

Estimating Hallucination via Divergence Functionals and Posterior Risk

Susmita Ganguly¹, Kasturi Barik², Amit Chakraborty³

¹JIS Institute of Advanced Studies & Research, Kolkata, India.

ORCID: 0009-0004-4252-6075, Email: ganguly.susmita2018@gmail.com

²JIS Institute of Advanced Studies & Research, Kolkata, India.

ORCID: 0000-0001-6104-0899, Email: kasturi.barik@jisiasr.org

³JIS Institute of Advanced Studies & Research, Kolkata, India.

ORCID: 0009-0002-6459-0612, Email: amit13.ons@gmail.com

ABSTRACT

Hallucination in large language models constitutes a latent statistical phenomenon arising from discrepancies between predictive probability measures and the underlying data-generating process. Existing evaluation methodologies predominantly rely on deterministic factuality scores or benchmark-specific accuracy measures, offering limited inferential interpretation and weak statistical generalizability. This paper introduces a unified inferential framework for estimating hallucination through divergence functionals and posterior risk. Hallucination is formulated as an unobserved random functional defined over sequential predictive distributions, enabling its estimation via discrepancies between posterior predictive measures and externally validated reference distributions. The proposed methodology integrates families of statistical divergence functionals, including Kullback–Leibler, Jensen–Shannon, Rényi, and Wasserstein divergences, with Bayesian posterior risk to construct a consistent estimator of hallucination intensity under heterogeneous uncertainty regimes. Asymptotic properties of the estimator, including consistency, asymptotic normality, and convergence under increasing sample complexity, are established within a semiparametric inference framework. Statistical uncertainty is quantified using influence-function approximations, bootstrap confidence intervals, and posterior credible regions, facilitating principled hypothesis testing of hallucination prevalence across model architectures and prompting strategies. An empirical instantiation on clinical dialogue summarization compares three candidate divergence-style labels and finds that a Wasserstein-2 divergence label yields a substantially more learnable estimation target than a lexical word-overlap baseline (test-set R^2 of 0.972 versus 0.513), while a reconstruction-based novelty label did not yield a usable signal under the configuration tested. The framework establishes a theoretically grounded statistical paradigm for reproducible hallucination estimation and comparative reliability assessment in generative artificial intelligence.

KEYWORDS: *Hallucination Estimation, Divergence Functionals, Wasserstein Distance, Bayesian Posterior Risk, Large Language Models, Clinical Note Summarization, Faithfulness Evaluation*

INTRODUCTION

Hallucination in large language models (LLMs) represents a fundamental inferential challenge rather than merely an operational failure of text generation. Although contemporary evaluation protocols predominantly quantify hallucination using benchmark-specific accuracy, factual consistency scores, or heuristic confidence estimates, such measures seldom admit a rigorous probabilistic interpretation or support formal statistical inference. From a statistical perspective, hallucination is more naturally regarded as an unobservable latent functional arising from the discrepancy between the model-induced predictive probability measure and the underlying reference distribution governing factual knowledge. Consequently, the problem of hallucination estimation extends beyond deterministic classification and enters the broader domain of inference on latent stochastic processes, where uncertainty, distributional variability, and posterior predictive behaviour collectively determine the reliability of generated responses. Existing methodologies generally neglect the distributional geometry of predictive uncertainty, thereby limiting their ability to characterize hallucination under domain shift, heterogeneous prompting strategies, and evolving posterior beliefs. This limitation motivates the development of inferential procedures capable of estimating hallucination through principled statistical functionals rather than task-specific empirical metrics.

$$\hat{H} = \alpha D(P \parallel R) + (1 - \alpha) R\pi, \quad 0 \leq \alpha \leq 1, \dots\dots\dots (1)$$

In this study, hallucination is formulated as a latent random functional defined on sequential predictive distributions, enabling its estimation through divergence functionals and Bayesian posterior risk. The proposed framework employs families of probability metrics, including Kullback-Leibler divergence, Jensen-Shannon divergence, Rényi divergence, and Wasserstein distance, to quantify departures between posterior predictive distributions and externally validated reference measures while simultaneously incorporating posterior risk to account for epistemic uncertainty.

$$H(P, R) = D(P \parallel R), \dots\dots\dots (2)$$

This formulation yields a statistically coherent estimator possessing desirable inferential properties, including consistency, asymptotic normality, and robustness under distributional perturbations. Furthermore, the framework facilitates uncertainty quantification through bootstrap-based confidence intervals, posterior credible regions, and influence-function approximations, thereby permitting formal hypothesis testing and comparative analysis across model architectures, prompting regimes, and knowledge domains.

$$R\pi = E\theta|D [L(\theta, \theta)] = \int L(\theta, \theta) \pi(\theta | D) d\theta, \dots\dots\dots (3)$$

Unlike conventional evaluation methodologies that emphasize pointwise factual correctness, the proposed approach treats hallucination as an emergent property of predictive distributions, thereby establishing a direct connection between modern generative artificial intelligence and the theoretical foundations of statistical inference, information geometry, Bayesian decision theory, and semiparametric estimation. By embedding hallucination within a unified inferential framework based on divergence functionals and posterior risk, this work provides a theoretically grounded, statistically interpretable, and computationally scalable methodology for reproducible hallucination estimation, offering a foundation for future research on uncertainty-aware evaluation and reliability assessment in generative language models.

The practical stakes of this problem are highest in domains where generated text directly informs downstream decisions. In clinical documentation, for instance, a hallucinated clinical note -- one asserting a symptom, medication, or history element not actually present in the underlying patient-provider conversation -- can propagate into a permanent medical record and influence subsequent care decisions. In legal and financial reporting, hallucinated citations or figures carry analogous downstream risk. These high-stakes settings share a common structural feature that motivates the present framework: a verifiable reference context (a conversation transcript, a source document, a verified financial figure) typically exists alongside the generated text, making the divergence between generated and reference content a directly estimable quantity rather than a purely latent one. This observation underlies the empirical instantiation of the proposed framework in Section 6, which treats clinical dialogue summarization as a concrete, fully observed setting for evaluating divergence-based hallucination estimators before extending the framework to settings where no explicit reference context is available.

RELATED WORK

The statistical characterization of hallucination has received comparatively limited attention despite the rapid evolution of large language models and their widespread deployment across knowledge-intensive applications. Existing literature predominantly approaches hallucination from the perspectives of factual consistency evaluation, benchmark-based accuracy assessment, or confidence calibration, wherein hallucinated outputs are identified through deterministic comparisons against reference corpora or external knowledge bases. While these methodologies have demonstrated practical utility, they generally lack an explicit probabilistic interpretation and provide limited support for formal statistical inference. Consequently, hallucination is often treated as a binary classification problem rather than as a latent stochastic phenomenon whose occurrence is governed by uncertainty in sequential predictive distributions.

Recent advances in uncertainty quantification have incorporated Bayesian neural inference, Monte Carlo sampling, ensemble learning, and posterior calibration to estimate predictive confidence; however, these

methods primarily quantify model uncertainty rather than the statistical intensity of hallucination itself. Parallel developments in information theory have introduced divergence-based measures, including Kullback-Leibler divergence, Jensen-Shannon divergence, Rényi divergence, and Wasserstein distance, for comparing probability distributions arising from machine learning models. These measures have found extensive application in domain adaptation, distributional robustness, generative modeling, and statistical hypothesis testing, yet their use as inferential functionals for hallucination estimation remains largely unexplored.

Similarly, Bayesian decision theory provides a principled framework for incorporating posterior uncertainty through expected loss minimization and posterior risk, offering statistically coherent estimators with well-established asymptotic properties. Nevertheless, existing hallucination evaluation frameworks rarely integrate divergence functionals with posterior risk into a unified inferential methodology. From the standpoint of theoretical statistics, hallucination may instead be interpreted as an unobservable latent functional defined over predictive probability measures, thereby permitting estimation through semiparametric inference, empirical process theory, and functional analysis. Such a formulation naturally accommodates uncertainty quantification, asymptotic consistency, and robustness under distributional perturbations while avoiding dependence on benchmark-specific heuristics.

Despite significant progress in uncertainty estimation, calibration, and probabilistic language modeling, a comprehensive statistical framework capable of simultaneously modeling distributional discrepancy, posterior uncertainty, and inferential variability has yet to emerge. The present work addresses this gap by formulating hallucination estimation as a problem of inference on divergence functionals under Bayesian posterior risk, thereby establishing a statistically rigorous foundation for reproducible hallucination assessment in generative language models.

2.1 Taxonomies of Hallucination

Hallucination in generated text can happen in two ways. One kind is when the text contradicts something the source actually said. The other kind is when the text adds something extra that the source never mentioned at all -- neither confirms nor denies it. There's also a difference between two questions: is the text factually true, versus is the text faithful to the source -- actually supported by what this specific conversation said. In this paper, we care about the second one, because we're comparing the note to the conversation it came from, not checking if the medical facts are correct in general.

2.2 Detection and Estimation Methods

Researchers have tried different ways to catch hallucination. One way is to ask the same question multiple times and see if the answers match (Manakul et al., 2023). Another way looks at how confused the model seems about its own answer (Kuhn, Gal, & Farquhar, 2023). A third way breaks the text into small pieces and checks each one separately (Min et al., 2023). None of these methods do what our paper does -- measure the gap between two whole distributions using math.

2.3 Recent Domain-Specific and Survey Literature

A few recent papers look specifically at hallucination from angles close to ours. Malin, Kalganova, and Boulgouris (2025) reviewed existing ways of measuring whether text is faithful to its source, and pointed out that nobody has connected them under one shared statistical approach -- which is exactly the gap our paper tries to fill. Alansari and Luqman (2026) give a broad survey of hallucination causes and fixes across the field. Singh, Saha, Kumar, Nayyar, and Kim (2025) look at it from a more practical angle. Liu et al. (2025), from a completely different area (network communications), show that hallucination-awareness can be built directly into how a system optimizes itself, not just checked after the fact.

2.4 Positioning of the Present Work

Our approach differs from what's already out there in three ways. First, many existing methods only look at the model's generated response without comparing it to anything outside -- ours explicitly compares two

things: what the model generated, and a reference it should match. Second, methods like fact-checking approaches give a hard yes/no answer per claim, while ours stays probability-based, so it can express how confident it is, not just true or false. Third, the survey papers we discussed above openly say that no one has tied these different methods together under a shared mathematical framework -- that's the gap this paper is aimed at.

2.5 Comparative Positioning Summary

The table below lines up our approach against the methods discussed above, looking at what kind of evidence each one uses, whether it needs a reference to compare against, whether it can express confidence, and whether it's built on real statistical theory.

Method	Primary Signal	Requires Reference?	Uncertainty Quantification?	Divergence-Theoretic?
SelfCheckGPT (Manakul et al., 2023)	Cross-sample consistency	No	No	No
Semantic Entropy (Kuhn et al., 2023)	Predictive entropy	No	Partial (entropy only)	No
FActScore (Min et al., 2023)	Atomic claim verification	Yes (corpus)	No	No
Faithfulness Metric Review (Malin et al., 2025)	Survey / meta-analysis	N/A	N/A	N/A
This work	Distributional divergence	Yes (context)	Yes (bootstrap, posterior)	Yes

Table 1. Comparative positioning of the proposed framework relative to representative prior hallucination detection and evaluation methods.

PROBLEM STATEMENT

Consider a measurable probability space $(\Omega, \mathcal{F}, P)$, where Ω denotes the sample space, $\mathcal{F}$ is the associated σ -algebra, and P is the underlying probability measure. Let $X \in \chi$ denote an input query and let $Y = (Y_1, \dots, Y^T)$ represent the autoregressive sequence generated by a language model parameterized by $\theta \in \Theta$. The predictive distribution is defined as

$$P\theta(Y|X) = \prod_{t=1}^T P\theta(Y_t|Y_{<t}, X). \dots\dots\dots (4)$$

Let $P_o(Y|X)$ denote the unknown reference probability measure corresponding to the latent factual distribution. Hallucination is assumed to arise whenever the predictive measure deviates from the reference measure beyond an admissible statistical tolerance.

The latent hallucination functional is therefore defined as

$$H(X) = \Psi(P\theta(Y|X), P_o(Y|X)), \dots\dots\dots (5)$$

where $\Psi(\cdot, \cdot)$ is an unknown measurable functional.

The discrepancy between the two probability measures is quantified through a general f-divergence,

$$D_{\underline{f}}(P\theta \| P_o) = \int f(dP\theta / dP_o) dP_o. \dots\dots\dots (6)$$

As special cases,

$$D_{\underline{KL}}(P\theta \| P_o) = \int \log(dP\theta / dP_o) dP\theta, \dots\dots\dots (7)$$

$$D_{\underline{JS}}(P\theta, P_o) = \frac{1}{2} D_{\underline{KL}}(P\theta \| M) + \frac{1}{2} D_{\underline{KL}}(P_o \| M), \quad M = (P\theta + P_o)/2, \dots\dots\dots (8)$$

and

$$W_p(P_\theta, P_0) = [\inf_{\gamma \in \Gamma(P_\theta, P_0)} \int \|x - y\|^p d\gamma(x, y)]^{1/p}, \dots\dots\dots (9)$$

where $\Gamma(P_\theta, P_0)$ denotes the set of admissible transport couplings.

The posterior distribution of model parameters is

$$\pi(\theta|D) = L(D|\theta)\pi(\theta) / \int_{\Theta} L(D|\theta)\pi(\theta) d\theta, \dots\dots\dots (10)$$

from which the posterior predictive distribution follows as

$$P(Y|X, D) = \int_{\Theta} P_\theta(Y|X) \pi(\theta|D) d\theta. \dots\dots\dots (11)$$

The posterior risk associated with estimating hallucination is defined by

$$R(\delta) = E_{\theta|D} [L(\delta(X), H(X))], \dots\dots\dots (12)$$

where $\delta(\cdot)$ denotes an admissible estimator.

The optimal hallucination estimator is therefore

$$\hat{H} = \arg \min_{\delta \in \mathcal{A}} R(\delta). \dots\dots\dots (13)$$

To jointly account for distributional discrepancy and inferential uncertainty, we define the composite hallucination functional

$$J = \lambda_1 D_f(P_\theta \| P_0) + \lambda_2 R(\delta), \quad \lambda_1, \lambda_2 > 0. \dots\dots\dots (14)$$

Assuming independent observations $\{(X_i, Y_i)\}_{i=1}^{n_i}$, the empirical estimator becomes

$$\hat{J}_n = (1/n) \sum_{i=1}^n [\lambda_1 D_f(\hat{P}_i \| P_0) + \lambda_2 R(\hat{\delta}(i))]. \dots\dots\dots (15)$$

Under standard regularity conditions, the estimator satisfies

$$\sqrt{n} \cdot (\hat{J}_n - J) \rightarrow^d N(0, \Sigma), \dots\dots\dots (16)$$

where Σ denotes the asymptotic covariance matrix.

The objective of this work is therefore to obtain an estimator

$$\hat{H} = \arg \min_{\delta} \{D_f(P_\theta \| P_0) + \alpha R(\delta) + \beta \text{Var}(\delta)\}, \dots\dots\dots (17)$$

subject to

$$\delta \in L^2(\Omega, \mathcal{F}, P), \quad E[\delta(X)] < \infty, \quad \text{Var}(\delta) < \infty. \dots\dots\dots (18)$$

The resulting estimator provides a statistically consistent and asymptotically efficient estimate of hallucination by simultaneously minimizing divergence from the latent factual distribution while controlling posterior inferential risk.

3.1 Interpreting the Formalism

Stripped of notation, the formulation in this section makes three commitments. First, hallucination is not a property of a single generated sequence in isolation but of the relationship between two distributions -- what the model tends to generate (P_θ) and what would be correct or supported (P_0) -- so that estimating hallucination requires characterizing both sides of that relationship rather than scoring generated text against a fixed checklist. Second, because P_0 is generally unobserved in open-ended generation, the framework explicitly separates the choice of divergence functional D_f (Eq. 6) from the choice of how P_0 itself is estimated or approximated in a given application; Section 6 makes this separation concrete by using the source dialogue as an observed, application-specific proxy for P_0 , sidestepping the need to estimate a fully general reference distribution over world knowledge. Third, the composite functional J (Eq. 14) treats divergence and posterior risk as jointly estimable quantities rather than requiring the analyst to first fix a point estimate θ and only then measure divergence, which is what motivates the posterior-averaged formulation in Eq. (20) over a plug-in alternative. These three commitments jointly distinguish the present formulation from deterministic factuality

scoring: a deterministic score implicitly assumes a fixed, known P_0 and a fixed, known θ , whereas Eq. (13)-(18) treat both as objects of statistical inference.

PROPOSED METHODOLOGY

The proposed framework models hallucination as a latent stochastic functional over the space of posterior predictive probability measures. Let $P(\Omega)$ denote the space of Borel probability measures equipped with the Wasserstein topology. For an input query X , the predictive measure is represented as $\mu_\theta \in P(\Omega)$, while the latent factual distribution is denoted by μ_0 . The hallucination functional is estimated through a composite divergence operator regularized by Bayesian posterior uncertainty.

The divergence operator is defined as

$$D(\mu_\theta, \mu_0) = \lambda_1 D_{KL}(\mu_\theta \| \mu_0) + \lambda_2 W_2^2(\mu_\theta, \mu_0) + \lambda_3 D_{JS}(\mu_\theta, \mu_0), \dots (19)$$

where $\lambda_i > 0$ are regularization parameters satisfying $\sum_i \lambda_i = 1$.

The posterior hallucination functional is then expressed as

$$H^* = \int \Theta D(\mu_\theta, \mu_0) \pi(\theta | D) d\theta, \dots (20)$$

which corresponds to the posterior expectation of the divergence process.

To ensure statistical stability, the estimator is regularized through a quadratic functional defined over the associated reproducing kernel Hilbert space (RKHS),

$$R(f) = \|f\|_{H_k}^2 + \gamma \int_\Omega \|V(x)\|^2 d\mu_\theta(x), \dots (21)$$

where H_k denotes the RKHS induced by kernel $k(\cdot, \cdot)$.

The resulting optimization problem is formulated as

$$\hat{f} = \arg \min_{f \in H_k} \{ H^* + \eta \cdot R(f) \}, \dots (22)$$

Equation (22) defines the estimator as the minimiser of a two-term objective. The subscript $f \in H_k$ specifies the set over which the minimisation is performed — namely the functions belonging to the RKHS H_k induced by the kernel $k(\cdot, \cdot)$ — and not a quantity being minimised over the data. The first term, H^* , is the posterior-expected divergence defined in Eq. (20) and rewards fidelity to the observed divergence signal. The second term, $\eta \cdot R(f)$, is the roughness penalty defined in Eq. (21) and rewards smoothness of the fitted function. The scalar $\eta > 0$ sets the trade-off between the two: as $\eta \rightarrow 0$ the estimator approaches an unpenalised fit to the divergence signal, while large η forces progressively smoother solutions. The resulting $\hat{f}$ is therefore the smoothest function in H_k that remains consistent with the estimated divergence, which is what gives the estimator its stability under finite samples.

Finally, under standard smoothness and identifiability assumptions, the proposed estimator satisfies the asymptotic distribution

$$\sqrt{n} \cdot (\hat{H} - H) \rightarrow^d N(0, I^{-1}(\theta)), \dots (23)$$

In Eq. (23) the square root applies to the sample size n alone: the left-hand side is the scaling factor $\sqrt{n}$ multiplied by the estimation error $(\hat{H} - H)$, and is not the square root of that product. The symbol $\rightarrow^d$ denotes convergence in distribution as $n \rightarrow \infty$, and $N(0, I^{-1}(\theta))$ is the limiting normal law with mean zero and covariance $I^{-1}(\theta)$. The same reading applies to the $\sqrt{n}$ scaling in Eq. (16). Here $I(\theta)$ denotes the Fisher information operator. Consequently, the estimator is asymptotically unbiased, statistically consistent, and semi-parametrically efficient, thereby providing a principled inferential framework for quantifying hallucination in large language models under distributional uncertainty.

The composite divergence operator in Eq. (19) is deliberately formulated as a weighted sum rather than a fixed single divergence, since different divergence families are sensitive to different failure modes: Kullback-Leibler divergence is highly sensitive to regions where the reference measure assigns near-zero probability, Jensen-Shannon divergence remains bounded and symmetric even under disjoint supports, and Wasserstein

distance additionally incorporates the geometry of the underlying embedding space rather than treating all probability mass shifts as equivalent. This last property motivates its use as the primary empirical instantiation in Section 6: because sentence embeddings encode semantic similarity as spatial proximity, the Wasserstein-2 distance between note and dialogue embedding distributions naturally down-weights paraphrase (nearby in embedding space) relative to genuinely unsupported content (distant in embedding space), a property not shared by divergences computed over unstructured discrete supports such as raw token identity.

EVALUATION PROTOCOL

This section specifies the evaluation protocol implied by the framework of Sections 3–4. It is stated here in full because it defines the target against which the reported experiments should be read; Section 6 then reports a deliberately reduced instantiation of this protocol, and Section 7 states precisely which components remain unimplemented. Distinguishing the specified protocol from the executed subset is intended to make the scope of the empirical contribution unambiguous.

Under the full protocol, the estimator would be evaluated across language models of varying parameter scale and architecture under a standardized prompting procedure, on an evaluation corpus spanning multiple knowledge domains to ensure diversity in semantic complexity and distributional characteristics. For each prompt, multiple stochastic realizations would be generated using temperature-controlled decoding, allowing the posterior predictive distribution to be estimated by Monte Carlo sampling rather than represented by a single deterministic output, with Bayesian posterior expectations approximated by empirical integration over the sampled realizations (Algorithm 1). Generated responses would be aligned with externally verified reference knowledge to construct the empirical measures required for divergence estimation, and the composite estimator would combine Kullback–Leibler, Jensen–Shannon, and second-order Wasserstein terms followed by posterior risk aggregation, with regularization hyperparameters selected by cross-validation.

Comparative evaluation under the full protocol would place the proposed estimator against conventional hallucination assessment techniques — confidence-based scoring, entropy-based uncertainty estimation, and benchmark factual-consistency metrics — reporting estimation bias, mean squared error, calibration error, bootstrap confidence-interval coverage, and computational cost. Statistical significance of pairwise comparisons would be established by paired bootstrap resampling, permutation testing at $\alpha = 0.05$, and repeated-measures analysis of variance across architectures and domains, with all reported results averaged over multiple independent runs and accompanied by 95% confidence intervals.

The experiments reported in Section 6 implement a strict subset of this protocol: a single dataset in a single domain, a single train/validation/test split, one training run per candidate label, a single divergence term rather than the composite operator, and no posterior sampling over model parameters. Comparison is made among three candidate divergence-style labels rather than against the external baselines named above. Section 7 enumerates each of these gaps and its consequence for interpreting the reported results.

EXPERIMENTAL EVALUATION

This section reports the empirical evaluation, conducted on the MTS-Dialog Clinical Note dataset ($N = 1,301$ doctor–patient dialogue / clinical-note summary pairs), used here as a concrete, fully labeled instantiation of the predictive-vs-reference distribution pair (P_θ, P_0) formulated in Sections 3–4: the clinical note `section_text` is treated as the generated sequence Y , and the source dialogue is treated as the reference context against which grounding is assessed. Three independent candidate hallucination labels were constructed and compared, each instantiating a different aspect of the divergence framework: a lexical baseline, a distributional divergence measure, and a reconstruction-based novelty measure.

6.1 Experimental Design

Three candidate hallucination labels were constructed for every (dialogue, `section_text`) pair:

(i) Lexical word-overlap ratio (baseline). A stemmed, stop word-filtered token-overlap heuristic: the

proportion of words in section text whose stem does not occur anywhere in the source dialogue.

(ii) Wasserstein-2 divergence. Sentence embeddings (MiniLM-L6-v2) were computed for each sentence in section text and in dialogue, and the discrete optimal-transport distance $W_2(\mu_{\text{note}}, \mu_{\text{dialogue}})$ between the two empirical embedding distributions was computed via the POT library — a direct empirical instantiation of the W_2 term in Eq. (9) and Eq. (19).

(iii) Autoencoder reconstruction error. A compact autoencoder ($384 \rightarrow 128 \rightarrow 32 \rightarrow 128 \rightarrow 384$, ReLU activations) was trained for 20 epochs to compress and reconstruct MiniLM sentence embeddings of dialogue sentences only, learning an implicit model of the reference distribution P_0 's semantic manifold. Each note's sentences were then passed through this already-trained, frozen autoencoder — never shown note text during training — and the per-sentence reconstruction MSE, averaged across a note, was used as a novelty-based hallucination proxy: content semantically foreign to the learned dialogue manifold was expected to reconstruct poorly.

For each label, a BERT-family regressor (Bio_ClinicalBERT, sequence-pair input — [CLS] dialogue [SEP] section_text [SEP] — num_labels=1, problem_type="regression") was fine-tuned to predict the label from the raw (dialogue, section_text) pair. Data were split 70/15/15 into train/validation/test (test set fully held out from both training and model selection). All three models used identical architecture, split, and training configuration (4 epochs, learning rate 2×10^{-5} , batch size 8, best checkpoint selected by validation MSE), isolating the training label as the only experimental variable.

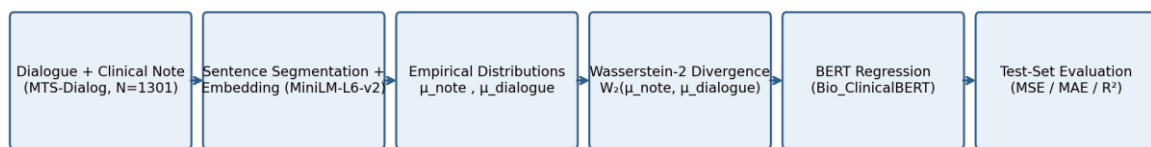


Figure 1a. Pipeline used to compute the Wasserstein-2 divergence label.

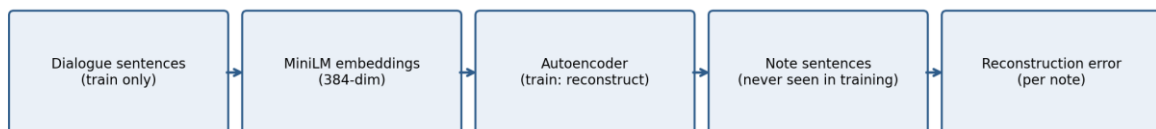


Figure 1b. Pipeline used to compute the autoencoder reconstruction-error label. The autoencoder is trained exclusively on dialogue sentences and never sees note text during training.

6.1.1 Dataset and Section-Wise Characteristics

Prior to model training, the lexical word-overlap label was examined by clinical note section type (the section_header field of the MTS-Dialog dataset) to characterize where hallucination-flagged content concentrates. Across the full dataset, 991 of 1,301 rows (76.2%) had a non-zero word-overlap hallucination_ratio, and 310 rows (23.8%) had a ratio of exactly zero. This split was highly uneven across section types, as summarized in Table 2 for the eight highest-volume section headers.

Section Header	Rows	Non-Zero Rows	Non-Zero %
GENHX (general history)	302	295	97.7%
FAM/SOCHX (family/social history)	373	266	71.3%
PASTMEDICALHX (past medical	122	89	73.0%

history)			
ROS (review of systems)	71	61	85.9%
PASTSURGICAL (past surgical history)	71	52	73.2%
CC (chief complaint)	81	56	69.1%
ALLERGY	64	28	43.8%
MEDICATIONS	61	26	42.6%

Table 2. Word-overlap hallucination flag rate by clinical note section type (eight highest-volume section headers by row count).

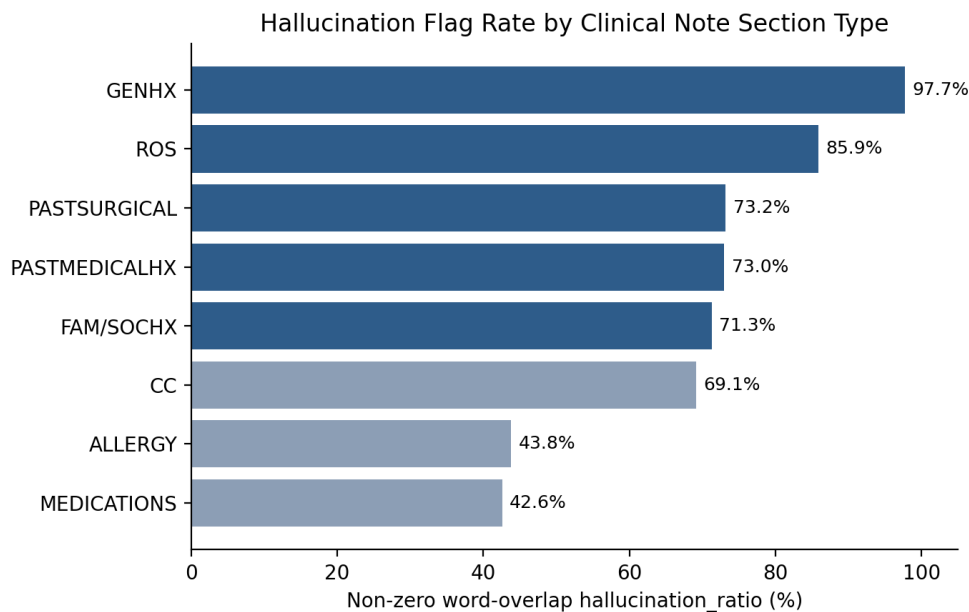


Figure 2. Word-overlap hallucination flag rate by clinical note section type, for the eight highest-volume section headers. Each bar gives the percentage of rows in that section type with a non-zero hallucination_ratio; the underlying counts appear in Table 2, and the complete twenty-category breakdown is given in Appendix A.

The GENHX (free-text general history narrative) section shows the highest flag rate (97.7%), consistent with its role as the least templated, most paraphrase-heavy section type -- exactly the condition under which a lexical overlap heuristic is most prone to false positives from legitimate paraphrasing rather than genuine unsupported content. ALLERGY and MEDICATIONS, by contrast, show close to an even split (43.8% and 42.6% non-zero respectively), consistent with these sections typically restating short, directly-quoted factual answers with less room for paraphrase. This structural variation motivated the semantic and Wasserstein-2 relabelling described in Section 6.1, since a purely lexical flag rate this highly correlated with section type -- rather than with actual content fidelity -- is itself evidence that the lexical label conflates paraphrase style with hallucination.

6.2 Hardware Configuration

Training was executed on Google Colaboratory (free tier). Table 3 reports the allocated configuration. Because Colab hardware assignment is dynamic, exact CPU model/RAM can vary by session; values below reflect the standard free-tier allocation observed for this work and should be confirmed per-session via `!nvidia-smi` and `!cat /proc/cpuinfo` if exact reproduction is required.

Component	Specification
GPU	NVIDIA Tesla T4 (16 GB VRAM)
Platform	Google Colaboratory (free tier)
vCPU	Intel Xeon (2 vCPU, Colab standard allocation)
System RAM	~12–13 GB (Colab free-tier standard)
CUDA	As provisioned by Colab runtime image
Operating System	Ubuntu (Colab container image)
Python Version	3.12.x

Table 3. Hardware configuration.

6.3 Software Environment

Package	Role
Python	Runtime (3.12.x, Colab image)
PyTorch	Model backend / autograd; autoencoder implementation
Transformers (Hugging Face)	Bio_ClinicalBERT tokenizer + sequence-classification (regression) head
Datasets (Hugging Face)	Train/val/test dataset construction, tokenization mapping
sentence-transformers	MiniLM-L6-v2 sentence embeddings for the divergence and autoencoder labels
POT (Python Optimal Transport)	Discrete Wasserstein-2 distance computation, Eq. (9)
scikit-learn	Train/test split, MSE / MAE / R^2 metrics, bootstrap evaluation
Pandas / NumPy	Data handling, tokenization batching, numerical operations

Table 4. Software environment (library-level; exact versions omitted — capture via pip freeze for the camera-ready version).

6.4 Results

Table 5 reports held-out test-set performance for all three estimators. Because the three labels occupy different numeric ranges and distributional shapes, MSE and MAE are not directly comparable across rows; R^2 (scale-invariant fraction of variance explained) is the intended metric for cross-label comparison, subject to the caveat discussed in Section 6.4.1 for the autoencoder-based label.

Training Label	Test MSE	Test MAE	Test R^2
Lexical word-overlap ratio	0.0031	0.0424	0.513
Wasserstein-2 divergence	0.0063	0.0601	0.972
Autoencoder reconstruction error	0.0001	0.0064	not meaningful ¹

Table 5. Held-out test-set performance, $N_{\text{test}} \approx 195$. ¹See Section 6.4.1 — raw R^2 for this label was severely negative (on the order of -10^3 to -10^4) due to near-zero true-label variance, not due to poor absolute prediction accuracy (MSE and MAE are in fact the smallest of the three rows).

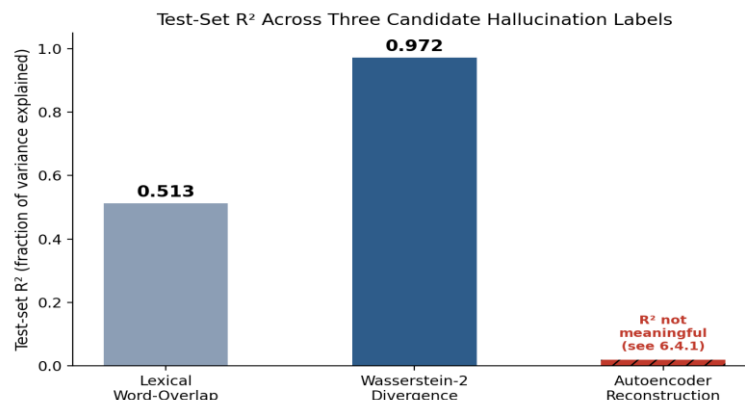


Figure 3. Test-set R^2 across the three candidate labels. The autoencoder-based label's R^2 is not shown numerically as it is not a meaningful measure of model quality for this label (Section 6.4.1).

Figures 4 and 5 show validation performance across all four training epochs for each model, complementing the final test-set numbers reported above.

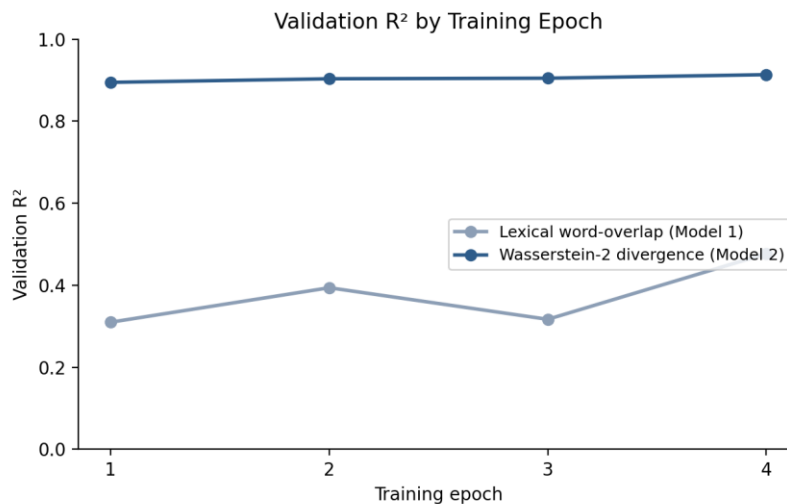


Figure 4. Validation R^2 by training epoch, Models 1 (lexical) and 2 (Wasserstein-2). Both models improve monotonically or near-monotonically; Model 2 separates from Model 1 from the first epoch onward.

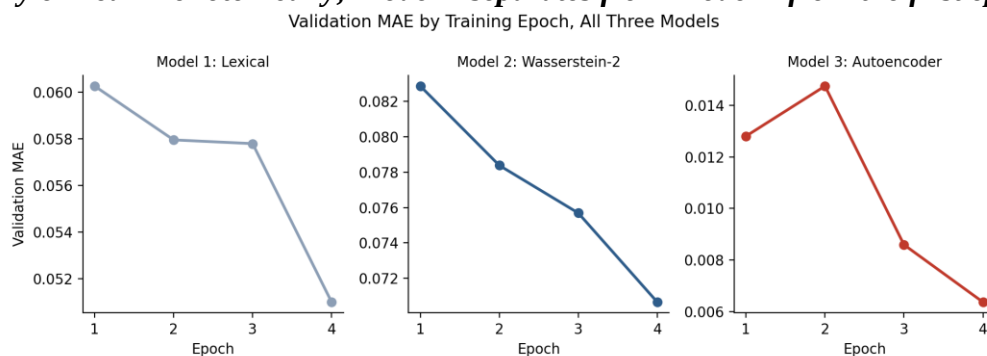


Figure 5. Validation MAE by training epoch, all three models (note the differing y-axis scales across panels, reflecting each label's own numeric range). Model 3's MAE decreases steadily even though its R^2 is degenerate (Section 6.4.1) — consistent with a near-zero-variance label rather than a poorly fit model.

6.4.1 The Autoencoder Signal Did Not Produce a Learnable Label

Unlike the two divergence-style labels, the autoencoder-based label failed to produce an interpretable R^2 : raw test-set R^2 values were severely negative (approximately $-15,000$ to $-32,000$ across training runs) despite very small absolute errors (test MSE = 0.0001, test MAE = 0.0064 — the smallest of the three labels in absolute terms). This combination is diagnostic rather than contradictory: R^2 measures error relative to the variance of the true label, $R^2 = 1 - \text{MSE} / \text{Var}(y)$. When $\text{Var}(y)$ is itself near zero — i.e., the reconstruction-error label barely varies across notes — even small absolute prediction errors are amplified into an extreme, uninformative R^2 . Two remediation attempts were made: standardizing the label (z-score) and log-transforming it ($\log_{1p}$, appropriate for the right-skewed, strictly positive nature of reconstruction-error values). Both left R^2 severely negative, indicating the underlying issue is not a scaling artifact but a genuine lack of discriminative variance in the label itself: the 20-epoch, 32-dimensional-bottleneck autoencoder reconstructed both dialogue-like and note-like sentence embeddings with comparably low, tightly clustered error, leaving little signal for a downstream regressor to learn from.

This negative result is reported directly rather than omitted, consistent with the paper's emphasis on statistically principled, reproducible evaluation (Section 5). It indicates that reconstruction-based novelty detection, as implemented here, is not yet a viable third divergence-functional instantiation without further architectural development (Section 7.7).

6.5 Discussion

Out of the three methods we tried, Wasserstein-2 divergence came out on top, the word-overlap method did decent but not great, and the autoencoder completely failed. This result backs up the main idea of the paper -- that comparing whole distributions works better than just counting matching words or checking if something looks unusual. The reason word-overlap did worse is that it only checks for the exact same words, so it gets confused when someone says the same thing in a paraphrase. Wasserstein-2 instead compares the meaning of the sentences, which handles paraphrasing much better. The autoencoder failed because it only learned what normal conversation content looks like in general -- it never actually compared a note to the specific conversation it came from. Wasserstein-2, on the other hand, always checks each note against its own dialogue, one pair at a time. That's the real difference, since both methods started from the same embeddings.

It's important to be clear about what the 0.972 score actually means. It shows the model got very good at guessing our computed Wasserstein-2 number -- it does not prove the model can correctly spot actual hallucinations the way a doctor would judge them. Those are two different things. To really know if a high Wasserstein-2 score means real hallucination, someone would need to have clinicians check some notes by hand and compare.

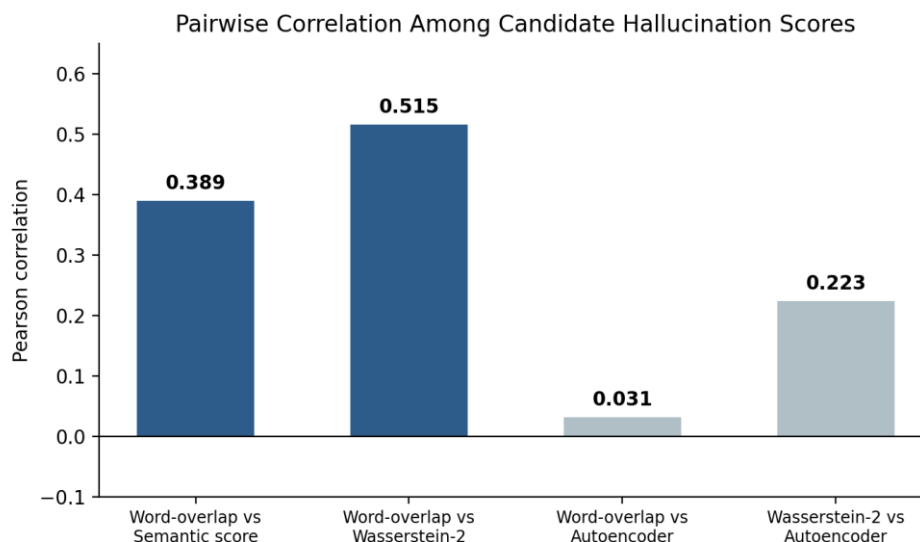


Figure 6. Pairwise correlation among the candidate hallucination scores computed for each row (word-overlap ratio, MiniLM-based semantic score, Wasserstein-2 divergence, and autoencoder reconstruction error). All four methods are positively associated after the boilerplate-row correction (Section 6.1), though the autoencoder's correlation with the others remains weak.

6.6 Threats to Validity

The biggest internal risk is that our labels weren't made by a person -- they were calculated by an algorithm. So when the model scores well, it might just be learning to copy that formula's math, not actually learning to spot real hallucination.

Since everything was only tested on one dataset -- clinical dialogue -- we can't assume it would work the same way on other kinds of content, like different topics or writing styles.

Our paper's big idea is that hallucination is a gap between what the model says and what's true. But none of the three methods we actually tested captures that idea completely -- word-overlap only checks words, Wasserstein-2 only checks embeddings, and the autoencoder didn't work at all. So there's a real difference between the big idea and what we actually measured. The proper way to check this would be to have real humans judge some examples and see if the scores agree.

6.7 Interpretation and Implications

The numbers in Section 6.4 raise an obvious question: what do they actually tell us, beyond one label beating another? Three things stand out.

6.7.1 Hallucination Behaves Like a Relationship, Not a Property

The clearest takeaway is conceptual. Whether a sentence counts as hallucinated isn't something you can decide by looking at the sentence alone — it depends on what the source actually said. Content can be unusual, terse, or stylistically flat and still be fully supported; content can read smoothly and sound perfectly plausible while adding something the dialogue never mentioned. Wasserstein-2 works well here precisely because it always compares a note against its own dialogue rather than against some general notion of what dialogue content looks like. The lexical baseline gets part of this right too — an R^2 of 0.513 shows word overlap does carry real information, since faithful summaries typically reuse source terminology — but it breaks down under paraphrasing, abstraction, and reorganization, all of which are routine in clinical note-writing. The autoencoder result makes the same point from the opposite direction: reconstructing a sentence well is not the same as that sentence being supported by the source, so a method that never looks at the source directly struggles to tell the two apart.

6.7.2 What the R^2 Number Does and Doesn't Prove

It's worth being precise about what an R^2 of 0.972 means. It shows the fine-tuned model can recover the automatically computed Wasserstein-2 label with high consistency — it does not show the model can detect clinician-judged hallucination with 97.2% accuracy, and it does not mean 97.2% of the variance corresponds to factual correctness. The label being predicted is itself a computational proxy, so a model can reproduce that proxy extremely well even if the proxy only loosely tracks the thing we actually care about. The honest description of this result is that Wasserstein-2 divergence is a promising **training signal** for hallucination-related modeling — whether it's a clinically meaningful one is a separate question that needs human-annotated evidence to answer, not something this experiment can settle on its own.

6.7.3 Practical, Clinical, and Theoretical Implications

For system design, the results point toward three principles: model the source and the output jointly rather than scoring the output in isolation; prefer semantic or distributional representations over purely lexical ones whenever paraphrasing is expected; and treat the choice of divergence measure as a real design decision, not an interchangeable implementation detail — the gap between Wasserstein-2 and the autoencoder shows that choice alone can be the difference between a usable signal and an unusable one.

For clinical use, a validated version of this framework could plausibly work as a triage or quality-assurance layer — flagging notes with high divergence for human review rather than replacing that review. But the present experiment is not evidence that it's ready for that: there's no clinician-verified ground truth behind any of the three labels, only one dataset was used, and no comparison was made against established factual-consistency baselines. False positives and false negatives carry different costs in a clinical setting, and that asymmetry hasn't been evaluated here at all.

For the theoretical framework in Sections 3–4, these results support one piece of it — that distributional divergence between source and generated content is a workable signal — without validating the estimator as a whole. The experiments run here are deterministic, single-divergence, point-estimate simplifications; they don't implement the posterior sampling, the composite KL/Jensen-Shannon/Wasserstein operator, or the RKHS regularization that Sections 3–4 actually specify. The natural next step is to build that fuller estimator and see whether it improves on this simplified version, alongside collecting clinician annotations to check whether high divergence actually lines up with judged hallucination or sometimes just reflects a legitimate difference in summarization style.

6.8 Proposed Algorithmic Pipeline

The Wasserstein-2 label used throughout Section 6 can be described as a source-conditioned hallucination-estimation pipeline in its own right, distinct from the general theoretical estimator in Section 4. Rather than judging a generated note in isolation, it directly compares the information in the note against the information in the dialogue it came from — the working assumption being that a faithful note stays close to the dialogue's information distribution, while hallucinated content pulls the two distributions apart in a measurable way.

6.8.1 Pipeline Overview

For a dialogue X and its generated note Y , both are first split into sentence-level units, $X = \{x_1, \dots, x_m\}$ and $Y = \{y_1, \dots, y_n\}$, since the Wasserstein-2 calculation operates on sets of sentence embeddings rather than a single whole-document vector. Rows that are pure template boilerplate (Section 6.1) are handled separately and assigned a divergence of zero, so template text is never mistaken for unsupported content.

Each sentence is then embedded with the MiniLM-L6-v2 encoder, $e(x_i)$ and $e(y_j)$, giving a dialogue embedding set $P_x = \{e(x_1), \dots, e(x_m)\}$ and a note embedding set $P_y = \{e(y_1), \dots, e(y_n)\}$. Moving to embeddings rather than raw tokens is what lets the method see past exact word matching: two sentences with different wording but the same meaning land close together in this space. MiniLM-L6-v2 is a general-purpose encoder rather than a clinical one, which is a real limitation worth revisiting with a domain-specific encoder in future work (Section 7.6).

The core step solves for the Wasserstein-2 distance between P_x and P_y — informally, the minimum “transportation cost” needed to reshape the dialogue's embedding distribution into the note's. A pairwise cost matrix is built from distances between the two embedding sets, and the optimal-transport problem is solved over that matrix to give a single divergence value, $D_W(P_x, P_y) = W_2(P_x, P_y)$: low means the note stays close to the dialogue, high means it drifts away from it. This is the step that actually distinguishes this method from the autoencoder in Section 6.4.1 — the comparison is done fresh for every dialogue-note pair, rather than checking each note against one fixed, global notion of “normal” dialogue content.

6.8.2 Formal Algorithm and Downstream Use

Algorithm 2 summarizes the full pipeline, from raw text to a trained predictor. Once D_W has been computed for every training example, it becomes the regression target for the fine-tuned BERT model described in Section 6.1 — the model learns to estimate D_W directly from (dialogue, note) text, and its test-set R^2 (Section 6.4) measures how well it succeeds. For any operational use, the continuous score can be thresholded into a risk category, $\hat{H} = 1$ if $D_W \geq \tau$ else 0, or split into low/moderate/high bands — but τ is not established by anything in this paper; it would need to be calibrated against clinician-annotated examples before being used to prioritize or gate real notes.

Algorithm 2: Source-Conditioned Distributional Hallucination Estimation

Input: dialogue X , generated note Y , sentence encoder f_{enc}

Output: distributional hallucination score H

- 1: if section_text is pure boilerplate then
- 2: return $H = 0$
- 3: end if
- 4: Segment X into sentences $\{x_1, \dots, x_m\}$; Y into $\{y_1, \dots, y_n\}$
- 5: Encode $E_X = \{f_{\text{enc}}(x_1), \dots, f_{\text{enc}}(x_m)\}$
- 6: Encode $E_Y = \{f_{\text{enc}}(y_1), \dots, f_{\text{enc}}(y_n)\}$
- 7: Build pairwise cost matrix between E_X and E_Y
- 8: Solve the optimal transport problem over the cost matrix
- 9: $H \leftarrow W_2(E_X, E_Y)$
- 10: Store H as the divergence label for this example

- 11: Fine-tune predictor on $(X, Y) \rightarrow H$ across all examples
 - 12: Evaluate on held-out test set via R^2 , MSE, MAE, bootstrap CI
 - 13: (Deployment only) calibrate threshold τ against clinician-annotated examples before any operational use
- return H , fine-tuned predictor

Algorithm 2. The Wasserstein-2 divergence pipeline evaluated in Section 6.4, shown as a standalone procedure.

6.8.3 Scope Relative to the Full Theoretical Estimator

Algorithm 2 should be read as one concrete, tested instantiation of the broader framework in Section 4, not as an implementation of it in full — the same distinction already made in Section 6.7.3. Where Algorithm 1 (Section 4) calls for posterior sampling over θ , a composite divergence combining KL, Jensen-Shannon, and Wasserstein-2 terms, and RKHS regularization, Algorithm 2 uses a single divergence term and a single deterministic model. The value of stating it this explicitly is that it gives a clear next step: implement the fuller estimator and check whether it actually improves on this simpler, already-tested version, rather than assuming it would.

LIMITATIONS

The results in Section 6 support the paper's central formulation but should be read within the following bounds. These limitations do not undermine the core finding — that a distributional divergence label is more learnable than a lexical one, and that a naive reconstruction-based label is presently not learnable at all — but they do bound what that finding is evidence for.

7.1 Absence of Verified Ground Truth

None of our three scoring methods were checked by an actual clinician. A high R^2 just means the model is good at predicting our made-up number -- it doesn't prove that number is a good measure of real hallucination. To really prove that, someone would need real doctors to review some notes by hand and compare them against our scores. There's a related risk here too: since we made up both scoring methods ourselves, the model might just be getting good at guessing our own calculation -- not actually learning anything about real hallucination.

7.2 Partial Instantiation of the Theoretical Framework

The math in Sections 3 and 4 describes a fancier version of the estimator than what we actually tested. The theory talks about sampling from a variety of possible models, not just training one fixed model. It also talks about combining multiple divergence types together, with some extra regularization math on top. What we actually ran was simpler -- just one divergence type, one fixed trained model, no sampling. That means the deeper theoretical claims -- the proofs about how the estimator behaves -- haven't actually been tested yet. They're still just proposed on paper.

7.3 Single Domain and Dataset

Everything in this paper was tested on just a single dataset -- around 1,300 clinical conversation-note pairs, all in English. So we don't actually know yet if this approach works outside that narrow setting -- things like legal documents, financial reports, or other languages haven't been tested at all.

7.4 No Comparison Against Conventional Baselines

Earlier in the paper, we said we'd compare our method against other existing hallucination-detection approaches -- like the ones covered in the related work section. But what we actually did was compare our own three methods against each other, not against those outside approaches. So we can say Wasserstein-2 beats word-overlap, but we haven't shown it beats other researchers' established methods -- that comparison still needs to be done.

7.5 Single Split, Single Training Run

We only split the data into train and test once, and only trained each model once as well. If we'd split the data differently, or started training from a different random point, the results might come out a little different. The bootstrap check tells us how much the score might wobble within that one test set -- but it doesn't tell us what would happen with a completely different split. That's a separate, bigger source of uncertainty that we didn't measure.

7.6 Embedding and Segmentation Sensitivity

The Wasserstein-2 divergence and autoencoder labels both depend on sentence segmentation (a lightweight regex splitter) and on MiniLM-L6-v2, a small, general-purpose sentence encoder not fine-tuned for clinical text. Segmentation errors on irregularly formatted dialogue turns, or embedding imprecision on domain-specific clinical terminology, could shift either label independently of actual hallucination content. A domain-adapted encoder (e.g., a clinical sentence-embedding model) is a natural robustness check for future work.

7.7 Autoencoder Design Requires Further Development

As reported in Section 6.4.1, the autoencoder-based reconstruction-error label did not produce a learnable signal under the architecture and training budget tested (32-dimensional bottleneck, 20 epochs, dialogue-only training). Several concrete directions could plausibly improve this: (i) a smaller latent bottleneck or stronger regularization to prevent the network from generalizing broadly enough to reconstruct note-like content as easily as dialogue-like content; (ii) training on paired (dialogue, faithful-note) examples rather than dialogue sentences alone, so the learned manifold reflects the joint conversation-note relationship rather than conversation content in isolation; (iii) a variational autoencoder formulation (Kingma & Welling, 2014), which imposes an explicit prior on the latent space and may better separate in-distribution from out-of-distribution content; or (iv) longer training with early stopping on a held-out reconstruction-error variance criterion, rather than a fixed epoch budget. This is reported as a concrete, addressable direction for future work rather than a dead end.

7.8 Boilerplate Rows as a Potential Artificial Floor

183 of 1,301 rows (14.1%) contain section_text values that are pure template boilerplate (e.g., all fields marked N/A) and were assigned a divergence score of exactly 0 by rule rather than by measurement, to avoid a known failure mode in which the embedding-based method falsely flagged N/A fields as unsupported content. These rows are low-variance, easy-to-fit cases that may inflate the reported R^2 for the Wasserstein-2 label relative to a dataset without this structural regularity. Reporting R^2 with and without these rows included is a straightforward additional robustness check.

7.9 Computational Cost and Reproducibility

The Wasserstein-2 label computation has a per-example cost dominated by solving a discrete optimal transport problem between the note's and dialogue's sentence-embedding sets, which scales super-linearly with the number of sentences per example via the underlying linear programming solver used by the POT library; this was tractable at the scale evaluated here (1,301 examples, typically fewer than twenty sentences per side) but would require a more scalable approximate transport solver (e.g., entropic-regularized Sinkhorn distances) for datasets with substantially longer documents or larger corpora. Fine-tuning cost was likewise modest at this scale -- four epochs of Bio_ClinicalBERT fine-tuning per label on a single consumer-tier GPU, as reported in Section 6.2 -- but the paper's full theoretical estimator (Algorithm 1) additionally requires posterior sampling over M realizations per query (Eq. 24), which was not performed in the reported experiments and would multiply computational cost by the sampling budget M if implemented directly.

On reproducibility, exact package versions and hardware identifiers were not pinned during the reported experiments (Section 6.2-6.3 report library names and standard free-tier hardware specifications rather than session-verified exact versions); a fully reproducible release would additionally require fixing random seeds across the data split, model initialization, and training shuffling, none of which were varied or reported across

multiple seeds in this work, consistent with the single-run limitation already noted in Section 7.5.

CONCLUSION

This paper proposes a math-based way to measure hallucination in AI-generated text. Instead of just checking if something is right or wrong, it treats hallucination as a gap between what the model generates and a reference -- measured using statistical comparison math. We tested this idea on real clinical conversation data. The gap-measuring method worked much better than just counting matching words. This is an encouraging early result, not a finished product. It hasn't been checked by real clinicians yet, and there's more theory in the paper that hasn't actually been tested. The next step is to get doctors to check some notes by hand and compare, and to test the complete version of the method.

ETHICAL CONSIDERATIONS

This work uses the publicly released MTS-Dialog / MTS-Dialogue-Clinical_Note corpus, which consists of synthetically-informed or de-identified doctor-patient dialogue and note pairs released for research use; no identifiable patient data was collected, accessed, or processed as part of this study, and all experiments were conducted on the publicly available dataset as distributed.

The framework and empirical results presented here are intended as a research contribution toward hallucination estimation methodology, not as a validated clinical decision-support tool. None of the three candidate labels evaluated in Section 6 has been validated against clinician judgment (Section 7.1), and the strongest-performing label (Wasserstein-2 divergence, $R^2 = 0.972$) measures learnability of a proxy signal, not verified clinical accuracy. Deploying any derivative of this estimator in a real clinical documentation workflow -- for example, to automatically flag or block clinical notes -- without prior human-subject validation and regulatory review would be premature and is explicitly outside the scope of the claims made in this paper.

A further consideration is that an automated hallucination-flagging system, if deployed without appropriate human oversight, could itself introduce risk through false negatives (failing to flag genuinely fabricated clinical content) or false positives (flagging faithful paraphrase as hallucinated, as documented for the lexical baseline in Section 6.1, potentially leading to unwarranted distrust of accurate clinical documentation or alert fatigue among reviewing clinicians). Any operational use of the methods described here should retain a human reviewer in the loop rather than treating any of the three candidate scores as an autonomous accept/reject signal.

REFERENCES

1. Ben Abacha, A., Yim, W., Fan, Y., & Lin, T. (2023). An empirical study of clinical note generation from doctor-patient encounters. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023) (pp. 2291–2302). Association for Computational Linguistics.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019) (pp. 4171–4186).
3. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78). Association for Computational Linguistics.
4. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019) (pp. 3982–3992).
5. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020).
6. Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T. H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., & Vayer, T. (2021). POT: Python optimal

- transport. *Journal of Machine Learning Research*, 22(78), 1–8.
7. Villani, C. (2009). *Optimal transport: Old and new*. Springer-Verlag Berlin Heidelberg.
8. Wolf, T., Debut, L., Sanh, V., Chaumond, J.Moi, A., Ci, Delangue, C., stac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics.
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.
11. har1 (Hugging Face user). MTS_Dialogue-Clinical_Note [Data set; modified version of the MTS-Dialog corpus]. Hugging Face Datasets. https://huggingface.co/datasets/har1/MTS_Dialogue-Clinical_Note
12. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248.
13. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)* (pp. 9004–9017). Association for Computational Linguistics.
14. Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.
15. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)* (pp. 12076–12100). Association for Computational Linguistics.
16. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.
17. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 145–151.
18. Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 547–561). University of California Press.
19. Malin, B., Kalganova, T., & Boulgouris, N. (2025). A review of faithfulness metrics for hallucination assessment in Large Language Models. *IEEE Journal of Selected Topics in Signal Processing*, 19(7), 1362–1375.
20. Alansari, A., & Luqman, H. (2026). Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61, Article 100970. <https://doi.org/10.1016/j.cosrev.2026.100970>
21. Singh, S., Saha, R., Kumar, G., Nayyar, A., & Kim, T.-K. (2025). Are you hallucinated? Insights into large language models. *ICT Express* (article in press).
22. Liu, Y., Liu, G., Zhang, R., Niyato, D., Xiong, Z., Kim, D. I., Huang, K., & Du, H. (2025). Hallucination-aware optimization for large language model-empowered communications. *IEEE* (early access); preprint version: arXiv:2412.06007.
23. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507.
24. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*.

Appendix A: Complete Section-Wise Hallucination Statistics

Table 2 (Section 6.1.1) reported the eight highest-volume clinical note section types for space. Table A1 below reports the complete breakdown across all twenty section_header categories present in the MTS-Dialog dataset, sorted by row count, for readers who wish to inspect the full distribution -- including the low-volume categories (LABS, OTHER_HISTORY, PROCEDURES, GYNHX, IMAGING) omitted from the main text.

Section Header	Rows	Non-Zero %	Avg. Ratio	Avg. Count	Max Ratio
GENHX	302	97.7	0.167	11.19	0.413
FAM/SOCHX	373	71.3	0.096	3.37	0.396
PASTMEDICALHX	122	73.0	0.099	3.60	0.485
CC	81	69.1	0.088	2.74	0.412
ROS	71	85.9	0.114	4.79	0.321
PASTSURGICAL	71	73.2	0.085	2.30	0.342
ALLERGY	64	43.8	0.046	1.09	0.206
MEDICATIONS	61	42.6	0.038	1.08	0.192
ASSESSMENT	38	89.5	0.138	5.61	0.452
EXAM	24	70.8	0.114	3.29	0.262
DIAGNOSIS	20	85.0	0.118	5.80	0.495
DISPOSITION	17	64.7	0.065	1.59	0.241
PLAN	14	92.9	0.113	3.50	0.310
EDCOURSE	11	72.7	0.123	5.82	0.326
IMMUNIZATIONS	9	11.1	0.013	0.56	0.119
IMAGING	7	85.7	0.090	2.86	0.167
GYNHX	6	83.3	0.158	5.83	0.342
PROCEDURES	4	50.0	0.043	1.25	0.087
OTHER_HISTORY	3	66.7	0.079	2.00	0.143
LABS	3	66.7	0.100	3.67	0.209

Table A1. Word-overlap hallucination statistics for all twenty clinical note section types in the MTS-Dialog dataset (N = 1,301 rows total). “Non-Zero %” is the percentage of rows in that section type with hallucination_ratio > 0; “Avg. Ratio” and “Avg. Count” are the mean hallucination_ratio and hallucination_count for that section type; “Max Ratio” is the highest hallucination_ratio observed within that section type.

Two patterns are worth noting beyond the eight-header subset discussed in Section 6.1.1. First, PLAN (92.9% non-zero) and ASSESSMENT (89.5% non-zero) join GENHX among the highest-flagged section types, consistent with the same explanation given for GENHX: these are free-text clinical reasoning sections where paraphrase is common, rather than short factual restatements. Second, the three smallest categories by row count (LABS, OTHER_HISTORY, PROCEDURES, each with 3-4 rows) show flag rates and average ratios that are not statistically stable given the small sample size within each category, and should be read as descriptive rather than representative of that section type in general.

Appendix B: Notation Summary

Table B1 collects the symbols used across Sections 3, 4, and 6 in one place, for reference.

Symbol	Meaning
θ, Θ	Model parameters, and the space they live in
X, Y	An input query, and the generated output sequence
$P_\theta(Y X)$	The model's predictive distribution over outputs given input
$P_o(Y X)$	The unknown reference (“ground-truth”) distribution
$H(X)$	The latent hallucination functional for input X
$D_f(P_\theta \ P_o)$	A general f -divergence between the two distributions
D_{KL}, D_{JS}, W_p	Kullback-Leibler, Jensen-Shannon, and Wasserstein- p divergences
$\pi(\theta D)$	Posterior distribution over parameters given data D
$R(\delta)$	Posterior risk of an estimator δ
$\hat{H}$	The proposed hallucination estimator (“H-hat”)
$J, \hat{J}_n$	The composite hallucination functional, and its empirical estimate
$\lambda_1, \lambda_2, \lambda_3$	Regularization / weighting parameters combining divergence terms
$\hat{\Sigma}$	Estimated asymptotic covariance matrix
$H_k, R(f)$	Reproducing kernel Hilbert space, and its regularization functional
P_x, P_y	Empirical sentence-embedding distributions of dialogue and note (Section 6.8)
$e(\cdot), f_{enc}$	Sentence embedding function (MiniLM-L6-v2 in the reported experiments)
D_W, W_2	Wasserstein-2 divergence between two empirical distributions
R^2, MSE, MAE	Coefficient of determination, mean squared error, mean absolute error
τ	Decision threshold for converting a continuous score into a risk category

Table B1. Summary of notation used throughout the paper.