HARE: an entity centric evaluation framework for histopathology reports

Anonymous ACL submission

Abstract

Medical domain automated text generation is an active area of research and development; however, evaluating the clinical quality of generated reports remains a challenge, especially in instances where domain-specific metrics are lacking, e.g. histopathology. We propose **HARE (Histopathology Automated Report Evaluation)**, a novel entity-centric framework, composed of a benchmark dataset, a NER model and a novel metric, which prioritizes clinically relevant content by aligning critical histopathology entities between reference and generated reports. To develop the HARE benchmark, we curated a **golden dataset** of 1,196 de-identified diagnostic histopathology reports annotated with domain-specific entities and a **silver dataset** of 1,830 automatically annotated reports from The Cancer Genome Atlas (TCGA). We fine-tuned GatorTronS, a domain-adapted language model to develop HARE-NER which achieved the highest F1-score (0.812) among the tested NER models. The proposed HARE metric outperformed traditional metrics including ROUGE and Meteor, as well as radiology metrics RaTEScore and RadGraph-XL, with the highest correlation to expert evaluations (higher than the second best method, RadGraph-XL, by Pearson $r = 0.061$, Spearman $\rho = 0.048$, Kendall $\tau = 0.066$). We will release HARE, datasets, and the NER model to foster advancements in histopathology report generation, providing a robust framework for improving the quality of histopathology reports.

1 Introduction

Medical report generation has become an increasingly active area of research in clinical natural language processing (NLP) with the goal of automating the creation of specialized clinical documents (Xu et al., 2024; Liu et al., 2025). Among various medical domains, radiology has witnessed the earliest and most notable advancements in automated report generation (Hyland et al., 2023; Nicolson et al., 2023; Wu et al., 2024; Bannur et al., 2024). This progress is partly attributed to the development of domain-specific evaluation metrics that prioritize clinical correctness (Smit et al., 2020; Jain et al., 2021; Delbrouck et al., 2024; Zhao et al., 2024). Unlike general-purpose metrics such as BLEU and ROUGE, these specialized metrics assess the accuracy of radiologically significant entities and findings, thereby offering a more clinically meaningful measure of report quality (Lin, 2004; Papineni et al., 2002; Zhao et al., 2024) and facilitating the development of accurate generative models.

In contrast, the field of histopathology, which involves the microscopic examination of tissue samples to diagnose diseases such as cancer, still relies only on general-purpose lexical metrics for evaluating automatically generated reports (Chen et al., 2023; Guo et al., 2024; Tan et al., 2024; Chen et al., 2024). Histopathology reports are semi-structured, terminology-intensive documents that provide detailed microscopic evaluations of tissue samples, playing a crucial role in disease diagnosis and guiding treatment decisions. Histopathology reports encompass multiple sections, including descriptions of anatomical sites, cellular morphology, tumor classification, staging, further analyzes (e.g. immunohistochemistry (IHC) markers, special stains, or in situ hybridization (ISH)), and the final diagnosis.

Figure 1 shows the difference between the word embeddings of radiology reports (from MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016) and histopathology reports (used in this study). Histopathology word embedding has many areas that are uncovered by radiology word embeddings, making the histopathology reports unsuitable for radiology report evaluation metrics. Conventional lexical evaluation metrics such as METEOR and BERTScore as well as clin-

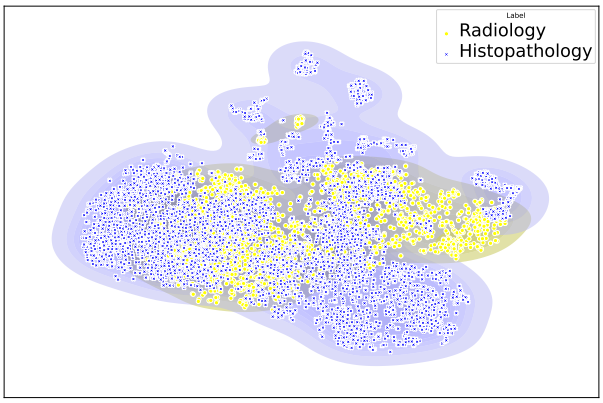

Figure 1: Scatter and density plot of word embeddings for radiology and histopathology reports. The radiology reports are 1,000 randomly sampled reports from MIMIC-CXR dataset and IU-X-ray dataset (Johnson et al., 2019; Demner-Fushman et al., 2016). The histopathology reports are 1,000 randomly sampled reports from both golden and silver datasets used in this study. Reports are embedded using a BERT-base model and reduced to two dimensions using PCA. The density regions highlight where words from each category are concentrated, with "Radiology" shown in yellow and "Histopathology" in blue. This visualization emphasizes the distinct distribution of vocabulary between the two domains.

ical relevance based evaluation metrics designed for radiology reports are insufficient for assessing the quality of automatically generated histopathology reports, as they fail to capture the nuanced histopathological details essential for accurate diagnosis and patient management (Banerjee and Lavie, 2005; Zhang et al., 2019; Smit et al., 2020; Delbrouck et al., 2024; Zhao et al., 2024).

This challenge is further compounded by the scarcity of publicly available datasets for specifically histopathology named entity recognition (NER), which limits the ability to train robust models tailored to the complexities of histopathological language. There is only one NER model and the dataset for pathology reports, but the model and the dataset are not publicly available (Zeng et al., 2023). This gap underscores the need for an entity-centric evaluation metric that can capture the unique characteristics of histopathology reports.

To address this gap, we introduce HARE (Histopathology Automated Report Evaluation): a novel, entity-focused metric designed to assess the clinical quality of generated histopathology reports. In Figure 2, the process of computing the score is demonstrated. HARE captures domain-specific entities (e.g., anatomical sites, microscopic findings, IHC markers, descriptor for final diagnosis) from both candidate and reference reports and quantifies their alignment via a cosine similarity measure (Rahutomo et al., 2012). Our approach is grounded in a comprehensive annotation effort on 1,196 real-world diagnostic histopathology reports, where domain experts labeled critical histopathology entities. We then trained and compared various BERT-based models on these gold annotations, selecting the best-performing model to automatically annotate an additional 1,830 publicly available histopathology reports from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015). This silver annotation set both increases the diversity of available training examples and enhances the model's ability to generalize across different report styles.

By emphasizing the presence and correctness of domain-specific entities, HARE provides a more clinically oriented benchmark than existing lexical metrics. We validated its effectiveness by demonstrating the higher correlation between HARE scores and expert-derived evaluations of generated reports than multiple other available metrics. By releasing both our silver-annotated TCGA dataset and the final trained NER model (which we call HARE-NER), we aim to encourage further research in histopathology NLP and to improve the clinical utility and reliability of automated report-generation systems.

The primary contributions of this paper are as follows:

1. **Introduction of a New Metric (HARE):** We propose a domain-specific evaluation metric for histopathology report generation that focuses on the detection and alignment of significant histopathology entities. To our knowledge, it is the first dedicated metric for this purpose.
2. **Histopathology Score Dataset:** We collect and provide expert histopathologist scores for automatically-generated reports, demonstrating the real-world validity of our HARE metric.
3. **HARE-NER:** We develop a NER model specialized in histopathology (HARE-NER), capable of recognizing critical domain-specific entities such as IHC markers, anatomical sites and descriptor (for final diagnosis), filling a gap where there is no publicly available histopathology-focused NER model.

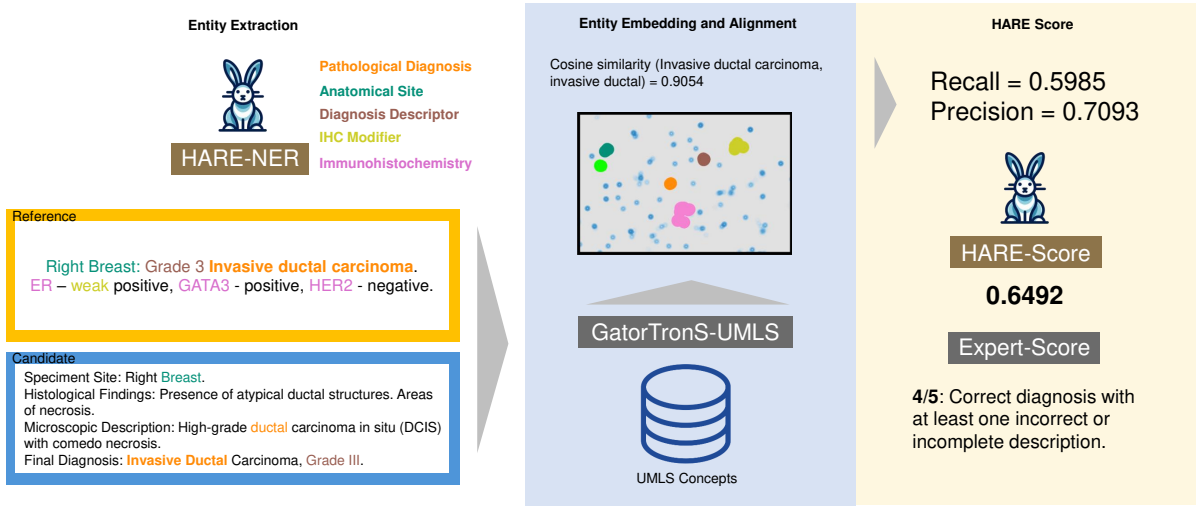

Figure 2: Illustration of the process of computing the HARE score, a novel entity-centric metric for evaluating histopathology report generation.

4. **Open Source:** We will release (1) the silver-annotated TCGA dataset, (2) the final trained NER model as well as the alignment model, and (3) HARE score computation code to facilitate further research and development in both NER and report generation in the histopathology domain.

2 Related Work

While several evaluation metrics have been proposed for radiology, the field of histopathology remains underexplored. Two most recent notable contributions in radiology emphasize the design of domain-specific metrics that capture clinical significance: **RadGraph-XL** and **RaTEScore** (Jain et al., 2021; Zhao et al., 2024).

2.1 RadGraph-XL

RadGraph-XL (Delbrouck et al., 2024) is a large-scale, expert-annotated dataset created for extracting clinical entities and relations from radiology reports. Building upon its predecessor, RadGraph-1.0 (Jain et al., 2021), RadGraph-XL expands annotations to cover multiple anatomy-modality pairs, including chest CT, abdomen/pelvis CT, and brain MRI, in addition to existing chest X-ray data. The dataset consists of over 2,300 reports annotated with 410,000 entities and relations, significantly enhancing its scale and diversity.

RadGraph-XL underscores the importance of clinically relevant entities and relationships in domain-specific metrics. This principle directly informs our work, as we extend it to the histopathology domain by focusing on uniquely critical entities such as features of the histopathological report including pathological diagnosis and IHC marker data.

2.2 RaTEScore

RaTEScore (Zhao et al., 2024) is a domain-specific evaluation metric designed to assess the quality of radiology report generation. Unlike general-purpose metrics such as BLEU or ROUGE, RaTEScore prioritizes clinical accuracy through entity-level assessments. It employs a named entity recognition (NER) module to extract key medical entities (e.g., anatomy, abnormalities, diseases) and a synonym disambiguation encoding module to address challenges such as medical synonymy and negation. The final metric is computed using the cosine similarity of entity embeddings, with adjustments made to reflect the clinical relevance of specific entity types.

To support its development, RaTEScore introduced two foundational resources:

1. **RaTE-NER**: A large-scale dataset for medical NER, covering nine imaging modalities and 22 anatomical regions.
2. **RaTE-Eval**: A benchmark for evaluating metrics, including sentence- and paragraph-level human ratings, as well as comparisons involving synthetic reports.

RaTEScore demonstrated superior alignment with human preferences, achieving the highest correlation scores in evaluations on public datasets such as ReXVal and the RaTE-Eval benchmark. Inspired by RaTEScore's methodology, our proposed

HARE metric adapts the principles of entity recognition and embedding similarity to the histopathology domain, addressing unique challenges such as the interpretation of pathological diagnosis and IHC findings.

2.3 Limitations in Existing Metrics

Although RadGraph-XL and RaTEScore have significantly advanced the evaluation of radiology reports, their applicability is limited to specific modalities (e.g., chest X-rays) and radiological contexts. They do not address the unique linguistic and clinical knowledge of histopathology, which involve detailed morphological assessments and IHC findings.

HARE addresses these limitations by introducing an entity-aware evaluation framework tailored specifically to the histopathology domain. By emphasizing the detection and alignment of domain-specific entities, HARE provides a clinically relevant metric to assess the quality of generated histopathology reports.

3 Methods

In this section, we describe the development of HARE (Histopathology Automated Report Evaluation), a domain-specific evaluation metric designed to assess the clinical quality of generated histopathology reports. Our methodology involves dataset preparation and annotation, named entity recognition (NER) model training, and the design of the HARE metric.

3.1 Dataset Preparation and Annotation

We curated two datasets to support the development of HARE: a manually annotated golden dataset and an automatically annotated silver dataset.

3.1.1 Golden Dataset

We collected 1,196 fully de-identified/anonymized histopathology reports from the pathology department of a large teaching hospital. The reports were from cases across a range of tissue types and diagnoses. The reports were annotated by expert histopathologists using the Inception annotation tool (Klie et al., 2018). Disagreements were resolved by a senior histopathologist. The annotations focused on histopathology-specific entities, including:

- **Anatomical Site:** Entities describing specific tissue regions or locations, such as *breast, lung, kidney, lymph nodes etc.*.
- **Immunohistochemistry Markers:** The presence of immunohistochemical markers such as *CK20, CDX2, ER, PR, Ki-67*.
- **Pathological diagnosis:** The pathological diagnosis, such as *classical Hodgkin lymphoma*.
- **Diagnosis Descriptor:** Provides descriptive characteristics of the pathological diagnosis is uncertain, e.g., 'raises the possibility of'.
- **IHC Modifier:** Used to modify immunohistochemical annotations, e.g., 'patchy' or 'strong.

Entity Type	Golden	Silver
Immunohistochemistry	7,464	180
IHC Modifier	1,398	504
Pathological diagnosis	925	2,101
Anatomical Site	754	664
Diagnosis Descriptor	255	289

Table 1: Entity annotation statistics for the Golden and Silver datasets.

3.1.2 Silver Dataset

To increase diversity, we automatically annotated 1,830 publicly available histopathology reports from the previously published HistGen training and evaluation dataset, which is originally sourced from The Cancer Genome Atlas (TCGA) (Guo et al., 2024; Tomczak et al., 2015). These reports were then further annotated using the best performing model trained with the golden dataset. We then extracted sentences with histopathological descriptions, specifically microscopic findings and final diagnosis characteristics. The breakdown of the number of annotations for the Golden and Silver datasets are summarized in table 1.

3.2 HARE-NER Training

The NER task is critical to HARE, as it identifies domain-specific entities from histopathology reports. We followed a systematic process to train and evaluate multiple BERT-based NER models optimized for the histopathology domain.

We experimented with several transformer-based architectures, including PathologyBERT and GatorTronS, which are pre-trained on clinical corpora, and BiomedBERT which was trained with PubMed articles as well as general domain models (BERT and DeBERTa models)2 (Devlin, 2018; He et al., 2021; Santos et al., 2023; Yang et al., 2022). PathologyBERT is the only publicly available model that is trained with pathology reports

General Domain	Model Size
BERT(Devlin, 2018)	110M 340M
DeBERTa-v3(He et al., 2021)	70M 435M
Biomedical Domain	**Model Size**
PathologyBERT(Santos et al., 2023)	110M
BiomedBERT(Tinn et al., 2021)	110M 340M
SapBERT(Liu et al., 2020)	110M
GatorTronS(Yang et al., 2022)	345M

Table 2: Models tested for fine-tuning. The models are sorted in the order of size.

Split	Reports	Sentences	Words
Golden-Train	834	7,159	85,639
Golden-Test	362	1,790	36,000
Silver-Train	1,270	1,684	33,581
Silver-Test	560	706	15,696

Table 3: Statistics of the train and test datasets used. The **Golden Dataset** includes manually annotated reports, while the **Silver Dataset (TCGA Reports)** includes automatically annotated reports. **Reports** represents the number of reports, **Sentences** the total sentences, and **Words** the total words

but for document classification specifically for breast cancer (Santos et al., 2023). SapBERT is also included as it was further trained with BiomedBERT model for entity alignment to Unified Medical Language System (UMLS), a detailed and widely used ontology (Liu et al., 2020; Tinn et al., 2021; National Library of Medicine (US), 2024). These models were fine-tuned using the annotated golden dataset to recognize histopathology-specific entities. The annotated reports input tokens were split into sentences, and any sentences longer than 512 tokens were split during preprocessing. All models were implemented using the HuggingFace Transformers library (Wolf, 2019). Training was conducted on NVIDIA A5000 GPU. The training used an AdamW optimizer with a learning rate of $5e^{-5}$ and a batch size of 4 for 2 epochs. We evaluated the models using standard metrics: accuracy and F1-score. The evaluation was conducted on a held-out test set, 30%, from the golden dataset. The train and test split is shown in table 3. The best-performing model was selected to annotate the silver dataset. To improve generalization, the final NER model, HARE-NER, was trained on a combined dataset of golden and silver annotations. This final model serves as the backbone for extracting entities in the HARE metric.

3.3 Design of the HARE Metric

The HARE metric evaluates the quality of generated histopathology reports by measuring the alignment of clinically relevant entities between reference and candidate reports.

3.3.1 Entity Extraction

Using the trained HARE-NER model, entities are extracted from both the reference and candidate reports. For each token, the model outputs a probability distribution over entity classes, and only entities with confidence scores above a threshold of 0.7 are retained. This threshold ensures that uncertain predictions are excluded from the evaluation.

3.3.2 Entity Embedding and Alignment

Extracted entities are embedded in a high-dimensional space using contextual representations from GatorTronS. The embeddings are further fine-tuned using a UMLS-based SapBERT approach, which ensures semantic alignment of similar entities (e.g., *lymphovascular invasion* and *vascular invasion*). Cosine similarity is computed between all pairs of entities from the reference and candidate reports. For each entity, we calculate the maximum cosine similarity with the entities in the other set.

3.3.3 Scoring

The HARE metric calculates precision, recall, and F1-scores based on matched entities. The formulas for precision and recall are as follows:

$$\text{Recall} = \frac{1}{|\mathbf{E}_{\text{ref}}|} \sum_{\mathbf{e}_{\text{ref}} \in \mathbf{E}_{\text{ref}}} \max_{\mathbf{e}_{\text{cand}} \in \mathbf{E}_{\text{cand}}} \text{S}(\mathbf{e}_{\text{ref}}, \mathbf{e}_{\text{cand}})$$

$$\text{Precision} = \frac{1}{|\mathbf{E}_{\text{cand}}|} \sum_{\mathbf{e}_{\text{cand}} \in \mathbf{E}_{\text{cand}}} \max_{\mathbf{e}_{\text{ref}} \in \mathbf{E}_{\text{ref}}} \text{S}(\mathbf{e}_{\text{cand}}, \mathbf{e}_{\text{ref}})$$

Where:

- $\mathbf{E}_{\text{ref}}$: Set of embeddings for reference entities.
- $\mathbf{E}_{\text{cand}}$: Set of embeddings for candidate entities.
- $\text{S}(\mathbf{u}, \mathbf{v})$: Cosine similarity between embeddings $\mathbf{u}$ and $\mathbf{v}$.

The HARE score is then calculated as the harmonic mean of precision and recall, also referred to as the F1-score:

$$\text{HARE Score (F1)} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This ensures that both precision and recall are considered equally, providing a balanced measure of the alignment between ground truth and predicted entities. A higher HARE score indicates better alignment, reflecting both accurate and comprehensive entity matching.

3.4 Validation of the HARE Metric

To validate HARE, we conducted an expert evaluation of machine-generated histopathology reports. We generated reports using GPT4O and GPT4O-mini using whole slide images (WSI) downloaded from TCGA (Hurst et al., 2024). Due to the volume of the images, we processed to lower resolution and resized the image to 1024 by 1024 pixels. In total, 75 randomly selected images were downloaded and used for generating reports. For each image, eight sets of reports were generated with different levels of specimen site information provided. In total, 600 reports were compared to the ground truth reports. Experts compared generated reports to ground truth (original) reports and assigned scores based on diagnostic accuracy and histopathological detail to ensure an objective evaluation of the model's performance in generating histopathology reports from WSI.

The following is the scoring system and the rationale for each score level:

- **Scores 5 (Perfect match with ground truth)**: This score is assigned to reports that are identical to the reference report in terms of both diagnostic accuracy and histopathological descriptions.
- **Scores 4 (Perfect match diagnosis with at least one wrong description)**: This score is assigned to reports that correctly identify the diagnosis, but contain at least one minor error in histopathological descriptions. These errors may involve inaccurate terminology, missing morphological features. Although these reports provide a reliable diagnosis, an incomplete or incorrect description reduces their overall quality.
- **Scores 3 (Correct diagnosis)**: This score is assigned to reports that accurately determine the correct diagnosis but do not provide any of the detailed histopathological descriptions in the ground truth.
- **Scores 2 (Broadly correct diagnosis)**: This score is assigned when reports correctly identify the general disease category but do not specify the exact diagnosis. For example, a report may correctly classify a tumor as malignant but does not differentiate between specific subtypes. These reports provide a useful but incomplete diagnosis, which limits their clinical applicability.
- **Scores 1 (Incorrect diagnosis with a few histopathological descriptions match with ground truth)**: This score is assigned when the report fails to provide the correct diagnosis but includes practical histopathological descriptions that align with the reference report. While some microscopic features are correctly described, the overall diagnostic conclusion is incorrect, greatly reducing the clinical reliability and utility of the report.
- **Scores 0 (Incorrect diagnosis with no histopathological description matches with ground truth)**: This score is assigned to reports that provide neither a correct diagnosis nor any histopathological descriptions that align with the ground truth. These reports fail to recognize key pathological features and do not contribute to an accurate clinical assessment, making them completely unreliable.

3.4.1 Metric Comparison

HARE scores were compared to expert scores using Pearson's correlation coefficient, Spearman's correlation coefficient, and Kendall's τ. Additionally, we benchmarked the metric against traditional lexical metrics (BLEU, ROUGE, METEOR, BERTScore) and radiology-specific metrics (RadGraph-XL, RaTEScore) (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Zhang et al., 2019; Delbrouck et al., 2024; Zhao et al., 2024).

4 Results and Discussion

4.1 GatorTronS Achieves Superior Performance in Named Entity Recognition

Our experiments demonstrated that GatorTronS outperforms other models, both general-purpose and biomedical, in extracting entities from histopathology reports. As shown in Table 4, GatorTronS achieved the highest accuracy (0.967) and F1-score (0.795), surpassing the next-best model, DeBERTa-v3-large (F1 = 0.788). This result underscores the efficacy of GatorTronS in ad-

Model	Accuracy	F1
DeBERTa-v3-xsmall	0.952	0.662
BERT-base	0.962	0.751
SapBERT	0.961	0.755
PathologyBERT	0.963	0.755
BiomedBERT-base	0.962	0.759
BERT-large	0.963	0.763
BiomedBERT-large	0.965	0.780
DeBERTa-v3-large	0.965	0.788
GatorTronS	**0.967**	**0.795**

Table 4: Model selection results based on evaluation accuracy and F1-score on the golden test set. Models are sorted by F1-score.

dressing the complexities inherent to histopathology text. Its extensive pre-training on large-scale synthetic clinical corpora provides it with a comprehensive understanding of domain-specific language, abbreviations, and nuanced terminology. This ability is particularly critical in histopathology, where specialized expressions describing tissue morphology and disease subtypes are prevalent. GatorTronS model even outperformed DeBERTa-large model which was approximately 90M parameters larger. A similar sized biomedical model, BiomedBERT-large, was also better than BERT-large model. Moreover, training GatorTronS on both highly curated "golden" annotations and automatically generated "silver" annotations broadened its exposure to diverse reporting styles, improving its generalizability and resulting in an F1-score increase to 0.812 5.

Model	Accuracy	F1
DeBERTa-v3-xsmall	0.957	0.716
BERT-large	0.960	0.733
SapBERT	0.965	0.774
PathologyBERT	0.966	0.779
BERT-base	0.965	0.780
BiomedBERT-base	0.966	0.784
DeBERTa-v3-large	0.967	0.792
BiomedBERT-large	0.968	0.801
GatorTronS	**0.971**	**0.812**

Table 5: Merged results based on evaluation accuracy and F1-score on the merged dataset. Models are sorted by F1-score.

An additional factor contributing to GatorTronS's superior performance is its model size. As the largest model among the biomedical models tested, GatorTronS benefits from greater representational capacity, enabling it to capture complex relationships in text more effectively.

While smaller biomedical models such as PathologyBERT and BiomedBERT-base demonstrate domain adaptation benefits and outperform general-purpose models of comparable size, such as DeBERTa-xsmall and BERT-base, they fall short when compared to larger general-purpose models like BERT-large and DeBERTa-v3-large. This finding emphasizes the importance of scaling model size as well as domain adaptation to achieve state-of-the-art performance in specialized domains like histopathology.

4.2 Majority of Generated Reports Lack Clinical Alignment

Despite advances in text generation methods, expert evaluations reveal a significant misalignment between system-generated reports and clinical requirements. As shown in Table 6, 369 out of 600 generated reports (61.5%) received a score of 0 and 71 reports received a score of 1 (11.8%), indicating 73.3% of the reports had an incorrect diagnosis. Only eight reports attained a score of 4, while none achieved the perfect score of 5. Scores with partially correct diagnosis, broadly correct diagnosis, and correct diagnosis (Score 2, 3, and 4) accounted for 160 reports (26.7%). When we compared the HARE and other scores to expert scores, we excluded reports with 0 scores to have more balanced representation of the scores. These findings highlight a significant limitation in the diagnostic accuracy of the generative model utilized, with a substantial proportion of reports failing to predict reliable pathological interpretations.

The high percentage of incorrect diagnoses and the lack of accurate microscopic descriptions can be attributed to several factors. One major limi-

Score	Count
0	369
1	71
2	90
3	62
4	8
5	0

Table 6: Distribution of expert evaluation scores for generated histopathology reports. Scores represent the degree of alignment with the reference reports, with higher scores indicating better alignment.

tation can be the use of a single, low-resolution WSI, which could restrict the model's ability to discern detailed morphological features essential for histopathological evaluation. Histopathologists analyze WSIs at multiple magnification levels (low-power magnification for architectural patterns, high-power for cellular details such as nuclear atypia, and mitotic figures), which is crucial to distinguish benign from malignant. This limitation can hinder the model's capacity to generate precise microscopic descriptions and accurately differentiate pathological entities. Furthermore, only one WSI was provided per case, whilst in most cases multiple WSI were utilized as part of the diagnostic process. Finally, critical contextual information (e.g., clinical history or anatomical site information) was not provided all the time.

4.3 HARE Outperforms Existing Metrics in Capturing Clinical Relevance

Method	r	ρ	τ
ROUGE-L	0.048	0.030	0.025
BLEU	0.078	0.106	0.099
METEOR	0.265	179	0.136
BERTScore	0.203	0.180	0.141
RaTEScore	0.372	0.350	0.276
RadGraph-XL	0.427	0.425	0.351
HARE (Ours)	**0.488**	**0.473**	**0.417**

Table 7: Comparison of evaluation methods based on Pearson correlation (r) Spearman (ρ) and Kendall's τ. Methods are sorted by Kendall's τ

As shown in Table 7, HARE achieved the highest Pearson correlation ($r = 0.488$), Spearman correlation ($\rho = 0.473$) and Kendall τ ($\tau = 0.417$) with expert scores. These results surpass those of RadGraph-XL, the second-best metric, which achieved $r = 0.427$, $\rho = 0.425$ and $\tau = 0.640$. Lexical metrics such as ROUGE-L ($r = 0.048$, $\rho = 0.030$, $\tau = 0.025$) and BLEU ($r = 0.078$, $\rho = 0.106$, $\tau = 0.099$) performed poorly, further underscoring their inability to evaluate clinically relevant content.

HARE's effectiveness originates from its focus on histopathology entity-level alignment, which ensures that key clinical features, such as pathological diagnosis, are appropriately prioritized. Unlike traditional lexical metrics, HARE incorporates semantic similarity measures tailored to pathology-specific terminology by incorporating descriptor and modifier entities, making it robust to linguistic variations. By capturing both semantic and clinical correctness, HARE offers a more accurate and reliable evaluation of generated histopathology reports.

The implications of HARE's performance are significant. Its strong correlation with expert evaluations indicates that it is a reliable proxy for clinical relevance and accuracy of the generated reports. As such, HARE can guide iterative improvements in report generation models, ensuring that future systems better align with clinical requirements. However, a Pearson correlation of 0.497 suggests that there is still room for improvement in evaluation metrics. Future work could explore larger annotated datasets sourced from multiple hospitals to incorporate more diverse writing styles, pathological entities and features, with more annotation of a greater variety of pathological entities and histomorphological features.

5 Conclusion

In this work, we proposed HARE, a novel entity-centric evaluation metric specifically designed to assess the clinical quality of machine-generated histopathology reports. HARE addresses the critical gap in domain-specific evaluation by prioritizing clinical relevance over traditional lexical overlaps. Through a combination of golden and silver annotated datasets and leveraging the powerful GatorTronS model for named entity recognition, HARE effectively aligns with expert evaluations, outperforming existing metrics such as BLEU, ROUGE, and RaTEScore.

Our findings reveal that even the proprietary multimodal large language models, such as GPT4O, struggle to produce clinically accurate histopathology reports. Although we have not tested a comprehensive list of models trained for histopathology reports such as HistGen and WsiCaption, HARE can be a robust framework for evaluating these models (Guo et al., 2024; Chen et al., 2024). HARE's superior performance underscores the importance of domain-specific evaluation metrics in bridging the gap between automated report generation and clinical expectations. By making HARE publicly available, along with the silver-standard annotations and NER model, we aim to facilitate advancements in both report generation and evaluation methodologies in histopathology and related fields.

Limitation

While HARE demonstrates strong alignment with expert assessments, several limitations remain. The reliance on silver-standard annotations in training may introduce noise, potentially impacting the generalization of the HARE-NER model in complex or unseen contexts. Future efforts could focus on improving the quality of silver annotations through semi-supervised or human-in-the-loop approaches.

The scope of this study is also limited to evaluating the generated text at an entity level. Future work could incorporate relation extraction and cross-entity consistency checks to better capture higher-order clinical reasoning in generated histopathology reports. Finally, while HARE aligns well with expert scores, its correlation with clinical outcomes remains unexplored, which could be a key area of future investigation.

Broader Impacts and Ethics Statement

All histopathology reports used in this study were de-identified to protect patient privacy and ensure compliance with ethical and legal standards. No personally identifiable information (PII) was used in the development of the HARE framework. Our work does not raise any major ethical concerns. HARE is designed for evaluation and research purposes only and is not intended for direct use in clinical decision-making.

While HARE provides a reliable metric for evaluating the quality of generated histopathology reports, it does not address potential biases or hallucinations in the underlying text generation models. Therefore, any use of automated text generation systems in clinical workflows should include rigorous human oversight to mitigate risks, such as incorrect diagnoses or misleading conclusions.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Pingyi Chen, Honglin Li, Chenglu Zhu, Sunyi Zheng, Zhongyi Shui, and Lin Yang. 2024. Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–556. Springer.

Pingyi Chen, Honglin Li, Chenglu Zhu, Sunyi Zheng, and Lin Yang. 2023. Mi-gen: Multiple instance generation of pathology reports for gigapixel whole-slide images. *arXiv preprint arXiv:2311.16480*.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. 2024. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12902–12915.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc.*, 23(2):304–310.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. 2024. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199. Springer.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. 2023. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Xinyao Liu, Junchang Xin, Qi Shen, Zhihong Huang, and Zhiqiong Wang. 2025. Automatic medical report generation based on deep learning: A state of the art survey. *Computerized Medical Imaging and Graphics*, page 102486.

National Library of Medicine (US). 2024. UMLS Knowledge Sources. Release 2024AB. Bethesda (MD): National Library of Medicine (US); 2024 November 6 [cited 2025 Jan 21]. Available from: http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *arXiv preprint arXiv:2307.09758*.

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*, pages 311–318.

Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. 2012. Semantic cosine similarity. In *The*

7th international student conference on advanced science and technology ICAST, volume 4, page 1. University of Seoul South Korea.

Thiago Santos, Amara Tariq, Susmita Das, Kavyasree Vayalpati, Geoffrey H Smith, et al. 2023. Pathologybert-pre-trained vs. a new transformer language model for pathology domain. In *AMIA annual symposium proceedings*, volume 2022, page 962.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Jing Wei Tan, SeungKyu Kim, Eunsu Kim, Sung Hak Lee, Sangjeong Ahn, and Won-Ki Jeong. 2024. Clinical-grade multi-organ pathology report generation for multi-scale whole slide images via a semantically guided medical text foundation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–35. Springer.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jinge Wu, Yunsoo Kim, Daqian Shi, David Cliffton, Fenglin Liu, and Honghan Wu. 2024. Slava-cxr: Small language and vision assistant for chest x-ray report automation. *arXiv preprint arXiv:2409.13321*.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, et al. 2024. Overview of the first shared task on clinical text generation: Rrg24 and" discharge me!". *arXiv preprint arXiv:2409.16603*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, et al. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Ken G Zeng, Tarun Dutt, Jan Witowski, GV Kranthi Kiran, Frank Yeung, Michelle Kim, Jesi Kim, Mitchell Pleasure, Christopher Moczulski, L Julian Lechuga Lopez, et al. 2023. Improving information extraction from pathology reports using named entity recognition. *Research Square*.

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

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. *medRxiv*, pages 2024–06.

Appendix

A Word cloud representations of radiology and histopathology reports

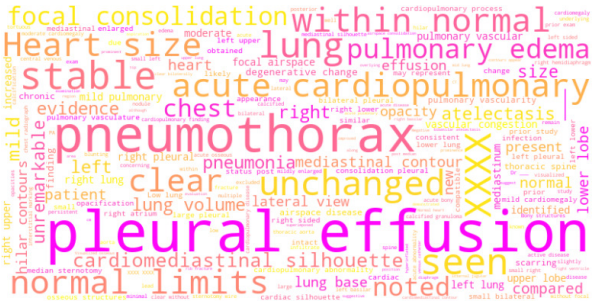

Figure 3: Word clouds of radiology reports. The radiology reports are 1,000 randomly sampled reports from MIMIC-CXR dataset and IU-X-ray dataset (Johnson et al., 2019; Demner-Fushman et al., 2016). The size of each word represents its relative frequency in the corresponding category.

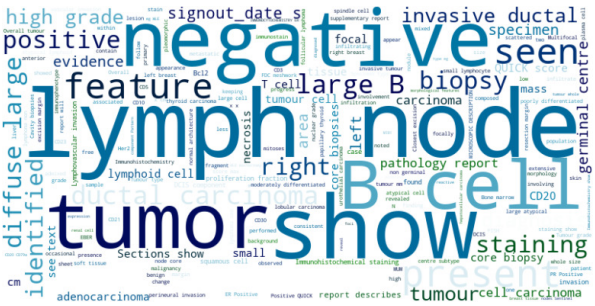

Figure 4: Word clouds of histopathology reports. The histopathology reports are 1,000 randomly sampled reports from our golden and silver dataset. The size of each word represents its relative frequency in the corresponding category.

These visualizations provide insight into the linguistic differences between radiology and histopathology reports, highlighting the specialized vocabulary and diagnostic focus within each domain. Larger words represent higher relative frequency. The word cloud visualization for radiology reports highlights key terms such as "pleural effusion", "pneumothorax", "cardiopulmonary" and "atelectasis", indicating these are more common findings and diagnostic terminology used in radiology (see Figure 3). Figure 4 illustrates a word cloud generated from 1,000 randomly sampled histopathology reports from our golden and silver dataset. Frequent occurring terms such as "tumor", "lymph node", "B cell", "negative", "biopsy", and "staining", reflect key features and diagnostic language used in histopathology reports. Compared to radiology reports, histopathology reports exhibit more granular terminology related to cellular morphology and pathology-specific descriptors.