Hallucination filtering in radiology vision-language models using discrete semantic entropy

Info

PUBLISHED DATE
2026-02-20

JOURNAL
European Radiology

LINK
https://doi.org/10.1007/s00330-026-12384-z

Authors

Patrick Wienholt Sophie Caselitz Robert Siepmann Philipp Bruners Keno Bressem Christiane Kuhl Jakob Nikolas Kather Sven Nebelung Daniel Truhn

Abstract

Objective: To determine whether using discrete semantic entropy (DSE) to reject questions likely to generate hallucinations can improve the accuracy of black-box vision-language models (VLMs) in radiologic image-based visual question answering (VQA).

Materials and methods: This retrospective study evaluated DSE using two publicly available, de-identified datasets: the VQA-Med 2019 benchmark (500 images with clinical questions and short-text answers) and a diagnostic radiology dataset (206 cases: 60 computed tomography scans, 60 magnetic resonance images, 60 radiographs, 26 angiograms) with corresponding ground-truth diagnoses. GPT-4o and GPT-4.1 (Generative Pretrained Transformer) answered each question 15 times using a temperature of 1.0. Baseline accuracy was determined using low-temperature answers (0.1). Meaning-equivalent responses were grouped using bidirectional entailment checks, and DSE was computed from the relative frequencies of the resulting semantic clusters. Accuracy was recalculated after excluding questions with DSE > 0.6 or > 0.3. p values and 95% confidence intervals were obtained using bootstrap resampling and a Bonferroni-corrected threshold of $p < 0.004$ for statistical significance.

Results: Across 706 image-question pairs, baseline accuracy was 51.7% for GPT-4o and 54.8% for GPT-4.1. After filtering out high-entropy questions (DSE > 0.3), accuracy on the remaining questions was 76.3% (retained questions: 334/706) for GPT-4o and 63.8% (retained questions: 499/706) for GPT-4.1 (both $p < 0.001$). Accuracy gains were observed across both datasets and largely remained statistically significant after Bonferroni correction.

Conclusion: DSE enables reliable hallucination detection in black-box VLMs by quantifying semantic inconsistency. This method significantly improves diagnostic answer accuracy and offers a filtering strategy for clinical VLM applications.

Key points: Question Can DSE identify hallucination-prone questions and improve the reliability of black-box vision-language models in radiologic image-based VQA? Findings DSE filtering at a 0.3 threshold increased GPT-4o accuracy from 51.7% to 76.3% and GPT-4.1 from 54.8% to 63.8%, while answering fewer questions. Clinical relevance Integrating DSE as a black-box uncertainty filter enables selective answering and explicit

uncertainty display for radiology vision-language tools, supporting safer diagnostic use, mitigating hallucinations, and improving clinicians' trust in AI-assisted image interpretation.

Keywords: Data accuracy; Diagnostic imaging; Entropy; Generative artificial intelligence; Image interpretation (Computer-assisted).