Staged Alignment of Decoder Large Language Models for Oncology Tasks via Radiology-Pathology Note Pairing

Bo Liu1,3,4; Benedict Neo1,2; Yannet Interian2; Olivier Morin1; Steve E. Braunstein1; Janine Lupo3,4; Hui Lin1,3,5
1 Radiation Oncology, UCSF | 2 University of San Francisco | 3 Bioengineering, UCSF-UC Berkeley | 4 Radiology and Biomedical Imaging, UCSF | 5 Computational Precision Health, UCSF-UC Berkeley

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Purpose: General-purpose language models do not naturally link radiology and pathology reports that describe the same glioma. We developed a clinically aligned glioma embedding model (GEM) from paired reports.
Methods: We fine-tuned jina-embeddings-v3 (570M parameters) with Multiple-Negatives Ranking Loss using 97,398 radiology-pathology pairs from 4,250 patients. Cross-modal retrieval was tested in 850 held-out patients. Frozen embeddings were evaluated with patient-disjoint cross-validation for tumor type, MGMT methylation, and 1-year survival, against encoder baselines and zero-shot decoder LLMs.
Results: GEM improved retrieval Recall@1 from 6.3% to 35.6% and mean reciprocal rank from 12.6% to 46.9%. Tumor type accuracy/AUROC was 87.2 ± 1.3%/96.5 ± 0.8%; MGMT accuracy/AUROC was 70.5 ± 5.6%/77.0 ± 3.3%; and survival accuracy/AUROC was 78.6 ± 4.5%/76.4 ± 6.2%. GPT-4o showed poor AUROC for tumor type (35.1%) and survival (30.0%).
Conclusion: Radiology-pathology pairing yields a scalable, label-free signal for clinically aligned glioma embeddings.
Innovation/Impact: GEM uses electronic health record pairing structure to create a compact, reusable encoder for neuro-oncology decision support.

CONTACT

Bo Liu
UC San Francisco
Bo.liu3@ucsf.edu

INTRODUCTION

Glioma diagnosis and prognostication require clinicians to integrate MRI findings with histopathologic and molecular information (Ref. 1). Although these reports describe the same tumor, general-purpose language models typically treat them as unrelated text and may not capture their shared clinical meaning. We hypothesized that the natural radiology-pathology pairings already present in the electronic health record can provide a label-free signal for domain alignment. We therefore fine-tuned a compact text encoder on matched report pairs (Figure 2) to learn clinically aligned embeddings that transfer to diagnostic, molecular, and prognostic glioma tasks.

RESULTS

MNRL fine-tuning substantially improved cross-modal alignment: Recall@1 increased from 6.3% to 35.6%, Recall@10 from 24.3% to 71.1%, and MRR from 12.6% to 46.9%.
GEM achieved strong downstream performance:
Tumor type: 87.2 ± 1.3% accuracy; 96.5 ± 0.8% AUROC
MGMT methylation: 70.5 ± 5.6% accuracy; 77.0 ± 3.3% AUROC
1-year survival: 78.6 ± 4.5% accuracy; 76.4 ± 6.2% AUROC
GEM matched or exceeded strong encoder baselines and outperformed zero-shot decoder models in calibrated discrimination. GPT-4o showed near-chance AUROC for tumor type (35.1%) and survival (30.0%) despite comparable accuracy on selected tasks (Figure 4; Ref. 3).

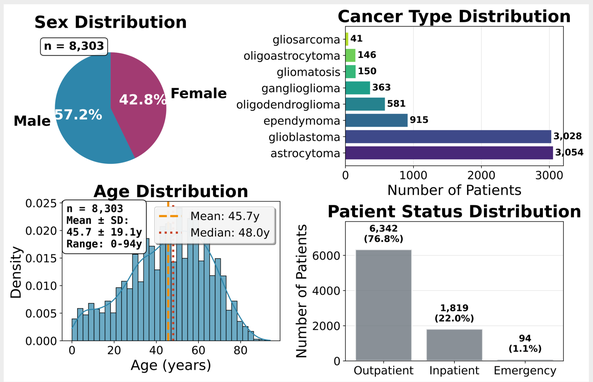

Figure 1. Demographic characteristics of the matched radiology-pathology cohort.

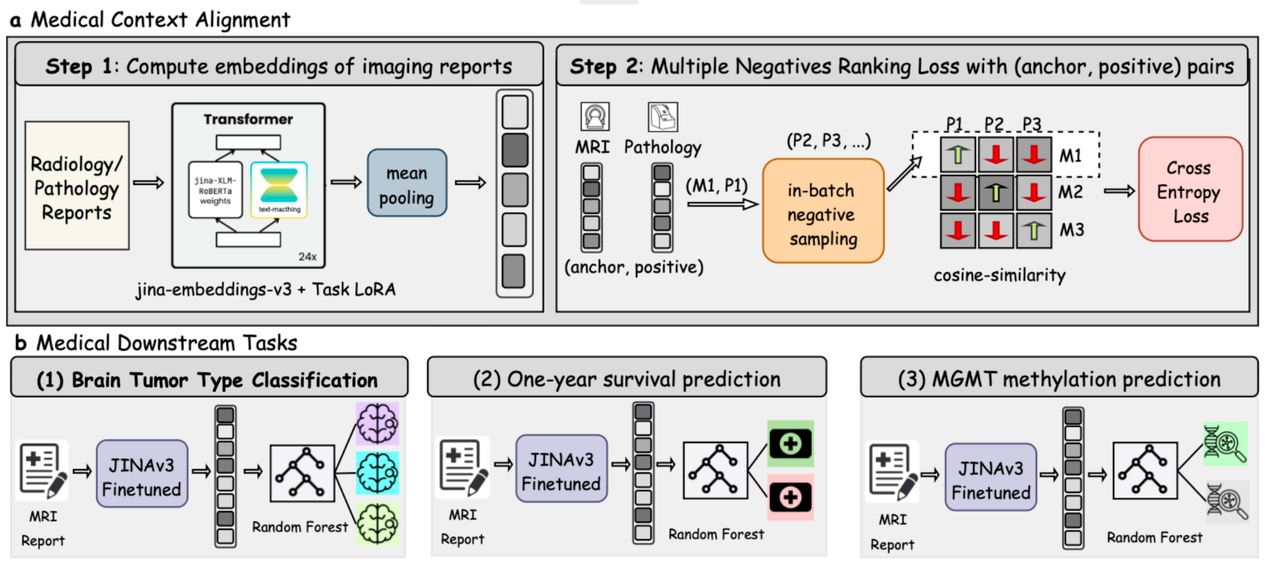

Figure 2. Workflow for GEM alignment and downstream classification. (a) Matched radiology-pathology reports are encoded with jina-embeddings-v3 (Ref. 2) and fine-tuned using MNRL with in-batch negatives. (b) Frozen GEM embeddings are evaluated with random forest classifiers for tumor type, 1-year survival, and MGMT methylation.

METHODS AND MATERIALS

We retrospectively identified glioma patients treated from 1995–2024 under IRB approval. The source corpus contained 90,134 clinical reports from 16,149 patients, including 64,093 MRI radiology reports and 26,041 pathology reports.

Radiology and pathology reports were matched within patient and tumor subtype using a many-to-many strategy, generating 97,398 positive pairs from 4,250 patients. We fine-tuned jina-embeddings-v3 (570M parameters; Ref. 2) with Multiple-Negatives Ranking Loss (Figure 2), using matched reports as positive pairs and in-batch mismatches as negatives.

Cross-modal retrieval was evaluated in a held-out cohort of 850 patients. Frozen embeddings were then evaluated with patient-disjoint stratified 5-fold cross-validation for 3-class tumor type, MGMT promoter methylation, and 1-year overall survival. Performance was compared with generalist and clinical encoder baselines and zero-shot decoder LLMs.

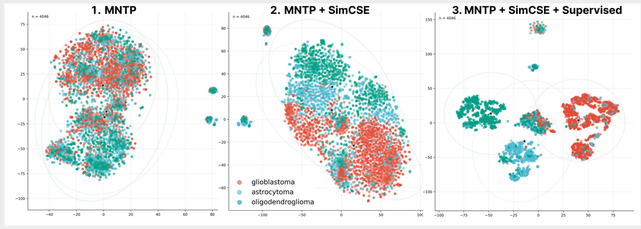

Figure 3. LLM2Vec t-SNE visualization of embeddings (Ref. 4) across MNTP, SimCSE, and supervised MNRL stages. Progressive alignment produces more compact subtype clusters and clearer separation among glioblastoma, astrocytoma, and oligodendroglioma.

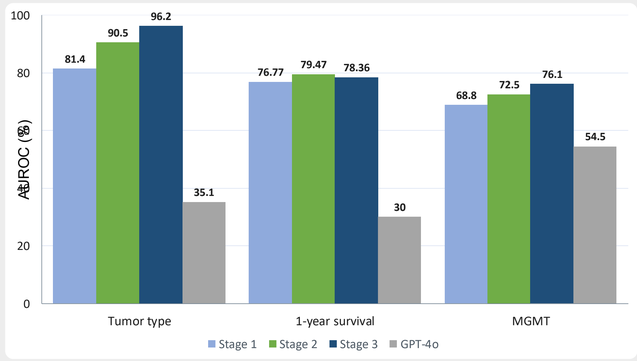

Figure 4. LLM2Vec AUROC by alignment stage versus zero-shot GPT-4o (Refs. 3–4)

DISCUSSION

Radiology-pathology alignment enables a compact encoder to learn clinically meaningful relationships between imaging descriptions and molecular diagnosis without task-specific labels. GEM improved retrieval and transferred across diagnostic, molecular, and prognostic endpoints, supporting the value of EHR-native cross-modal pairing for specialized oncology NLP.
In contrast, zero-shot decoder LLMs showed unstable or poorly calibrated AUROC despite occasional accuracy parity (Figure 4; Ref. 3), highlighting the limitation of generation-optimized models for clinical classification.
MNRL provided the strongest alignment among tested objectives (Figures 2–4; Refs. 2 and 4). These findings support clinically aligned encoders as an efficient foundation for neuro-oncology decision support; external and prospective validation remain necessary.

CONCLUSIONS

Fine-tuning jina-embeddings-v3 on matched radiology-pathology reports (Figure 2; Ref. 2) produced clinically aligned glioma embeddings that improved cross-modal retrieval and transferred across tumor type, MGMT methylation, and 1-year survival prediction. Natural EHR report pairing provides a scalable, label-free strategy for specialty model alignment. GEM offers an efficient foundation for neuro-oncology decision support, pending external and prospective validation.

REFERENCES

1. Louis DN, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*. 2021;23:1231-1251.
2. Sturua S, et al. jina-embeddings-v3: Multilingual embeddings with task LoRA. arXiv:2409.10173, 2024.
3. Hurst A, et al. GPT-4o System Card. arXiv:2410.21276, 2024.
4. Behnamghader P, et al. LLM2Vec: Large language models are secretly powerful text encoders. arXiv:2404.05961, 2024.

Funded by the Department of Defense Impact Award (W81XWH-22-1-0695), the Uncle Kory Foundation Seed Grant (2023–2024), the UCSF HDFCCC Computational Seed Grant (2024–2025), and the UCSF SPORE Brain Tumor Career Enhancement Program (2024–2025).