

## *Supporting Information*

### **Evaluating Hallucination and Diagnostic Reliability of LLMs on Medical Image-Based Multiple Choice Tasks**

Debapriya Hazra<sup>1</sup>, Shayani Mukherjee<sup>2</sup>, Suman Kumar<sup>3</sup>, Subhajit Chatterjee<sup>4</sup>, Prince Waqas Khan<sup>1, \*</sup>, Khizar Abbas<sup>5, \*</sup>

<sup>1</sup>School of Computing, Montclair State University, United States

<sup>2</sup>Department of Surgery, McGowan Institute for Regenerative Medicine, University of Pittsburgh, United States

<sup>3</sup>Department of Computer Science, Cleveland State University, United States

<sup>4</sup>Faculty of Engineering and Natural Sciences, Tampere University, Finland

<sup>5</sup>The College of Computing and Information Technology, University of Doha for Science and Technology, Doha, Qatar

\*Corresponding Authors: Khizar Abbas ([engr.khizarabbas14@gmail.com](mailto:engr.khizarabbas14@gmail.com))

and Prince Waqas Khan ([khanp@montclair.edu](mailto:khanp@montclair.edu))

**Table S1.** LLM reasoning for correct and incorrect selections (by case and model).

Model Name	Case	Reasoning for Correct Response	Reasoning for Wrong Response
ChatGPT o4-mini-high	Orbital Dermoid	A well-circumscribed superolateral orbital mass in the lacrimal fossa that is hyperintense on both T1 and T2 and nulls on fat-suppressed images, with homogeneous fat signal and no solid enhancing component gently displacing (but not invading) the globe—classic for a dermoid cyst.	
	Behcet Disease/syndrome		The coronal T2/FLAIR images demonstrate marked hyperintense signal and swelling of the bilateral (especially medial) temporal lobes. On the post-contrast T1 sequence there is patchy enhancement in these regions—an imaging pattern characteristic of viral (herpes-type) limbic encephalitis.
	colloid (mucinous) carcinoma		On mammography there is a focal architectural distortion with radiating spicules and a small central lucency at the site of a biopsy clip, rather than a discrete mass or clustered calcifications. On ultrasound the corresponding area appears as an ill-defined, hypoechoic zone with posterior shadowing, features characteristic of a radial scar (complex sclerosing lesion) rather than invasive carcinoma or simple fibrocystic changes.
	Lupus Myelitis	On sagittal T2-weighted images there are long-segment, centrally-located hyperintense lesions spanning multiple vertebral levels in the cervicothoracic cord (arrows), a pattern of longitudinally extensive transverse myelitis classically seen with SLE rather than the short-segment	

		<p>peripherally located plaques of MS or the isolated dorsal-column signal of B12 deficiency.</p>	
<p>ChatGPT 4o</p>	<p>Orbital Dermoid</p>	<p>The lesion is well-circumscribed and located in the superolateral orbit, consistent with the lacrimal fossa region.</p> <p>It shows high signal intensity on T2-weighted MRI and appears hyperintense with a fat component, suggesting the presence of fat or sebaceous material.</p> <p>The mass does not invade adjacent structures and has a benign appearance, typical for an orbital dermoid cyst, a congenital lesion commonly arising near the zygomaticofrontal suture.</p> <p>These features are classic for an orbital dermoid, differentiating it from lipomas (which would be purely fat-density throughout), pleomorphic adenomas (more irregular, often enhancing), and preauricular dermoids (which occur near the ear, not orbit).</p>	

	Behcet Disease/syndrome	<p>"Key Imaging Findings:</p> <p>Hyperintense signal on FLAIR and T2-weighted images involving the medial temporal lobes and inferior frontal lobes, particularly noticeable on axial and coronal views.</p> <p>Mild mass effect with preservation of gray-white matter differentiation in early stages.</p> <p>Post-contrast image shows mild enhancement, consistent with disruption of the blood-brain barrier.</p> <p>Reasoning: These imaging findings are classic for herpes simplex virus (HSV) encephalitis, the most common cause of viral encephalitis, which characteristically involves the mesial temporal lobes and orbital frontal regions. The pattern and location of involvement help differentiate it from MS (which is periventricular and multifocal), vasculitis (which tends to be more diffuse or involve small vessel infarcts), or Behçet's disease (which often shows brainstem and deep white matter involvement)."</p>
--	----------------------------	---

	colloid (mucinous)carcinoma		<p>Mammographic views (Images 1 &amp; 2) show a spiculated central lesion without a clearly defined mass, typical of a radial scar. The central lucency and radiating lines mimic carcinoma, but lack dense central mass.</p> <p>Ultrasound (Image 3) reveals a hypoechoic area with architectural distortion but no discrete solid mass—supporting a benign sclerosing lesion like a radial scar.</p> <p>Reasoning: Radial scars can closely mimic carcinoma radiologically, but their hallmark features are a central radiolucent core with long thin spicules and distortion without a dominant mass. The imaging pattern and absence of aggressive features (e.g., microcalcifications or high-density irregular mass) argue against malignancy or mucinous carcinoma.</p>
	Disseminated Coccidioidomycosis	<p>MRI and CT images show destructive changes of vertebral bodies, with evidence of vertebral body collapse and paravertebral soft tissue involvement.</p> <p>There is abnormal enhancement and edema within the vertebral bodies and discs, indicating osteomyelitis/discitis, and epidural extension suggestive of an infectious process.</p> <p>The imaging lacks features classic for metastatic carcinoma (e.g., multiple non-contiguous lesions) and doesn't match the expected findings for Pott's disease or idiopathic hypertrophic pachymeningitis.</p>	

Claude Sonnet 4	Orbital Dermoid	<p>The lesion demonstrates mixed signal intensity with both hyperintense (fat-like) and hypointense components on T2-weighted images, which is characteristic of dermoid cysts containing both fatty and proteinaceous material. The mass appears well-circumscribed and is located in the superolateral aspect of the left orbit, a typical location for orbital dermoids. The heterogeneous internal architecture with what appears to be fat-fluid levels or mixed tissue components is pathognomonic for dermoid cysts, distinguishing it from the more homogeneous appearance expected in lipomas or the different signal characteristics typical of pleomorphic adenomas.</p>	
	Behcet Disease/syndro me		<p>"Key Imaging Findings:</p> <p>Multiple bilateral hyperintense lesions in the deep white matter and periventricular regions (Images 1 and 3)</p> <p>Lesions appear to have a periventricular distribution with some extending perpendicular to the ventricles</p> <p>The coronal FLAIR image (Image 2) shows characteristic ovoid lesions in the white matter with the typical ""Dawson's fingers"" appearance</p> <p>Most Likely Diagnosis: Option A - Multiple Sclerosis (MS)</p> <p>The imaging findings are classic for MS, showing the characteristic periventricular white matter lesions with the typical distribution and morphology. The lesions</p>

		<p>appear hyperintense on FLAIR sequences and demonstrate the classic periventricular pattern that is pathognomonic for demyelinating disease, particularly multiple sclerosis. The bilateral, multifocal nature and specific anatomical distribution strongly support this diagnosis over the other inflammatory conditions listed."</p>
	colloid (mucinous)carcinoma	<p>Image 1 &amp; 2: Mammographic views showing a spiculated mass with radiating linear densities extending outward from a central point, creating a characteristic "star-burst" or "spoke-wheel" appearance.</p> <p>Image 3: Ultrasound of the left breast showing a hypoechoic area with irregular margins.</p> <p>Most likely diagnosis: Option A - Radial Scar</p> <p>Key imaging findings supporting this diagnosis:</p> <p>The mammographic images demonstrate the pathognomonic spiculated appearance with long, thin radiating lines extending from a central lucent or low-density area, creating the classic "black star" sign characteristic of radial scars. The ultrasound shows corresponding architectural distortion without a discrete solid mass, which is typical for this benign but high-risk lesion that can mimic malignancy on imaging.</p>

	Lupus Myelitis	<p>The sagittal T2-weighted images (Images 1-3) show extensive hyperintense (bright white) signal abnormality involving a long segment of the spinal cord, extending from the cervical through thoracic regions</p> <p>Image 4 appears to be a coronal STIR or T2-weighted sequence showing the longitudinal extent of the lesion with arrows pointing to areas of signal abnormality</p> <p>The lesion appears to span multiple vertebral segments (longitudinally extensive transverse myelitis pattern)</p> <p>The imaging demonstrates a longitudinally extensive transverse myelitis (LETM) pattern with T2 hyperintense signal changes spanning multiple spinal cord segments. This appearance is characteristic of lupus myelitis, which typically presents as extensive cord involvement over 3 or more vertebral segments, distinguishing it from the shorter lesions typically seen in multiple sclerosis or the specific posterior column involvement seen in subacute combined degeneration.</p>	
--	----------------	--	--



Gemini 2.5 Pro	Orbital Dermoid	The axial and sagittal T2-weighted MRI images demonstrate a well-circumscribed, ovoid mass in the supero-temporal aspect of the left orbit, which appears markedly hyperintense (bright), consistent with a cystic lesion. The T1-weighted images show the lesion to be largely hypointense (dark). This specific location (supero-temporal orbit, often near the frontozygomatic suture) and the appearance of an encapsulated, non-infiltrating cystic mass are classic for an orbital dermoid cyst.	
	Behcet Disease/syndrome	The images demonstrate a focal, ovoid lesion that is hyperintense (bright) on FLAIR and T2-weighted sequences and shows enhancement after contrast administration. The most critical finding is its specific location at the mesodiencephalic junction (the area between the midbrain and thalamus), which is a classic and highly characteristic site for parenchymal lesions in Neuro-Behçet's disease.	
	colloid (mucinous)carcinoma		The mammograms demonstrate a classic spiculated lesion with long, thin spicules radiating from a more radiolucent (darker) center, causing significant architectural distortion. This "black star" appearance, without a definite central high-density mass, is highly suggestive of a radial scar. The corresponding ultrasound shows an irregular, hypoechoic mass with indistinct margins that correlates with the mammographic finding.

	Lupus Myelitis		<p>The sagittal T2-weighted MRI images demonstrate a long, continuous segment of high signal intensity (hyperintensity) predominantly involving the posterior columns of the cervical and upper thoracic spinal cord. This specific pattern of involvement is a classic radiological sign of subacute combined degeneration, which is typically caused by vitamin B12 deficiency.</p>
--	----------------	--	---

## I. Case Examples Where Distractors Fooled the LLMs

**Case 1 (True: Celiac Artery Aneurysm; Model: Claude Sonnet 4, predicted Splenic Artery Aneurysm (Anatomical Distractor)).** The model's reasoning was: "Multiple serpiginous, tortuous vascular structures in the left upper quadrant near the spleen. Calcified vessel walls creating a characteristic 'eggshell' or rim-like calcification pattern. The tortuous vessels appear to be in the distribution of the splenic artery. Most Likely Diagnosis: Option D - Splenic Artery Aneurysm. The imaging shows the classic appearance of a splenic artery aneurysm with its characteristic tortuous, serpentine course and calcified walls. This is the most common visceral artery aneurysm and typically presents with this distinctive 'coiled spring' or 'string of beads' appearance on CT imaging due to the naturally tortuous course of the splenic artery, which becomes more pronounced when aneurysmal."

This reflects location anchoring and a vessel-identity swap: anchoring on "left upper quadrant near the spleen" led to assuming splenic-artery origin rather than verifying takeoff from the celiac trunk. It also shows prevalence-prior overreach leaning on "most common visceral artery aneurysm" as a tiebreaker instead of anatomy and cue conflation, where terms like "string of beads" (tortuosity/dysplasia) were over-generalized to justify "splenic" without confirming the parent vessel. In short, this is a mislocalization error amplified by a frequency heuristic: the explanation sounds plausible but points to the wrong artery.

**Case 2 (True: Behçet Disease/Syndrome; Model: o4-mini-high, predicted Viral Encephalitis (Semantic Distractor)).** The model's reasoning was: "The coronal T2/FLAIR images demonstrate marked hyperintense signal and swelling of the bilateral (especially medial) temporal lobes. On the post-contrast T1 sequence there is patchy enhancement in these regions—an imaging pattern characteristic of viral (herpes-type) limbic encephalitis."

It "template-matched" the textbook HSV pattern (bilateral medial temporal FLAIR hyperintensity + patchy enhancement) and treated that semantic label as decisive. The prevalence/teaching prior for "viral encephalitis" overrode case-specific cues for Behçet, and the model didn't actively seek disconfirming evidence (e.g., vascular/inflammatory context), so it reused stock HSV language with unwarranted certainty.

**Case 3 (True Diagnosis: Desmoplastic Small Round Cell Tumor; Model: Claude Sonnet 4, predicted Neuroblastoma [Visual Distractor]).** The model's reasoning was: "Ultrasound images showing spleen examination with Doppler flow studies CT scan of a 15-year-old male showing a large abdominal mass (measuring approximately 93.5 mm) Pancreatic ultrasound images

Key imaging findings:

Large heterogeneous abdominal mass in a 15-year-old male patient on CT

The mass appears to be in the retroperitoneal/abdominal cavity with significant size

Ultrasound shows involvement of multiple organs including spleen and pancreatic region

Most likely diagnosis: Option D - Neuroblastoma

Supporting findings:

The large heterogeneous abdominal mass in a pediatric patient (15 years old) is highly characteristic of neuroblastoma, which commonly presents as a retroperitoneal or abdominal mass in children

and adolescents. The imaging demonstrates a substantial mass lesion with the typical appearance and location consistent with neuroblastoma, which is one of the most common solid tumors in the pediatric age group.”

The model hallucinated the diagnosis due to over-reliance on age-related common tumor patterns and the visual appearance of a large heterogeneous abdominal mass, which are classic for neuroblastoma in children and adolescents. It disregarded other key context—such as multiorgan involvement and imaging characteristics that are atypical for neuroblastoma but more specific for desmoplastic small round cell tumor (DSRCT)—misled by dominant visual distractors.

**Table S2.** Distractor susceptibility by model (lower counts = more resistant).

LLM	Visual Distractor (VD) Option	Anatomical Distractor (AD) Option	Semantic Distractor (SD) Option
ChatGPT o4-mini-high	9	3	6
ChatGPT 4o	9	2	7
Claude Sonnet 4	6	7	3
Gemini 2.5 Pro	5	3	1

## II. Case Examples Where Distractors Fooled the LLMs

**Case 1:** Brain & Neuro (Body Part/Specialization); True diagnosis: Option D (Behçet Disease/Syndrome); Distractors configured: VD = Option A (Multiple Sclerosis); AD = Option B (Cerebral Vasculitis); SD = Option C (Viral Encephalitis).

- **Model outcomes:**

- ChatGPT o4-mini-high and ChatGPT 4o chose Option C (SD; Viral Encephalitis), while Claude Sonnet 4 chose Option A (VD; Multiple Sclerosis) all options were wrong. Only Gemini 2.5 Pro chose the correct Option D.

- **Reasoning by LLMs**

- **ChatGPT o4-mini-high:** The coronal T2/FLAIR images demonstrate marked hyperintense signal and swelling of the bilateral (especially medial) temporal lobes. On the post-contrast T1 sequence there is patchy enhancement in these regions—an imaging pattern characteristic of viral (herpes-type) limbic encephalitis.
- **ChatGPT 4o:** Key Imaging Findings: Hyperintense signal on FLAIR and T2-weighted images involving the medial temporal lobes and inferior frontal lobes, particularly noticeable on axial and coronal views. Mild mass effect with preservation of gray-white matter differentiation in early stages. Post-contrast image shows mild enhancement, consistent with disruption of the blood-brain barrier. Reasoning: These imaging findings are classic for herpes simplex virus (HSV) encephalitis, the most common cause of viral encephalitis, which characteristically involves the mesial temporal lobes and orbital frontal regions. The pattern and location of involvement help differentiate it from MS (which is periventricular and multifocal), vasculitis (which tends to be more diffuse or involve small vessel infarcts), or Behçet’s disease (which often shows brainstem and deep white matter involvement).
- **Claude Sonnet 4:** Key Imaging Findings: Multiple bilateral hyperintense lesions in the deep white matter and periventricular regions (Images 1 and 3) Lesions appear to have a periventricular distribution with some extending perpendicular to the ventricles. The coronal FLAIR image (Image 2) shows characteristic ovoid

lesions in the white matter with the typical "Dawson's fingers" appearance. Most Likely Diagnosis: Option A - Multiple Sclerosis (MS). The imaging findings are classic for MS, showing the characteristic periventricular white matter lesions with the typical distribution and morphology. The lesions appear hyperintense on FLAIR sequences and demonstrate the classic periventricular pattern that is pathognomonic for demyelinating disease, particularly multiple sclerosis. The bilateral, multifocal nature and specific anatomical distribution strongly support this diagnosis over the other inflammatory conditions listed.

- **Gemini 2.5 Pro:** The images demonstrate a focal, ovoid lesion that is hyperintense (bright) on FLAIR and T2-weighted sequences and shows enhancement after contrast administration. The most critical finding is its specific location at the mesodiencephalic junction (the area between the midbrain and thalamus), which is a classic and highly characteristic site for parenchymal lesions in Neuro-Behçet's disease.

**Case 2:** Breast (Body Part/Specialization); True diagnosis: Option A (Bilateral axillary lymphadenopathy secondary to SLE); Distractors configured: VD = Option C (Bilateral breast fibroadenomas), AD = Option B (Inguinal lymphadenopathy), SD = Option D (Granulomatous mastitis).

- **Model outcomes:**

- ChatGPT o4-mini-high, ChatGPT 4o, and Claude Sonnet 4 all chose Option D (SD), while Gemini 2.5 Pro chose the true Option A.

- **Reasoning by LLMs**

- **ChatGPT o4-mini-high:** The mammographic images reveal asymmetric increased density and architectural distortion in both breasts. No discrete mass is clearly visible, but there is ill-defined, dense parenchymal tissue, especially notable in the subareolar regions. Magnified views show irregular parenchymal changes without the presence of typical benign lesions such as well-circumscribed fibroadenomas. The imaging pattern, in conjunction with bilateral involvement and the absence of significant calcifications or mass effect, is most consistent with granulomatous mastitis, a rare inflammatory breast condition often mimicking malignancy but lacking a true neoplastic mass.
- **ChatGPT 4o:** On both CC and MLO views there are ill-defined, asymmetric areas of increased density in the subareolar region with tubular (“duct-like”) extensions and mild architectural distortion—but no discrete, circumscribed masses or popcorn calcifications. Magnification views show no true microcalcifications, supporting an inflammatory rather than neoplastic process.
- **Claude Sonnet 4:** The images show bilateral breast involvement with heterogeneous, irregular tissue changes and areas of increased density/enhancement. The pattern appears inflammatory rather than representing discrete masses typical of fibroadenomas. The bilateral nature and inflammatory appearance are characteristic of granulomatous mastitis, which can present as diffuse, irregular breast tissue changes with areas of inflammation and potential fluid collections. The imaging findings do not show the discrete, well-circumscribed masses expected with fibroadenomas, nor do they demonstrate clear lymphadenopathy in the axillary or inguinal regions that would support the other diagnostic options.

- **Gemini 2.5 Pro:** The mammographic images, particularly the axillary tail and spot compression views (images 3, 4, 5, 6), demonstrate multiple enlarged and dense lymph nodes in the axilla. The presence of these findings across multiple images suggests a bilateral process. This reactive lymphadenopathy can be associated with various systemic inflammatory conditions, including Systemic Lupus Erythematosus (SLE).

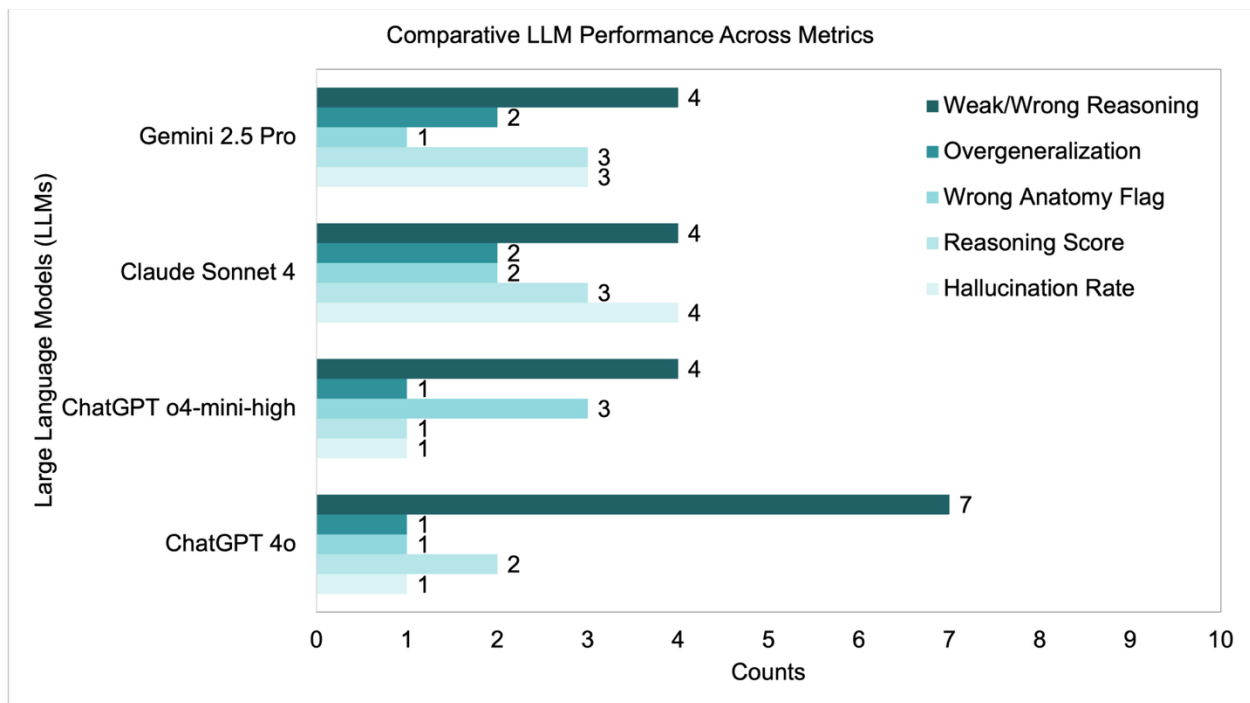
**Case 3:** Breast (Body Part/Specialization); True diagnosis: Option D – Colloid (mucinous) carcinoma; Distractors configured: VD = Option A (Radial scar); AD = Option C (Invasive lobular carcinoma of the axilla); SD = Option B (Fibrocystic changes with microcalcifications).

- **Model outcomes:**

- ChatGPT o4-mini-high, ChatGPT 4o, Claude Sonnet 4, and Gemini 2.5 Pro all chose Option A (VD), which is incorrect.

- **Reasoning by LLMs**

- **ChatGPT o4-mini-high:** On mammography there is a focal architectural distortion with radiating spicules and a small central lucency at the site of a biopsy clip, rather than a discrete mass or clustered calcifications. On ultrasound the corresponding area appears as an ill-defined, hypoechoic zone with posterior shadowing, features characteristic of a radial scar (complex sclerosing lesion) rather than invasive carcinoma or simple fibrocystic changes.
- **ChatGPT 4o:** Mammographic views (Images 1 & 2) show a spiculated central lesion without a clearly defined mass, typical of a radial scar. The central lucency and radiating lines mimic carcinoma but lack dense central mass. Ultrasound (Image 3) reveals a hypoechoic area with architectural distortion but no discrete solid mass—supporting a benign sclerosing lesion like a radial scar. Reasoning: Radial scars can closely mimic carcinoma radiologically, but their hallmark features are a central radiolucent core with long thin spicules and distortion without a dominant mass. The imaging pattern and absence of aggressive features (e.g., microcalcifications or high-density irregular mass) argue against malignancy or mucinous carcinoma.
- **Claude Sonnet 4:** Image 1 & 2: Mammographic views showing a spiculated mass with radiating linear densities extending outward from a central point, creating a characteristic "star-burst" or "spoke-wheel" appearance. Image 3: Ultrasound of the left breast showing a hypoechoic area with irregular margins. Most likely diagnosis: Option A - Radial Scar. Key imaging findings supporting this diagnosis: The mammographic images demonstrate the pathognomonic spiculated appearance with long, thin radiating lines extending from a central lucent or low-density area, creating the classic "black star" sign characteristic of radial scars. The ultrasound shows corresponding architectural distortion without a discrete solid mass, which is typical for this benign but high-risk lesion that can mimic malignancy on imaging.
- **Gemini 2.5 Pro:** The mammograms demonstrate a classic spiculated lesion with long, thin spicules radiating from a more radiolucent (darker) center, causing significant architectural distortion. This "black star" appearance, without a definite central high-density mass, is highly suggestive of a radial scar. The corresponding ultrasound shows an irregular, hypoechoic mass with indistinct margins that correlates with the mammographic finding.



**Fig. S1.** Comparative performance of LLMs across proposed evaluation metrics.