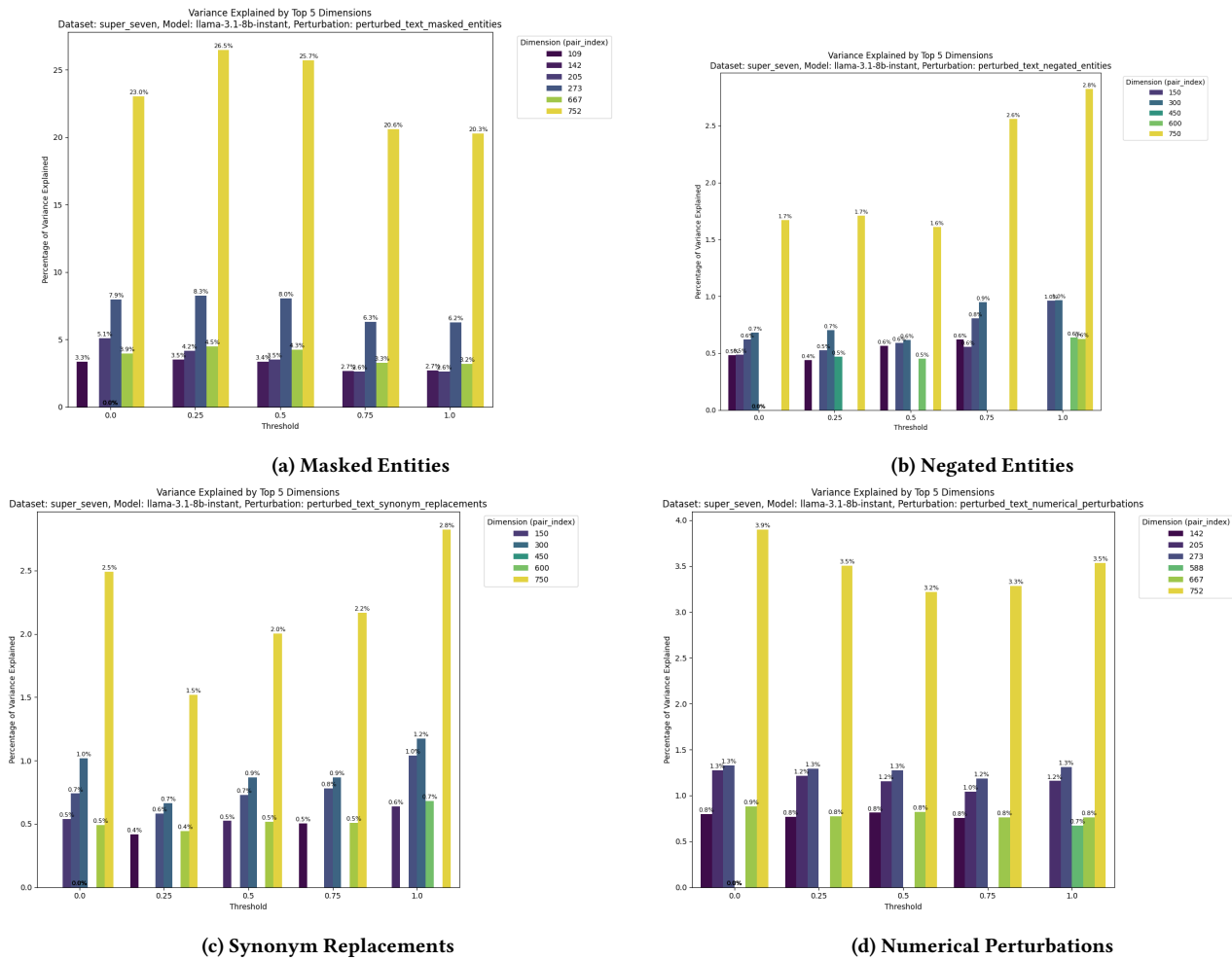


Figure 22: Variance explained by top 5 latent dimensions for llama318binstant across perturbation types.

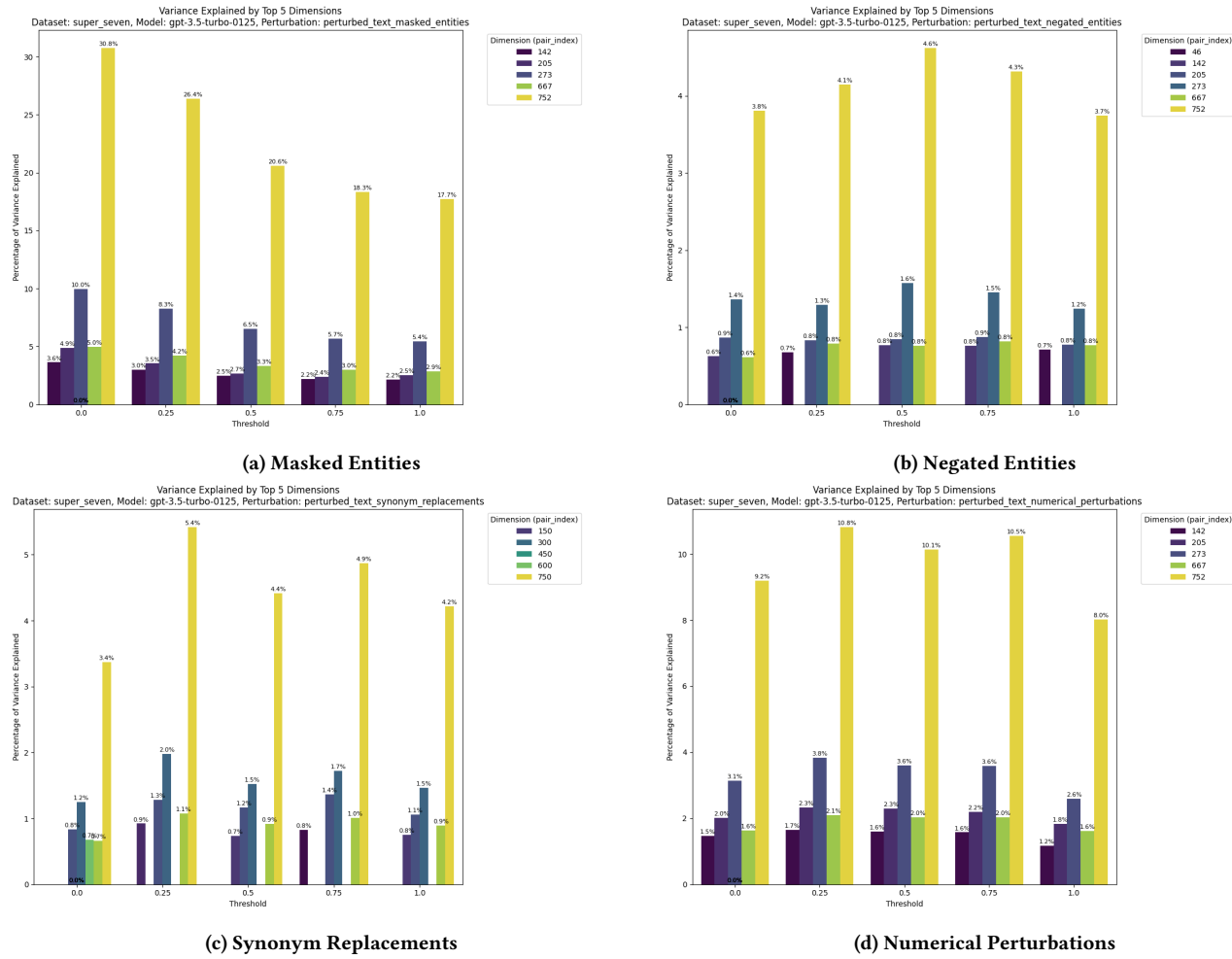


Figure 23: Variance explained by top 5 latent dimensions for gpt3\_5turbo0125 across perturbation types.

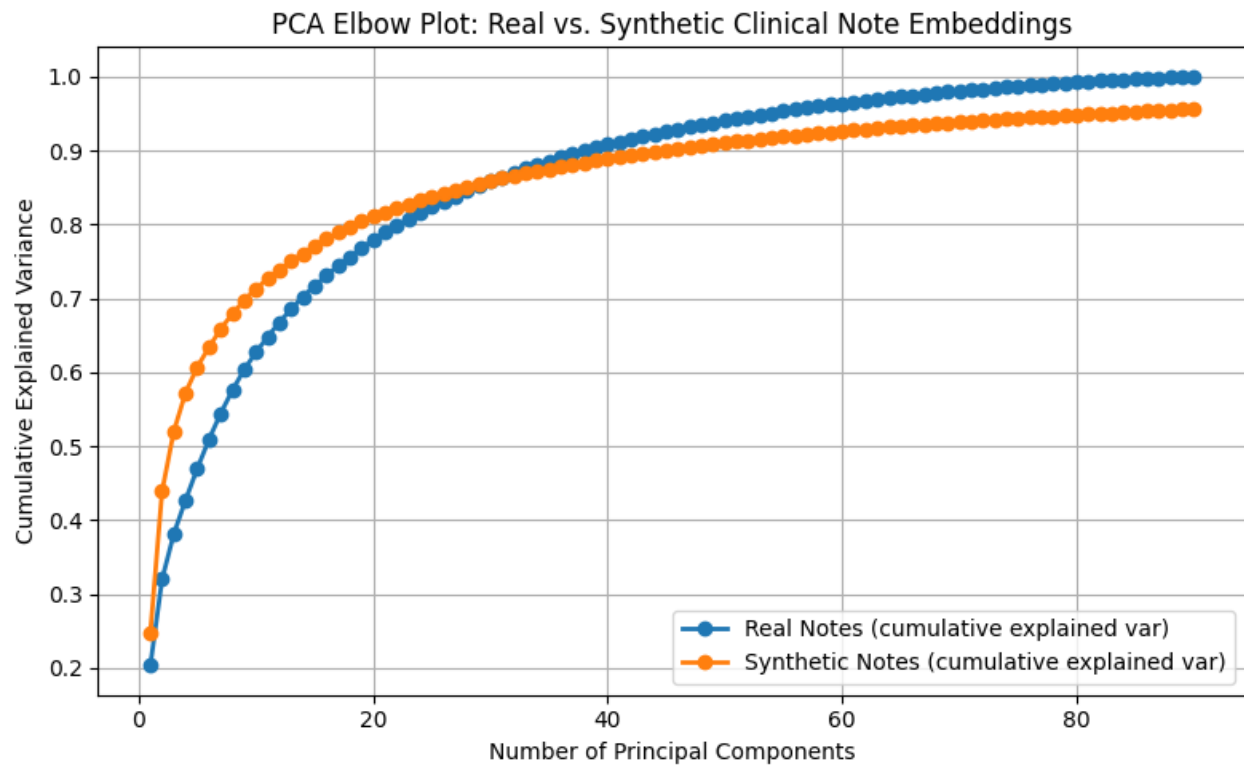


Figure 24: PCA Elbow Plot comparing cumulative explained variance for real (DiReCT) and synthetic clinical note embeddings. Real notes require more dimensions to capture 90% of variance, reflecting their greater complexity and redundancy.