HARE: an entity and relation centric evaluation framework for histopathology reports

Yunsoo Kim[1], **Michal W. S. Ong**[1], **Alex Shavick**[3], **Honghan Wu**[1,2], **Adam P. Levine**[1,4]
[1]University College London, London, UK
[2]University of Glasgow, Glasgow, UK
[3]Human Technopole, Milan, Italy
[4]University College London Hospitals NHS Foundation Trust, London, UK
yunsoo.kim.23@ucl.ac.uk

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 and relation centric framework, composed of a benchmark dataset, a named entity recognition (NER) model, a relation extraction (RE) model, and a novel metric, which prioritizes clinically relevant content by aligning critical histopathology entities and relations between reference and generated reports. To develop the HARE benchmark, we annotated 813 de-identified clinical diagnostic histopathology reports and 652 histopathology reports from The Cancer Genome Atlas (TCGA) with domain-specific entities and relations. We fine-tuned GatorTronS, a domain-adapted language model to develop HARE-NER and HARE-RE which achieved the highest overall F1-score (0.915) among the tested models. The proposed HARE metric outperformed traditional metrics including ROUGE and Meteor, as well as radiology metrics such as RadGraph-XL, with the highest correlation and the best regression to expert evaluations (higher than the second best method, GREEN, a large language model based radiology report evaluator, by Pearson $r = 0.168$, Spearman $\rho = 0.161$, Kendall $\tau = 0.123$, $R^2 = 0.176$, $RMSE = 0.018$). We release HARE, datasets, and the models at https://github.com/knowlab/HARE to foster advancements in histopathology report generation, providing a robust framework for improving the quality of 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. Such reports play 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 analyses (e.g. immunohistochemistry (IHC) markers, special stains, or in situ hybridization (ISH)), and a final diagnosis/conclusion.

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 not covered by radiology word embeddings, making radiology report

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9218–9233
November 4-9, 2025

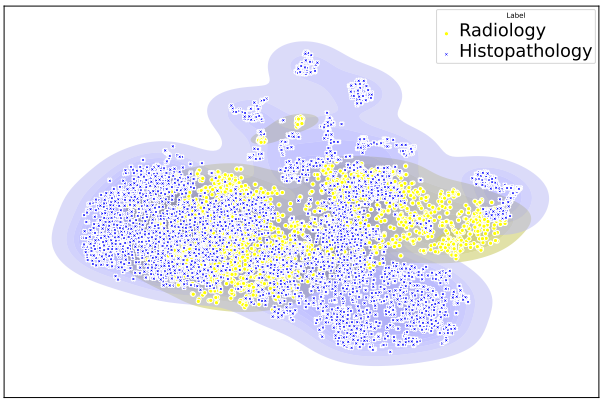

Figure 1: Scatter and density plot of word embeddings for radiology and histopathology reports. The radiology reports are 1,000 randomly sampled reports from MIMIC-CXR dataset and IU-X-ray dataset (Johnson et al., 2019; Demner-Fushman et al., 2016). The histopathology reports are 1,000 randomly sampled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

evaluation metrics unsuitable for histopathology reports. Conventional lexical evaluation metrics such as METEOR and BERTScore as well as clinical 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) and relation extraction (RE), which limits the ability to train robust models tailored to the complexities of histopathological language. There is only one NER model and dataset for pathology reports; however, these are not publicly available (Zeng et al., 2023). This gap underscores the need for an entity and relation 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 HARE score is demonstrated. HARE captures domain-specific entities (e.g., anatomical sites, IHC markers, descriptor and final diagnosis) and relationships between the entities 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 1,465 real-world diagnostic histopathology reports sourced from a large teaching hospital and from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015).

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 compared with multiple other available metrics. By releasing both our annotated dataset and the final trained models (which we call HARE-NER and HARE-RE), 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 HARE metric.
3. **HARE-NER and HARE-RE:** We develop a NER model and a RE model specialized in histopathology, capable of recognizing and relating 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 and RE model.
4. **Open Source:** We will release (1) the annotated dataset, (2) the final trained NER model, RE 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.

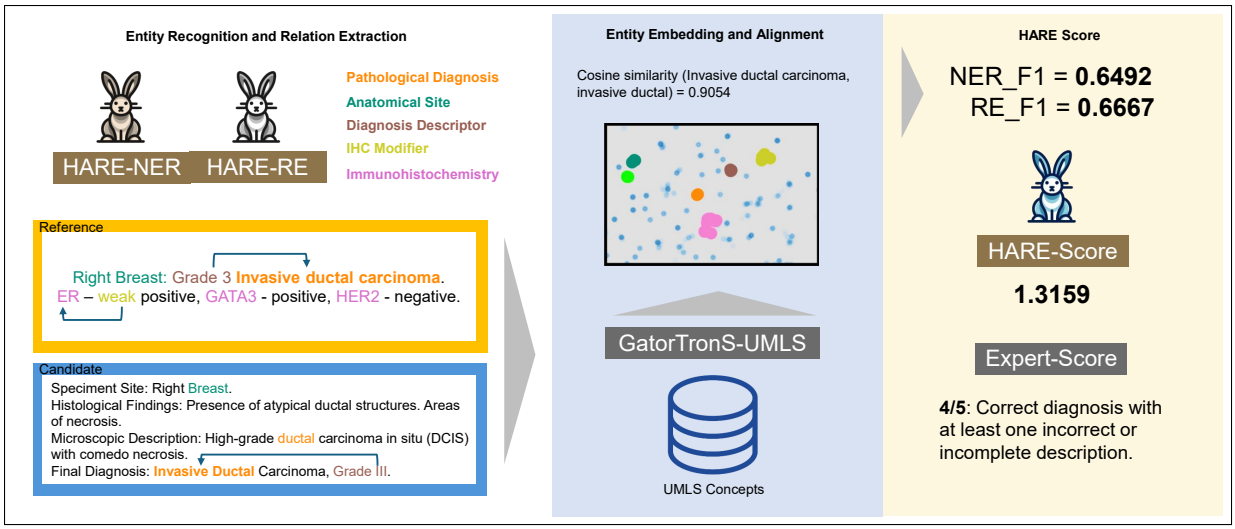

Figure 2: Illustration of the process of computing the HARE score, a novel entity and relation centric metric for evaluating histopathology report generation.

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 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, NER model and RE 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: reports collected from a hospital and publicly available reports from TCGA.

3.1.1 Hospital Dataset

We collected 813 fully de-identified/anonymized histopathology reports from the pathology department of a large teaching hospital. We ensured that the reports were free of any identifiable data through the use of Stanford AIMI's deidentification model and by manual review and redaction of identifiers by three histopathologists (Chambon et al., 2022). The reports were from cases across a range of tissue types and diagnoses with a partial focus on cases with lymphoma, breast cancer and cases in which several IHC markers were utilized as part of the diagnostic process. The reports were annotated by a junior histopathologist (with input from a senior histopathologist for clarification of challenging cases) using the Inception annotation tool (Klie et al., 2018). The annotations focused on histopathology-specific entities, including:

- **Anatomical Site:** Entities describing specific tissue regions or locations, such as *breast, lung, kidney, lymph node etc.*
- **Immunohistochemistry (IHC) 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 e.g., "raises the possibility of".
- **IHC Modifier:** Used to modify immunohistochemical annotations, e.g., "patchy" or "strong".

The relationships annotated were:

- **IHC Markers - IHC Modifier**
- **Diagnosis - Diagnosis Descriptor**

Type	Hospital	TCGA
IHC Markers	6,628	119
IHC Modifier	1,339	173
Pathological Diagnosis	885	882
Anatomical Site	747	794
Diagnosis Descriptor	247	475
Relations	1,745	653

Table 1: Entity and Relation annotation statistics for the Hospital and TCGA datasets.

3.1.2 TCGA Dataset

To increase diversity, we further annotated 652 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). The annotation was done in the same manner as the Hospital dataset but using the label studio as the annotation tool (Tkachenko et al., 2020-2025). We extracted sentences with histopathological descriptions, specifically IHC markers and final diagnosis characteristics. The breakdown of the number of annotations for the Hospital and TCGA datasets are summarized in table 1.

3.1.3 Annotator description

All annotations were performed by practicing physicians with formal training and accreditation in histopathology (at different stages of progression through the national pathology examination board). The two junior pathologists had 5 and 7 years of histopathology experience, alongside 13

and 12 years of clinical medical practice, respectively. The senior pathologist was board-certified, with 7 years of histopathology experience and 18 years of medical practice. Before initiating the annotation process, all annotators met to establish and agree on a standardized annotation protocol.

3.2 HARE-NER and HARE-RE Training

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. Models with two sizes indicate different pretrained model variants (e.g., BERT-base vs. BERT-large).

As shown in Table 2, we experimented with several transformer-based architectures, including PathologyBERT (Santos et al., 2023) and GatorTronS (Yang et al., 2022), which are pretrained on clinical corpora, and BiomedBERT (Tinn et al., 2021) which was trained with PubMed articles as well as general domain models (BERT (Devlin, 2018) and DeBERTa (He et al., 2021) models). PathologyBERT is the only publicly available model that is trained with pathology reports for document classification specifically for breast cancer. SapBERT (Liu et al., 2020) 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 (National Library of Medicine (US), 2024).

These models were fine-tuned using our annotated dataset for both NER and RE. For the NER task, we trained a token classification model based on the pre-trained encoder to recognize histopathology-specific entities. For the RE task, we trained a sequence classification model with entity markers (E1 and E2) based on the same pretrained encoders to capture relationships between extracted entities. E1 and E2 are placeholder tokens used to mark the two extracted entities involved in a candidate relation pair (they are abbreviations for entity 1 and entity 2, respectively). These are passed to the relation extraction model, which classifies the relation type via a sequence classification objective.

The annotated reports were split into sentences, and any sentence longer than 512 tokens was split during preprocessing to accommodate model input constraints. All models were implemented using the HuggingFace Transformers library (Wolf, 2019). Training was conducted on an NVIDIA A5000 GPU. For both NER and RE, we used an AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 4 for 2 epochs. Evaluation was performed using standard metrics, F1-score, for both tasks, with 10% of the data as a hold-out test set.

For relation extraction, the dataset required explicit construction of entity pairs. All positive samples (annotated entity relations) and an equal number of randomly sampled negative pairs were used to construct the training split. For the test split, three times as many negative samples as positive samples were sampled to ensure robust evaluation. The relation extraction model was evaluated using gold-standard entities, not predicted ones.

Details of the train and test splits are shown in Table 3. The best-performing models for NER and RE were selected as the backbone for extracting entities and relationships within the HARE metric.

Split	Samples	Tokens
NER-Train	2,181	127,553
NER-Test	243	13,855
Relation-Train	5,014	311,058
Relation-Test	1,068	66,769

Table 3: Statistics of the train and test datasets used for NER and RE tasks. **Samples** represents the number of samples and **Tokens** the total tokens (word-piece).

3.3 Design of the HARE Metric

The HARE metric evaluates the quality of generated histopathology reports by assessing both the alignment of clinically relevant entities and the relationships between them in the reference and candidate reports.

3.3.1 Entity and Relation Extraction

Entities are extracted from both reference and candidate reports using the trained HARE-NER model. For each token, the model outputs a probability distribution over entity classes; only entities with confidence scores above a threshold of 0.7 are retained, ensuring that uncertain predictions are ex-

cluded. Relations between recognized entities are then identified using the trained HARE-RE model, which predicts relation types for all candidate entity pairs. The same confidence threshold is applied to relation predictions to retain only high-confidence relations.

3.3.2 Entity Embedding and Alignment

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

3.3.3 Scoring

The HARE metric reports both entity- and relation-level alignment between candidate and reference reports. For entities, precision, recall, and F1-score are computed as follows:

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

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

where $\mathbf{E}_\text{ref}$ and $\mathbf{E}_\text{cand}$ are the sets of embeddings for reference and candidate entities, and $\text{S}(\mathbf{u}, \mathbf{v})$ is the cosine similarity between embeddings $\mathbf{u}$ and $\mathbf{v}$.

The F1-score for NER is then calculated as the harmonic mean of precision and recall:

$$\text{F1}_\text{e} = 2 \cdot \frac{\text{Precision}_\text{e} \cdot \text{Recall}_\text{e}}{\text{Precision}_\text{e} + \text{Recall}_\text{e}}$$

Relation extraction performance is quantified using the standard F1-score, computed by comparing the set of extracted relations (entity pairs and their predicted relation types) in the candidate report to those in the reference. Precision and recall are calculated based on the overlap of predicted and reference relations, and the relation F1-score is reported as:

$$\text{F1}_\text{r} = 2 \cdot \frac{\text{Precision}_\text{r} \cdot \text{Recall}_\text{r}}{\text{Precision}_\text{r} + \text{Recall}_\text{r}}$$

To obtain a comprehensive assessment, the final HARE score is defined as the sum of the entity and relation F1-scores:

$$\text{HARE Score} = \text{F1}_\text{e} + \text{F1}_\text{r}$$

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 GPT-4o and GPT-4o-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 or 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 some of the histopathological descriptions matching the ground truth)**: This score is assigned when the report fails to provide the correct diagnosis but includes 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 descriptions matching 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.

Histopathology reports are inherently complex and exhibit significant variability in writing styles across institutions and individual histopathologists, particularly in the microscopic description section. This variability introduces heterogeneity in report structure, making it challenging for models to learn consistent diagnostic patterns. Despite these differences, histopathologists generally reach a consensus on the final diagnosis, which carries the most clinical significance. Therefore, our evaluation places greater emphasis on the model's ability to generate correct diagnoses rather than the accuracy of microscopic descriptions.

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. We provide p-values for each correlation value. Additionally, we benchmarked the metric against traditional lexical metrics (BLEU, ROUGE, METEOR, BERTScore) and radiology-specific metrics (RadGraph-XL, RaTEScore, GREEN) (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Zhang et al., 2019; Delbrouck et al., 2024; Zhao et al., 2024; Ostmeier et al., 2024). We used only the single overall score for metrics such as BLEU, RadGraph-XL, and GREEN to compare our method with a state-of-the-art LLM-based evaluator for radiology reports. We also used GPT-4.1 as a judge to give a score of 0-5 based on the expert evaluation scheme for the candidate report based on the ground truth report (Hurst et al., 2024). The prompt used for this analysis can be found at the appendix Figure 7. We also performed regression analysis and provided R^2 and RMSE values to assess the predictive utility of each metric against expert scores. For all the metric comparison, we normalized the automated metric scores to a 0-1 scale and expert evaluation scores (originally 0–5) also normalized to 0-1.

4 Results and Discussion

4.1 Model Selection: GatorTronS

Model	NER	RE	Overall
PathologyBERT	0.771	0.798	0.785
BERT-base	0.833	0.798	0.816
BERT-large	0.825	0.798	0.811
DeBERTa-large	0.841	0.798	0.820
BiomedBERT-large	0.843	0.798	0.820
DeBERTa-xsmall	0.794	0.962	0.878
SapBERT	0.835	0.970	0.903
BiomedBERT-base	0.844	0.962	0.903
GatorTronS	**0.854**	**0.977**	**0.915**

Table 4: Model selection results based on NER and RE F1-scores on the test set. Models are sorted by Overall F1-score.

Our experiments demonstrated that GatorTronS outperforms other models, both general-purpose and biomedical, in extracting entities and relations from histopathology reports. As shown in Table 4, GatorTronS achieved the highest overall score (0.915) with NER F1 (0.854) and RE F1-score (0.977), surpassing the next-best model, BiomedBERT-base (Overall F1 = 0.903, NER F1 = 0.844, RE F1 = 0.962). Notably, models with an RE F1 of 0.798 failed to identify any relations for all test inputs, highlighting a limitation of these architectures in this domain.

This result underscores the efficacy of GatorTronS in addressing the complexities inherent in histopathology text. Its extensive pretraining 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

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

Table 5: 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.

expressions describing tissue morphology and disease subtypes are prevalent.

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.

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 5, 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. Reports often lacked diagnostic conclusions or included incorrect terminology, while others failed to capture essential histological findings. The scarcity of high-quality outputs underscores the challenge in generating nuanced and diagnostically accurate narratives. Errors in final diagnosis are particularly concerning as they can have significant clinical implications. 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 limitation 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 make an accurate pathological diagnosis. 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 WSIs were utilized as part of the actual diagnostic process to generate the ground truth report. Finally, critical contextual information (e.g., clinical history or anatomical site information) was not provided all the time. Notably, a subset of reports that included primary specimen site information demonstrated a slight improvement, achieving higher scores. This suggests that while the performance of current multimodal LLMs such as GPT-4o, is limited, when provided with additional clinical and anatomical context, the model's diagnostic reliability can sometimes be acceptable.

4.3 HARE Outperforms Existing Metrics in Capturing Clinical Relevance

Table 6 summarizes the performance of all evaluation metrics against expert pathologist scores using multiple statistical measures. HARE achieved the highest Pearson correlation (0.606), Spearman correlation (0.643) and Kendall τ (0.533), all with strong statistical significance. HARE also demonstrated the highest coefficient of determination ($R^2 = 0.368$) and the lowest root mean squared error (RMSE = 0.134), indicating both high alignment and predictive accuracy with respect to expert score.

These results surpass those of GREEN, the next-best metric, which leverages a large language model (RadLlama2-7b) as an evaluator. Moreover, HARE is significantly more computationally efficient: on 600 candidate reports, GREEN required 2 hours and 2 minutes for evaluation, while HARE completed the same analysis in 192 seconds on an A5000 24GB GPU. This efficiency, combined with robust performance, underscores HARE's practical viability and interpretability as an evaluation metric

Method	r	r p-val	ρ	ρ p-val	τ	τ p-val	R^2	RMSE
ROUGE-L	0.048	0.470	0.030	0.647	0.025	0.616	0.002	0.169
BLEU	0.077	0.241	0.106	0.108	0.099	0.107	0.006	0.168
GPT-4.1	0.177	0.007	0.173	0.008	0.146	0.008	0.031	0.166
BERTScore	0.203	0.002	0.180	0.006	0.141	0.005	0.041	0.165
METEOR	0.265	4.51e-05	0.179	0.006	0.136	0.007	0.070	0.163
RaTEScore	0.372	5.36e-09	0.350	4.81e-08	0.276	4.60e-08	0.138	0.157
RadGraph-XL	0.427	1.22e-11	0.425	1.43e-11	0.351	8.51e-11	0.182	0.153
GREEN	0.438	2.90e-12	0.482	7.58e-15	0.410	2.18e-13	0.192	0.152
HARE (Ours)	**0.606**	**1.39e-24**	**0.643**	**2.62e-28**	**0.533**	**1.51e-24**	**0.368**	**0.134**

Table 6: Comparison of evaluation methods based on Pearson correlation (r), Spearman (ρ), and Kendall's τ with their p-values, and regression performance (R^2 and RMSE). Methods are sorted by Pearson correlation r.

for histopathology report generation.

HARE significantly outperforms GPT-4.1 used as a judge in correlation with expert ratings. While GPT-4.1 is not appropriate for real clinical evaluation pipelines due to its proprietary nature and privacy constraints, our result confirms the superior performance of a domain specific report quality evaluator over an LLM based evaluator.

In contrast, although they are widely used in histopathology report evaluation, lexical metrics such as ROUGE-L ($r = 0.048$, $\rho = 0.030$, $\tau = 0.025$) and BLEU ($r = 0.078$, $\rho = 0.106$, $\tau = 0.099$) showed minimal correlation and high RMSE, further underscoring their inability to assess clinically relevant content in histopathology.

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. HARE can guide iterative improvements in report generation models, ensuring that future systems better align with clinical requirements.

5 Conclusion

In this work, we proposed HARE, a novel entity and relation 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. HARE effectively aligns with expert evaluations, outperforming existing metrics such as ROUGE and RaTEScore.

Our findings reveal that even proprietary multimodal large language models, such as GPT-4o, 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). In the future, we will include histopathology reports specific models for creating the human evaluation dataset as well as extend HARE models to a joint NER+RE model to further improve the performance and utility.

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 annotations and models, we aim to facilitate advancements in both report generation and evaluation methodologies in histopathology and related fields.

Limitation

Scope of clinical entities and relations: The current implementation of HARE primarily addresses a set of core histopathology entities and relatively simple binary relations. More nuanced or higher-

order clinical relationships, as well as rare or emerging entity types, remain underrepresented. Expanding both the entity and relation taxonomies to better reflect the complexity of real-world histopathology reporting is an important and interesting future work we plan to explore.

Negation and uncertainty handling: While HARE captures explicit clinical entities, it does not yet explicitly handle negation or uncertainty (e.g., "no evidence of malignancy," "cannot rule out invasion"). These linguistic phenomena are important for accurate clinical interpretation and could be incorporated into future extensions of the metric.

Breadth of expert evaluation models: For the generation of reports used in the expert evaluation, we utilized only two closed source models, GPT-4o and GPT-4o-mini. As the primary scope of this work was the development of the evaluation metric, a broader evaluation across more generative models remains to be explored in future work.

Broader Impacts and Ethics Statement

Histopathology reports used in this work were provided via a study registered with, and approved by, the NHS Health Research Authority (references 293404 and 23/LO/0253). All histopathology reports were fully de-identified to protect patient privacy and ensure compliance with ethical standards. No personally identifiable information 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.

Acknowledgement

The authors kindly acknowledge funding from an NIHR Academic Clinical Lectureship (APL), The Jean Shanks Foundation/The Pathological Society of Great Britain & Ireland (APL and MWSO) and Rosetrees Trust (APL and YK).

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

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*.

Pierre J Chambon, Christopher Wu, Jackson M Steinkamp, Jason Adleberg, Tessa S Cook, and Curtis P Langlotz. 2022. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *Journal of the American Medical Informatics Association*. Ocac219.

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*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit

Delbrouck. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA. Association for Computational Linguistics.

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.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

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

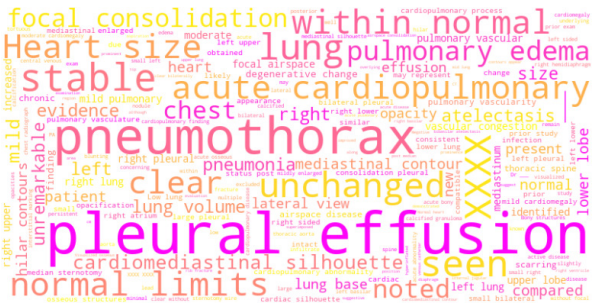

Figure 3: Word clouds of radiology reports. The radiology reports are 1,000 randomly sampled reports from MIMIC-CXR dataset and IU-X-ray dataset (Johnson et al., 2019; Demner-Fushman et al., 2016). The size of each word represents its relative frequency in the corresponding category.

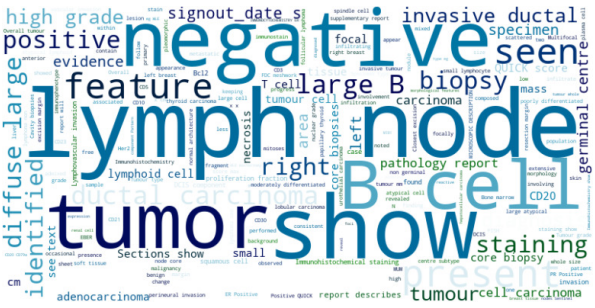

Figure 4: Word clouds of histopathology reports. The histopathology reports are 1,000 randomly sampled reports from our datasets. The size of each word represents its relative frequency in the corresponding category.

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 datasets. 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.

B Report Examples

We provide example annotated histopathology reports from both the Hospital and TCGA datasets. These examples illustrate not only the complexity and diversity of histopathology reporting, but also the breadth of clinically significant entities and inter-entity relationships captured by our annotation schema. Key entity types include pathological diagnosis, anatomical site, histological findings, immunohistochemistry markers, descriptors, and modifiers.

In addition to highlighting individual entities, these examples also depict the relationships between entities, such as associations between anatomical sites and diagnostic findings, or between immunohistochemistry results and corresponding pathological diagnoses. Modeling both entity-level information and their relationships is essential for accurately representing the clinical reasoning process in histopathology and for evaluating the fidelity of automated report generation.

Visualizing these examples demonstrates the level of annotation granularity and relational structure necessary for effective evaluation, and serves as a benchmark for downstream clinical NLP applications in entity recognition and relation extraction.

C Empirical anslysis of NER and RE errors

To empirically assess the impact of errors in NER and RE, we conducted an ablation study evaluating several variants of the HARE metric under different entity and relation confidence thresholding schemes. Specifically, we compared:

1. **HARE_ERROR**: HARE applied using only low-confidence (i.e., likely incorrect) NER and RE outputs by inverting the threshold
2. **HARE_No_Threshold**: HARE applied with no confidence threshold, including all predicted entities and relations
3. **HARE_0.7_Threshold**: our default approach, which applies a confidence threshold of 0.7 to retain only high-confidence entities and relations

Table 7 presents the results. The HARE_ERROR variant demonstrates very

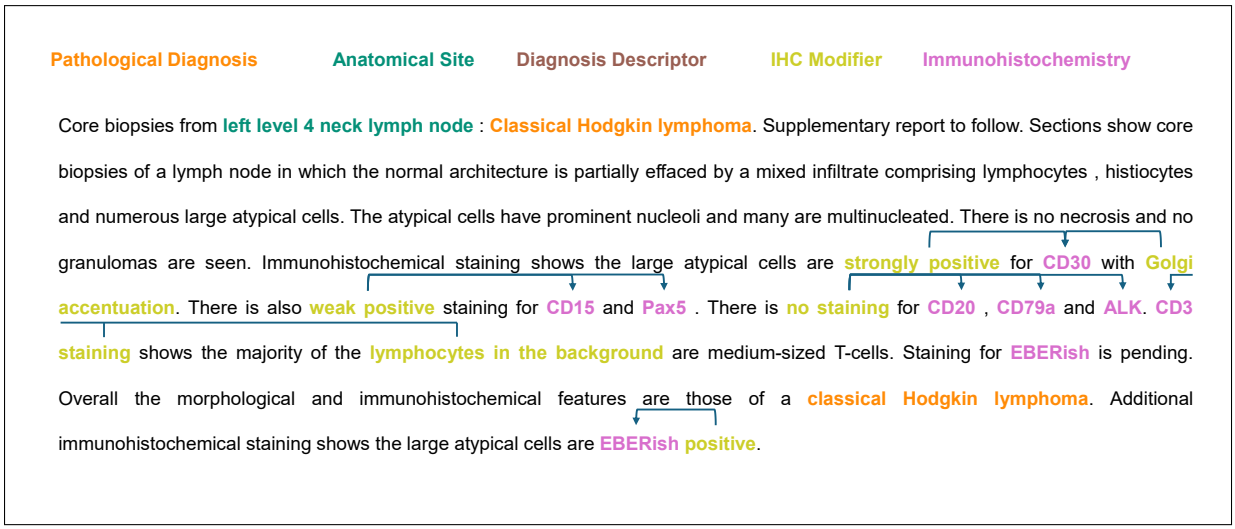

Figure 5: Example of an annotated histopathology report from the Hospital Dataset. The report details a diagnosis of classical Hodgkin lymphoma in a lymph node, with corresponding entity-level annotations highlighting pathological diagnosis, anatomical site, immunohistochemical findings, and key descriptors.

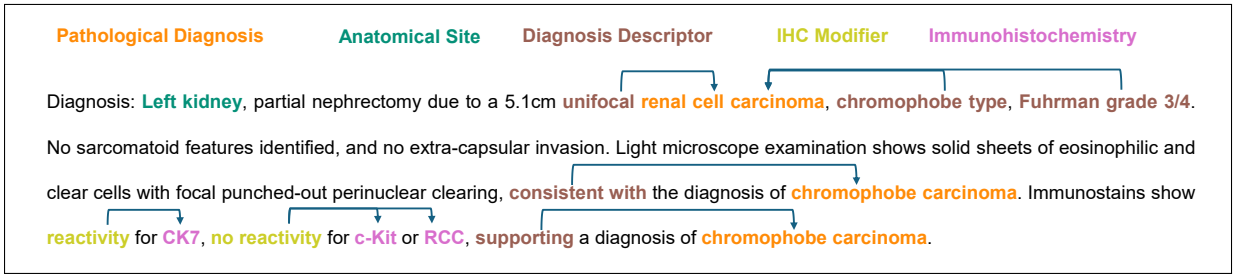

Figure 6: Example of an annotated histopathology report from the TCGA Dataset. The report presents a case of left kidney partial nephrectomy for chromophobe renal cell carcinoma, with entity annotations for pathological diagnosis, anatomical site, diagnostic descriptors, immunohistochemistry markers, and modifiers.

Method	r	r **p-val**	ρ	ρ **p-val**	τ	τ **p-val**	$\mathbf{R^2}$	**RMSE**
HARE_ERROR	0.026	0.693	-0.005	0.942	-0.005	0.934	0.001	0.169
HARE_No_Threshold	0.567	4.89e-21	0.629	7.86e-27	0.513	4.00e-24	0.321	0.139
HARE_0.7_Threshold	**0.606**	**1.39e-24**	**0.643**	**2.62e-28**	**0.533**	**1.51e-24**	**0.368**	**0.134**

Table 7: Comparison of evaluation methods based on Pearson correlation (r), Spearman (ρ), and Kendall's τ with their p-values, and regression performance (R^2 and RMSE). Methods are sorted by Pearson correlation r. HARE_ERROR is the one with inverted confidence threshold. HARE_No_Threshold is the one without threshold. HARE_0.7_Threshold is our method.

poor correlation with expert scores across all statistical measures, underscoring the critical importance of accurate entity recognition and relation extraction. Removing the threshold altogether (HARE_No_Threshold) moderately improves performance but still underperforms relative to our approach. The HARE_0.7_Threshold, our approach, achieves the highest correlation and lowest RMSE, validating our choice of threshold and the metric's design, which effectively mitigates the impact of noisy or uncertain predictions.

These findings highlight that HARE's strong correlation with expert assessments depends critically on accurate entity recognition and relation extraction. It also shows that the chosen confidence thresholding scheme is a key mechanism for maintaining robustness to NER and RE errors.

D GPT4.1 Prompt

We designed the prompt for GPT4.1 analysis for the expert evaluation of the machine-generated histopathology reports 7.

Prompt for GPT-4.1

Act as a histopathologist.
Review the following candidate report's similarity score to the ground truth report. Just report the numerical score.
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 or 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 some of the histopathological descriptions matching the 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 descriptions matching 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.

Ground Truth : {Ground Truth Report}
Candidate Report: {Candidate Report}

Figure 7: Prompt templates used for GPT-4.1 analysis.