BrainBench-Calibrated: KG-Guided, Evidence-Calibrated Diagnosis of LLM Reasoning Failures in Connectome Neuroscience

Anonymous ACL submission

Abstract

Evaluating LLMs in neuroscience requires more than factual recall: valid mechanistic claims depend on anatomical granularity, disease context, causal direction, and evidential scope. We introduce **BrainBench-Calibrated**, a literature-grounded benchmark for evidence-calibrated mechanistic reasoning in connectome neuroscience. The benchmark comprises multi-select questions and option-level claims spanning five neuroscience-specific reasoning paradigms: anatomical granularity, cross-path conflict, cross-disease differential reasoning, causal direction, and counterfactual intervention. To construct the benchmark, we use a knowledge graph only to generate structurally plausible candidate paths, while assigning final option labels against primary-literature evidence cards and a six-category Evidence-Gap Taxonomy through independent LLM auditing and risk-stratified adjudication. In closed-book evaluation, even the strongest contemporary LLMs fall well short of reliable performance, with overlapping confidence intervals among top systems. Option-level analyses further reveal distinct conservative and over-inclusive selection behaviors. Evidence-enhanced prompting improves several models substantially, but gains are uneven across models and reasoning paradigms, indicating that knowledge access helps but is not sufficient. Error analysis shows persistent vulnerabilities to hub stitching, granularity inheritance, and absolute overclaiming, highlighting ongoing difficulty in preserving evidential scope and anatomical constraints in neuroscience reasoning.

1 Introduction

Large Language Models have shown strong performance on biomedical and clinical question answering, yet evaluating their scientific reliability requires moving beyond factual recall. This is especially important in neuroscience, where many valid claims are not universal statements but conditionally supported mechanisms. A conclusion may hold only at a specific anatomical scale, disease stage, imaging modality, or evidence level. For example, a disruption observed at the level of a macroscale network does not automatically apply to each subregion within that network; two disorders may share the same functional network hub while arising from different pathological processes; and cross-sectional imaging associations cannot be treated as causal interventions. Therefore, the central evaluation question is not simply whether a model knows neuroscience terminology, but whether it can preserve the structural and evidential constraints under which a mechanistic claim is valid.

Existing benchmarks only partially address this problem. PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and related biomedical QA datasets primarily evaluate factual retrieval, reading comprehension, or exam-style clinical knowledge. Expert-level benchmarks such as GPQA (Rein et al., 2023), MMLU-Pro (Wang et al., 2024), LAB-Bench (Laurent et al., 2024), SciBench (Wang et al., 2023), and MedXpertQA (Zuo et al., 2025) increase task difficulty, but they are not designed to diagnose neuroscience-specific reasoning failures. BrainBench (Luo et al., 2024) is the closest domain-specific benchmark, but it evaluates whether models can distinguish original neuroscience abstracts from altered ones, emphasizing experimental outcome prediction rather than option-level mechanism verification. As a result, current evaluations provide limited insight into whether LLMs can distinguish supported neuroscience mechanisms from plausible but unwarranted extrapolations.

We argue that mechanistic reasoning in connectome neuroscience requires at least five distinct but complementary capabilities. First, **anatomical**

granularity reasoning asks whether a mechanism can be validly transferred across brain networks, regions, and subregions. Second, **cross-path conflict reasoning** asks whether a model can reconcile multiple graph paths that share hubs but imply different mechanisms. Third, **cross-disease differential reasoning** asks whether a model can distinguish disorders that involve overlapping networks but differ in regional, pathological, or biomarker signatures. Fourth, **causal-direction reasoning** asks whether a model can distinguish causes, consequences, correlations, and confounds based on the underlying evidence design. Fifth, **counterfactual reasoning** asks whether a model can reason about what would happen if a node, pathway, or mechanism were removed or altered. Together, these five paradigms form a diagnostic loop for neuroscience reasoning: localizing a mechanism, integrating network paths, discriminating diseases, interpreting causal evidence, and evaluating intervention-like alternatives.

A natural way to construct such structured tasks is to extract knowledge graphs from scientific literature and use graph paths to generate multi-hop questions. However, directly treating graph connectivity as ground truth is risky. We identify three recurring failure modes in automatically constructed neuroscience KGs. **Semantic collapse** compresses conditional findings into weak generic relations such as "associated with." **Hub splicing** connects disease mechanisms through shared network hubs even when the underlying evidence supports different pathological processes. **Knowledge incompleteness** omits important clinical or methodological constraints, causing valid claims to be mislabeled as unsupported. If KG topology determines labels, these errors can become pseudo-correct or pseudo-incorrect benchmark items.

BrainBench-Calibrated addresses this risk by decoupling structural plausibility from truth adjudication. The KG is used only as a candidate generator: it proposes plausible entity combinations and reasoning paths. Final option labels are assigned relative to primary-literature evidence cards. This design does not assume that the KG is error-free. Instead, graph-induced ambiguity becomes useful for constructing hard distractors, while evidence cards prevent graph topology from directly determining correctness.

Our work makes four contributions.

First, we formalize *evidence-grounded mechanistic claim verification* as a diagnostic evaluation task for connectome neuroscience. Unlike standard biomedical QA, the task requires models to select only those claims warranted by evidence under explicit anatomical, disease, temporal, and methodological constraints.

Second, we design five neuroscience-specific reasoning paradigms—anatomical granularity, cross-path conflict, cross-disease differential reasoning, causal direction, and counterfactual intervention—that operationalize recurring forms of mechanistic inference in brain science. Based on these paradigms, we construct BrainBench-Calibrated, a benchmark of 43 multi-select questions and 215 option-level claims.

Third, we propose a KG-guided construction framework that decouples structural plausibility from truth adjudication. Knowledge graphs are used to generate candidate reasoning paths and hard distractors, while final option labels are assigned through primary-literature evidence cards and a six-category Evidence-Gap Taxonomy.

Fourth, we evaluate 11 LLMs under both closed-book and evidence-enhanced settings. The results show that current models remain brittle in full-configuration mechanistic reasoning: errors are not explained solely by missing knowledge, but also by failures to preserve anatomical granularity, disease-specific boundaries, and evidential scope.

2 Related Work

Table 1 positions BrainBench-Calibrated among biomedical, scientific, and neuroscience benchmarks.

Biomedical and scientific LLM evaluation. Biomedical QA benchmarks such as PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), and MedMCQA (Pal et al., 2022) evaluate reading comprehension, exam knowledge, and clinical decision making. Recent resources such as MedXpertQA (Zuo et al., 2025), CARDBiomedBench (Bianchi et al., 2026), and MedR-Bench (Qiu et al., 2025) extend coverage to harder clinical and rare-disease settings, while GPQA (Rein et al., 2023), MMLU-Pro (Wang et al., 2024), and LAB-Bench (Laurent et al., 2024) test broader scientific reasoning. These benchmarks primarily supervise final answers and do not diagnose neuroscience-specific failures in anatomy, disease context, or causal evidence.

KG-grounded biomedical benchmarks. BioHopR (Kim et al., 2025) uses PrimeKG to build

Benchmark	Scale (#Q)	Domain	Neuroscience specific	Mech. reasoning focus	KG-guided construction	Diagnostic error taxonomy
PubMedQA (Jin et al., 2019)	1,000 expert	Biomedical QA	×	×	×	×
MedQA (Jin et al., 2021)	12,723 USMLE	Clinical	×	×	×	×
MedMCQA (Pal et al., 2022)	193,155 AIIMS/NEET	Clinical	×	×	×	×
CARDBiomedBench (Bianchi et al., 2026)	68k open-ended	Rare disease / biomed.	×	Partial	×	×
MedR-Bench (Qiu et al., 2025)	1,453 cases	Clinical reasoning	×	Partial	×	×
MedXpertQA (Zuo et al., 2025)	4,460 board-level	Multi-specialty	×	Partial	×	×
BioHopR (Kim et al., 2025)	10,127 279k ans.	Biomed. (PrimeKG)	×	Partial	✓	×
KG+Literature Biomarker QA (Martell et al., 2025)	24k synthetic Q/A	Therapeutic biomarkers	×	Partial	✓	×
GPQA (Rein et al., 2023)	448 PhD-level	Bio / Chem / Phys	×	Partial	×	×
MMLU-Pro (Wang et al., 2024)	12k multi-domain	General science	×	×	×	×
LAB-Bench (Laurent et al., 2024)	2,400 expert	Biology research	×	Partial	×	×
BrainBench (Luo et al., 2024)	200 binary	Neuroscience	✓	×	×	×
BrainBench-Calibrated (Ours)	43 Q 215 claims	Connectome Neuroscience	✓	✓	✓	✓

Table 1: Comparison with representative benchmarks across three categories: biomedical QA and KG-grounded biomedical benchmarks, expert-level scientific benchmarks, and neuroscience-specific benchmarks. **BrainBench-Calibrated** uniquely combines KG-guided construction, option-level mechanistic claim verification, and a diagnostic error taxonomy in connectome neuroscience. "Partial" indicates that the dimension is touched on but is not central to the benchmark's design.

multi-hop, multi-answer biomedical questions, exposing LLM weaknesses on one-to-many and many-to-many relations. Martell et al. (Martell et al., 2025) use PrimeKG relations to retrieve PubMed evidence and synthesize biomarker Q/A pairs with LLM-based classification, summarization, and filtering. These works motivate KG use for systematic coverage, but differ from our setting: BioHopR treats graph reachability as the answer source, and the biomarker framework relies on synthetic LLM-generated Q/A. BrainBench-Calibrated instead uses KG connectivity only to propose candidate mechanisms, with labels assigned from literature-grounded evidence cards.

Neuroscience and KG-assisted evaluation. BrainBench (Luo et al., 2024) is the closest neuroscience benchmark, but it focuses on distinguishing original from altered abstracts rather than verifying option-level mechanistic claims. General KG-assisted benchmarks such as KG-LLM-Bench (Markowitz et al., 2025), MINTQA (He et al., 2024), OneEval (Chen et al., 2025), and OKGQA (Sui et al., 2025) evaluate graph traversal, multi-source QA, or robustness to noisy graphs. In connectome neuroscience, however, graph proximity can create plausible but unsupported claims through shared hubs. We therefore combine KG-guided construction with evidence calibration and an explicit diagnostic error taxonomy.

3 Creation of BrainBench-Calibrated

3.1 Task Definition

BrainBench-Calibrated is formulated as a multi-select mechanistic claim verification task. Each instance contains a question stem q_i, five candidate options $O_i = \{o_{i1}, \ldots, o_{i5}\}$, and a gold label vector $y_i \in \{0, 1\}^5$. A candidate option is labeled positive if it is directly supported or entailed by the curated evidence under the stated anatomical, disease, temporal, and methodological constraints. It is labeled negative if it is contradicted, overgeneralized, directionally invalid, or insufficiently warranted by the curated evidence.

The labels are evidence-relative. A negative label does not necessarily indicate that a claim is false across the entire neuroscience literature; rather, it indicates that the claim is not warranted by the evidence used to construct the item. This formulation evaluates whether models can identify evidence-grounded mechanistic claims, instead of merely recalling isolated neuroscience facts.

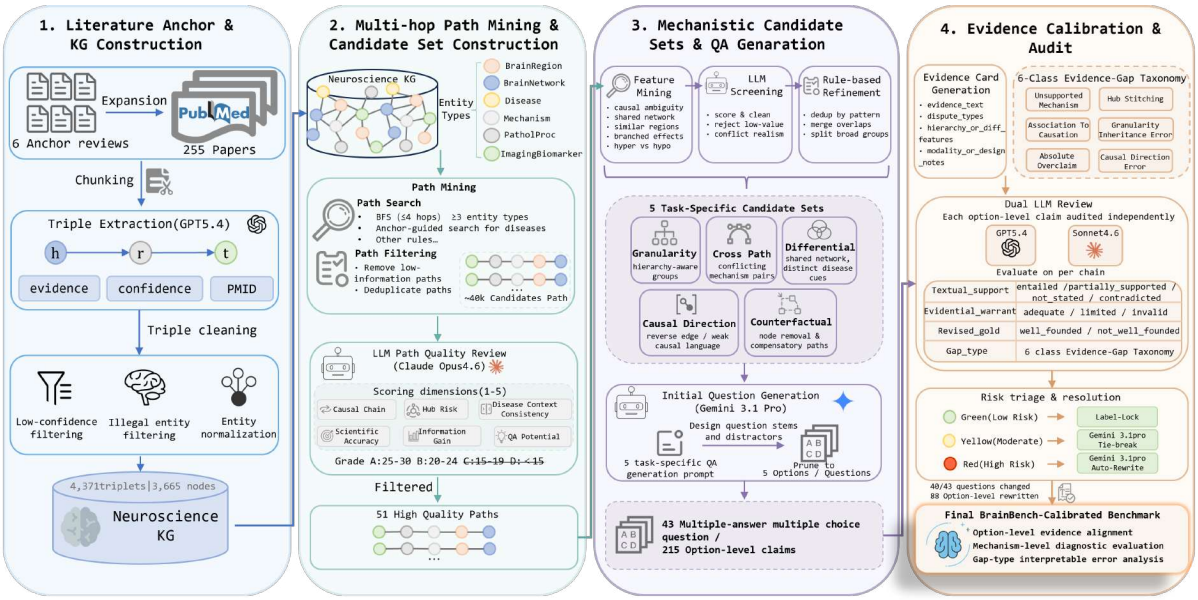

Figure 1: Construction pipeline of BrainBench-Calibrated.

3.2 Five Neuroscience Reasoning Paradigms

A central design choice of BrainBench-Calibrated is that questions are not generated from arbitrary KG paths. Instead, every item is assigned to one of five reasoning paradigms that correspond to recurring reasoning operations in connectome neuroscience. These paradigms translate domain-specific constraints into testable evaluation units.

This five-way design is important for two reasons. First, it reflects the structure of neuroscientific inference: valid mechanistic claims depend on anatomical hierarchy, disease specificity, evidence design, and intervention assumptions. Second, it enables diagnostic evaluation beyond aggregate accuracy, since different models may fail on different reasoning operations.

3.3 Overall Construction Framework

Figure 1 summarizes the full construction pipeline. BrainBench-Calibrated follows a KG-as-Candidate-Generator design. The pipeline contains four stages: literature anchoring and KG construction, multi-hop path mining, mechanistic question generation, and evidence calibration.

The key principle is to separate structural plausibility from scientific truth. The KG proposes candidate reasoning paths, but it never determines final correctness. This is necessary because LLM-extracted KGs can suffer from semantic collapse, hub splicing, and missing constraints. In our framework, these graph-induced ambiguities are useful for generating hard distractors, while final labels are assigned only through evidence cards grounded in primary literature.

3.4 Literature Corpus and KG Construction

We construct a focused corpus in macroscale connectome neuroscience. Six high-impact review papers serve as anchors, covering network communication, cross-disorder dysconnectivity, connectome-based neurodegeneration, structure-function coupling, and biomarker frameworks (Seguin et al., 2023; van den Heuvel and Sporns, 2019; Vogel et al., 2023; Fornito et al., 2015; Fotiadis et al., 2024; Kobayashi et al., 2026). From the reference lists of six anchor papers, we identified 255 additional studies that met our selection criteria: citation count > 80 and the presence of at least one predefined keyword related to macroscale brain networks, connectome alterations, neurodegenerative or neuropsychiatric disorders, or imaging biomarkers.

From standardized text chunks, LLMs extract structured triplets with typed entities, directed relations, source metadata, short evidence snippets, and relation explanations. After filtering low-confidence extractions, removing invalid entities, and normalizing entity names, the resulting KG contains 4,371 triplets and 3,665 nodes. The entity and relation schema, extraction prompts, and filtering rules are provided in Appendix G.

The resulting KG contains hub-like disease and network nodes, such as disease categories

and large-scale functional networks. These hubs are useful for generating challenging candidate paths, but they are also precisely where graph-topological reasoning can become misleading. Therefore, the KG is used only for candidate generation rather than truth adjudication. KG construction statistics and a visualization are provided in Appendix B.

3.5 Path Mining and Candidate Selection

We mine multi-hop paths using breadth-first search and bidirectional simple-path search. Candidate paths are scored using structural features, including hop count, entity rarity, and type diversity. We then apply an LLM-based quality screening step to assess whether a path can instantiate one of the five reasoning paradigms.

A path is retained only if it satisfies three conditions. First, it must correspond to at least one of the five neuroscience reasoning paradigms in Table 2. Second, its key entities and relations must be traceable to evidence snippets in the literature corpus. Third, it must support contrastive option construction, meaning that the path can yield both evidence-supported claims and plausible but evidence-invalid distractors. This filtering reduces approximately $40{,}000$ candidate paths to 51 retained reasoning pathways. Detailed path-scoring criteria are provided in Appendix G.2.

3.6 Question and Option Generation

For each retained reasoning pathway, we employ an over-generation strategy in which the LLM initially generates a larger pool of candidate options, typically 7 to 10 options, together with preliminary rationales and provisional labels. The generator receives four inputs: the retained path, associated evidence snippets, the target reasoning paradigm, and the intended distractor types. This larger pool is used only to increase the diversity of plausible distractors and minimal pairs.

We then prune the pool to exactly five options per question. Options are removed if they are duplicate, poorly grounded, evidence-inconsistent, or diagnostically uninformative during construction-time screening. The retained five options form the final select-all-that-apply format. Provisional labels are used only as construction aids and are not treated as final ground truth.

The number of correct options is not disclosed to the model. This format is stricter than single-choice QA because a model must recover the full set of supported claims while rejecting all unsupported ones. Distractors are designed to be scientifically plausible rather than trivially wrong. For example, a false option may preserve correct terminology but overextend a mechanism across anatomical scales, stitch two disease mechanisms through a shared hub, or remove necessary temporal or methodological qualifiers.

After generation, all options enter the evidence calibration pipeline. The final benchmark contains 43 questions and 215 option-level claims.

3.7 Evidence Cards and Evidence-Gap Taxonomy

Each option is paired with an evidence card. An evidence card records the option-level claim, source metadata, relevant evidence snippets, anatomical or disease constraints, modality or study-design notes, and a preliminary support judgment. The purpose of the evidence card is to force auditing decisions to be grounded in explicit literature evidence rather than model priors.

To support diagnostic evaluation, we define a six-category Evidence-Gap Taxonomy:

- `causal_direction_error`: reversing or misassigning the causal chain between pathology and network abnormalities;
- `association_to_causation`: interpreting correlational evidence as interventional causality;
- `hub_stitching`: merging disease mechanisms through shared network hubs without direct evidence;
- `unsupported_mechanism`: introducing mechanistic claims that are not supported by the literature evidence;
- `granularity_inheritance_error`: transferring a mechanism across anatomical levels without evidence;
- `absolute_overclaim`: removing necessary qualifiers and turning context-dependent findings into universal claims.

This taxonomy allows false options to be analyzed as specific reasoning failures rather than undifferentiated incorrect answers. A full evidence-card schema and examples are provided in Appendix F.

Paradigm	Reasoning operation	Typical invalid extrapolation
Granularity	Determining whether a mechanism established at one anatomical scale transfers to a network, region, or subregion.	Assuming that a network-level mechanism automatically applies to all subregions.
Cross-path conflict	Reconciling multiple KG paths that share hubs but imply different mechanisms or disease interpretations.	Merging two plausible paths through a shared hub without identifying moderating conditions.
Differential	Distinguishing disorders that involve overlapping networks but differ in regional signatures, biomarkers, or pathological mechanisms.	Treating shared network involvement as evidence for a shared disease mechanism.
Causal direction	Determining whether evidence supports cause, consequence, correlation, or confounding.	Converting correlational or cross-sectional evidence into causal intervention claims.
Counterfactual	Reasoning about alternative routes, compensatory mechanisms, or the effect of removing a node or pathway.	Assuming that removing one node necessarily abolishes the downstream mechanism.

Table 2: The five neuroscience-specific reasoning paradigms used to construct BrainBench-Calibrated.

3.8 Auditing, Adjudication, and Revision

Auditors receive the option text, evidence card, and provisional label (provided as the `original_gold` field), but are explicitly instructed to independently re-decide the label rather than confirm the provisional one. Each option is independently audited by two LLM auditors across five dimensions: textual support, evidential warrant, revised gold label, gap type, and recommended action. Textual support evaluates whether the claim is entailed, partially supported, not stated, or contradicted by the evidence, while evidential warrant assesses whether the evidence is sufficient to justify the mechanistic claim. An option is considered well-founded only when both dimensions are deemed adequate.

We then apply risk stratification. Green items are adopted when auditors agree on the label and evidence sufficiency. Yellow items involve partial disagreement, such as agreement on the final label but disagreement on the gap type or revision action. Red items involve fundamental disagreement about scientific validity or high-risk evidence gaps. Yellow and Red items are adjudicated by a strong model under evidence-card constraints.

Items marked for revision undergo closed-loop rewriting. During rewriting, we apply a label-locking mechanism: the intended support status of each option is fixed, and the wording is revised until it matches the evidence card. If an option cannot be made evidence-consistent, it is replaced or removed. At the question level, the revised item must preserve a valid multi-select structure with at least one positive and one negative option.

The auditing pipeline modified the option configurations of $40/43$ questions and rewrote 88 flawed options. As shown in Table 3, inter-auditor reliability is strong across audit dimensions. These results support the consistency of our auditing process, though we characterize BrainBench-Calibrated as literature-grounded and LLM-audited rather than expert-certified.

Audit dimension	Agreement (%)	Krippendorff's α
Textual support	71.2	0.776
Evidential warrant	79.5	0.872
Revised gold	88.4	0.789
Gap type	64.7	0.792
Risk signal	74.0	0.829

Table 3: Inter-auditor agreement across evidence auditing dimensions.

4 Experiments

4.1 Experimental Setup and Metrics

We evaluate 11 LLMs in a closed-book setting. Models receive only the question stem and options, with no access to external evidence. Each model receives a question stem and five candidate options, and must output the complete set of evidence-supported options. We report exact-match accuracy (EM), option-level F1, precision, recall, false positive rate (FPR), false negative rate (FNR), and the average number of selected options (AvgSel). Detailed dataset statistics are provided in Appendix A.

EM is the primary metric because BrainBench-Calibrated is a multi-select claim verification task: a scientifically valid answer requires selecting all supported claims while rejecting unsupported or over-extended ones. A single missing condition or an endorsed unsupported mechanism can change the scientific interpretation of the answer. How-

ever, EM is strict, since one option-level error makes the whole question incorrect. We therefore complement it with option-level metrics, which reveal whether models fail by missing supported claims, endorsing unsupported claims, or adopting overly conservative or over-inclusive selection strategies.

4.2 Closed-Book Performance

Table 4 reports closed-book performance. The strongest models answer only slightly more than half of the questions exactly, indicating that recovering the full evidence-grounded option configuration remains difficult. The top models have substantially overlapping Wilson confidence intervals, so we treat the ranking as descriptive rather than statistically definitive.

The option-level metrics reveal patterns that EM alone would hide. GPT-5.4 obtains the strongest option-level F1, despite not having the highest EM, suggesting that it makes fewer option-level errors overall but less often recovers the exact full configuration. In contrast, MiniMax M2.7, Gemini, Qwen, GLM, and MiMo show conservative behavior: they maintain relatively high precision and low FPR, but miss many supported claims. GPT-4o and DeepSeek show the opposite pattern, with higher recall but much higher FPR and AvgSel. Thus, failures on BrainBench-Calibrated are not uniform: some models under-select warranted mechanisms, whereas others over-endorse plausible but unsupported claims.

4.3 Evidence-Gap False Endorsement Analysis

Figure 2 provides a diagnostic view of false positive errors. Instead of counting all errors together, we normalize by the number of negative options in each evidence-gap type. This is important because different gap types have different base frequencies.

The heatmap shows that model errors are not dominated by simple causal inversion. Explicit causal-direction errors and association-to-causation errors have relatively lower false endorsement rates. More characteristic failures involve preserving the boundaries of valid mechanistic claims. Hub stitching is especially salient, although the category contains few examples and should be interpreted cautiously. Granularity inheritance and absolute overclaiming also appear repeatedly across models, indicating that LLMs often transfer mechanisms across anatomical levels or remove necessary contextual qualifiers. These patterns support the motivation for neuroscience-specific evaluation: models may handle causal vocabulary reasonably well while still failing to preserve anatomical hierarchy, disease specificity, and evidential scope.

4.4 Evidence-Enhanced Prompting

To distinguish missing knowledge from reasoning failure, we conduct an evidence-enhanced prompting experiment on four representative models. In this setting, models receive the same question and options plus relevant evidence snippets from the evidence cards. Gold labels and gap-type annotations are not provided.

Table 5 shows that evidence improves several models substantially, especially GPT-5.4 and Qwen 3.6-Plus. This indicates that closed-book performance partly reflects missing or inaccessible literature knowledge. However, the gains are uneven: some models improve only modestly, suggesting that evidence access alone does not guarantee correct mechanistic verification. Additional results, including the composite ranking, paradigm-level radar plot, and detailed evidence-enhanced prompting tables, are provided in Appendix E.

Overall, the evidence-enhanced experiment suggests that BrainBench-Calibrated probes both knowledge access and evidence-grounded reasoning: providing evidence helps, but models must still determine whether the evidence licenses a claim under the relevant anatomical, disease, and methodological constraints.

5 Conclusion

We introduced BrainBench-Calibrated, a literature-grounded diagnostic benchmark for mechanistic reasoning in connectome neuroscience. The benchmark is organized around five domain-specific reasoning paradigms: anatomical granularity, cross-path conflict, cross-disease differential reasoning, causal direction, and counterfactual intervention. These paradigms operationalize how neuroscientific mechanisms are constrained by anatomical hierarchy, disease context, evidence design, and intervention assumptions.

Methodologically, BrainBench-Calibrated decouples structural plausibility from truth adjudication. Knowledge graphs are used to gener-

Model	EM raw	EM (%)	Wilson 95% CI	Option-F1	Precision	Recall	AvgSel	FPR	FNR
Claude Opus 4.7	25/43	58.1	[43.3, 71.6]	67.6	73.8	62.3	1.5	12.3	37.7
OpenAI GPT-5.4	23/43	53.5	[38.9, 67.5]	70.4	76.9	64.9	1.5	10.9	35.1
Claude Sonnet 4.6	23/43	53.5	[38.9, 67.5]	66.7	75.4	59.7	1.4	10.9	40.3
Google Gemini Pro 3.1	23/43	53.5	[38.9, 67.5]	65.2	78.2	55.8	1.3	8.7	44.2
xAI Grok 4.20	20/43	46.5	[32.5, 61.1]	66.2	72.3	61.0	1.5	13.0	39.0
MiniMax M2.7	20/43	46.5	[32.5, 61.1]	67.2	84.3	55.8	1.2	5.8	44.2
DeepSeek V3.2	15/43	34.9	[22.4, 49.8]	61.4	57.3	66.2	2.1	27.5	33.8
Xiaomi MiMo V2 Pro	15/43	34.9	[22.4, 49.8]	56.2	77.3	44.2	1.0	7.2	55.8
Qwen 3.6-Plus	15/43	34.9	[22.4, 49.8]	55.7	75.6	44.2	1.1	8.0	55.8
Z.AI GLM-5.1	15/43	34.9	[22.4, 49.8]	60.8	79.2	49.4	1.1	7.2	50.6
OpenAI GPT-4o	12/43	27.9	[16.7, 42.7]	67.0	58.1	79.2	2.4	31.9	20.8

Table 4: Closed-book multi-select performance on BrainBench-Calibrated. EM is exact-match accuracy over 43 questions. Option-level metrics are micro-averaged over 215 claims. FPR and FNR are computed over negative and positive option labels, respectively.

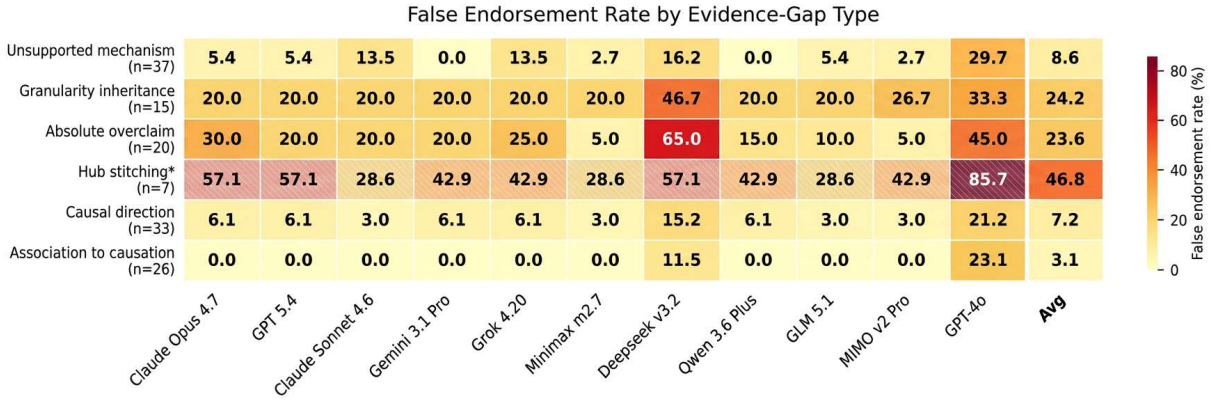

Figure 2: False endorsement rate (FER) by model and evidence-gap type. FER is the fraction of negative options of a given gap type selected by a model. This normalization controls for different base frequencies of gap types. Categories marked with * have low base frequency (n<10) and should be interpreted cautiously.

Model	Closed-book EM	+Evidence	Δ
OpenAI GPT-5.4	53.5	72.1	+18.6
Qwen 3.6-Plus	34.9	55.8	+20.9
xAI Grok 4.20	46.5	46.5	+0.0
DeepSeek V3.2	34.9	41.9	+7.0

Table 5: Exact-match Pass@1 before and after evidence-enhanced prompting. Models receive relevant evidence snippets but not gold labels or gap-type annotations.

ate candidate reasoning paths, while final labels are assigned relative to primary-literature evidence cards and a six-category Evidence-Gap Taxonomy. This design turns KG-induced plausibility into a source of challenging distractors without allowing graph connectivity to determine correctness.

Empirically, closed-book evaluation shows that current LLMs are not yet robust on full-configuration mechanistic reasoning. The best model answers 25/43 questions exactly, and confidence intervals overlap among top models. Option-level results reveal different failure regimes: some models are conservative and miss supported claims, while others are over-inclusive and endorse unsupported claims. Evidence-enhanced prompting improves several models, especially on granularity and differential reasoning, but does not eliminate the need to judge whether evidence licenses a claim under specific constraints.

For neuroscience, the main contribution of this work is not a leaderboard, but a structured evaluation framework for asking whether AI systems can preserve the conditional boundaries of mechanistic claims. Our results suggest that future scientific LLM evaluation should explicitly test anatomical granularity, disease specificity, causal evidential scope, and counterfactual validity.

Limitations

BrainBench-Calibrated has several limitations. First, the current release is intentionally small, with 43 questions and 215 option-level claims. It is suitable for diagnostic analysis but not for stable leaderboard ranking. The distribution across paradigms is also uneven, especially for Cross-path questions, so paradigm-level results should be interpreted descriptively.

Second, the benchmark is literature-grounded and LLM-audited rather than expert-certified. Independent auditing and evidence-card adjudication improve consistency, but they do not replace external neuroscience expert validation or human baseline evaluation.

Third, labels are evidence-relative. A negative option indicates contradiction, overgeneralization, directional invalidity, or insufficient support within the curated evidence set, rather than universal falsity across all neuroscience literature.

Fourth, closed-book performance conflates knowledge access and reasoning ability. Our evidence-enhanced prompting experiment partially addresses this issue, but it covers only a subset of models and may be sensitive to prompt format and evidence selection.

Finally, the benchmark focuses on English-language macroscale connectome neuroscience. Its findings may not generalize to cellular neuroscience, molecular neuroscience, non-English literature, or broader biomedical reasoning domains. Future versions should expand scale, include expert validation, diversify neuroscience subfields, and support evidence updates as the literature evolves.

Ethical Statement

BrainBench-Calibrated is intended solely for research evaluation of scientific reasoning systems and should not be used for clinical diagnosis, treatment recommendation, or as a measure of clinical competence. The benchmark is constructed from published literature and does not contain patient-level identifiable data. Because the labels are evidence-relative rather than expert-certified medical truth, model performance should be interpreted only as performance on evidence-grounded scientific reasoning tasks.

We will release benchmark questions, option-level labels, evidence-card metadata, prompts, and evaluation scripts for research use. Original source papers are not redistributed; evidence cards contain only short evidence snippets and source metadata. Users are responsible for complying with the access conditions and copyright policies of the original publications.

References

Owen Bianchi, Maya Willey, Chelsea X Alvarado, Benjamin Danek, Marzieh Khani, Nicole Kuznetsov, Anant Dadu, Syed Shah, Mathew J Koretsky, Mary B Makarious, Cory Weller, Kristin S Levine, Sungwon Kim, Paige Jarreau, Dan Vitale, Elise Marsan, Hirotaka Iwaki, Hampton Leonard, Sara Bandres-Ciga, and 5 others. 2026. Cardbiomedbench: a benchmark for evaluating the performance of large language models in biomedical research. *The Lancet Digital Health*, 8(1):100943.

Yongrui Chen, Zhiqiang Liu, Jing Yu, Lin Ren, Nan Hu, Xinbang Dai, Jiajun Liu, Jiazhen Kang, Shenyu Zhang, Xinda Wang, Keyan Ding, Pengfei Shen, Haolei Zhu, Hongjie Deng, Yisong Wang, Tongtong Wu, Sheng Bi, Wen Zhang, Tianxing Wu, and 5 others. 2025. Oneeval: Benchmarking llm knowledge-intensive reasoning over diverse knowledge bases. *arXiv preprint arXiv:2506.12577*.

Alex Fornito, Andrew Zalesky, and Michael Breakspear. 2015. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172.

Panagiotis Fotiadis, Linden Parkes, Kathryn A. Davis, Theodore D. Satterthwaite, Russell T. Shinohara, and Dani S. Bassett. 2024. Structure–function coupling in macroscale human brain networks. *Nature Reviews Neuroscience*, 25(10):688–704.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z Pan. 2024. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge. *arXiv preprint arXiv:2412.17032*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Yunsoo Kim, Yusuf Abdulle, and Honghan Wu. 2025. Biohopr: A benchmark for multi-hop, multi-answer reasoning in biomedical domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12894–12908.

Goh Kobayashi, Kosei Hirata, Maiko Ono, Kensaku Kasuga, and Yuhei Takado. 2026. The 2024 nia-aa biological definition of alzheimer's disease: linking biomarkers to clinical practice. *Frontiers in Dementia*, 5:1736297.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K. Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O. Cohen, Valentina Borghesani, Anton Pashkov, Daniele Marinazzo, Jonathan Nicholas, Alessandro Salatiello, Ilia Sucholutsky, Pasquale Minervini, Sepehr Razavi, Roberta Rocca, Elkhan Yusifov, Tereza Okalova, and 20 others. 2024. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9(2):305–315.

Elan Markowitz, Krupa Galiya, Greg Ver Steeg, and A. G. Galstyan. 2025. Kg-llm-bench: A scalable benchmark for evaluating llm reasoning on textualized knowledge graphs. *arXiv e-prints*.

Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Daniel Keitley, Maria Dermit, and Julien Fauqueur. 2025. Grounding QA generation in knowledge graphs and literature: A scalable LLM framework for scientific discovery. In *Will Synthetic Data Finally Solve the Data Access Problem?*

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Caio Seguin, Olaf Sporns, and Andrew Zalesky. 2023. Brain network communication: concepts, models and applications. *Nature Reviews Neuroscience*, 24(9):557–574.

Yuan Sui, Yufei He, Zifeng Ding, and Bryan Hooi. 2025. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12685–12701.

Martijn P. van den Heuvel and Olaf Sporns. 2019. A cross-disorder connectome landscape of brain dysconnectivity. *Nature Reviews Neuroscience*, 20(7):435–446.

Jacob W. Vogel, Nick Corriveau-Lecavalier, Nicolai Franzmeier, Joana B. Pereira, Jesse A. Brown, Anne Maass, Hugo Botha, William W. Seeley, Danielle S. Bassett, David T. Jones, and Michael Ewers. 2023. Connectome-based modelling of neurodegenerative diseases: towards precision medicine and mechanistic insight. *Nature Reviews Neuroscience*, 24(10):620–639.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Benchmark Statistics

Table 6 provides the overall statistics of BrainBench-Calibrated. The benchmark contains 43 multi-select questions and 215 option-level claims. Positive options denote claims supported by the curated evidence set, whereas negative options denote claims that are contradicted, over-generalized, directionally invalid, or insufficiently supported by the evidence.

Metric	Value
Final questions	43
Final option-level claims	215
Positive options	77
Negative options	138
Avg. correct options / question	1.79
1 correct option questions	14
2 correct option questions	24
3 correct option questions	5
Granularity questions	8
Cross-path questions	2
Differential questions	9
Causal-direction questions	16
Counterfactual questions	8
Options per question	5

Table 6: Overall dataset statistics for BrainBench-Calibrated.

Table 7 reports the distribution of questions and option-level claims across the five reasoning paradigms. The distribution is intentionally diagnostic rather than uniformly balanced. In particular, Cross-path questions are limited to 2 items, so paradigm-level results for this category should be interpreted descriptively.

Paradigm	#Q	#Opt	Pos	Neg	AvgCorr/Q
Granularity	8	40	16	24	2.00
Cross-path	2	10	2	8	1.00
Differential	9	45	17	28	1.89
Causal-direction	16	80	31	49	1.94
Counterfactual	8	40	11	29	1.38

Table 7: Statistics of different reasoning paradigms in the benchmark.

Table 8 shows how evidence-gap types are distributed across paradigms. This table complements the false endorsement heatmap in the main paper by showing the base frequency of each gap type.

B Construction Details

B.1 KG Construction Summary

Table 9 summarizes the construction pipeline. The large reduction from candidate paths to retained reasoning pathways reflects the diagnostic nature of the benchmark: paths are retained only when they can instantiate one of the five reasoning paradigms, are traceable to evidence snippets, and support contrastive option construction.

B.2 KG Visualization

Figure 3 visualizes the extracted connectome neuroscience KG. The graph contains hub-like disease and network nodes, which are useful for generating challenging candidate paths but risky if treated as evidence of scientific validity. This motivates our KG-as-Candidate-Generator design.

B.3 Path Retention Criteria

A candidate path is retained only if it satisfies three criteria. First, it must instantiate one of the five reasoning paradigms defined in the main paper. Second, its key entities and relations must be traceable to evidence snippets from the literature corpus. Third, it must support contrastive question construction, meaning that it can yield both evidence-supported claims and plausible but evidence-invalid distractors.

We use the retained paths only as structural scaffolds. They do not determine the final correctness of any option. All labels are assigned after evidence-card calibration.

C Evidence Calibration and Auditing Details

C.1 Risk Stratification

Table 10 summarizes the risk stratification protocol used during evidence auditing.

C.2 Revision Statistics

The auditing pipeline modified the option configurations of $40/43$ questions and rewrote 88 flawed options. Table 11 summarizes the main revision statistics.

D Experimental Details

D.1 Metric Definitions

For each question, exact match (EM) is 1 if the predicted option set exactly matches the gold option set and 0 otherwise. Option-level metrics are

Error Type	Gran.	Cross	Diff.	Causal	Counter.	Total
Unsupported mechanism	1	4	15	5	12	37
Granularity inheritance error	15	0	0	0	0	15
Absolute overclaim	0	2	5	1	12	20
Hub stitching	3	0	3	0	1	7
Causal direction error	0	2	3	24	4	33
Association to causation	5	0	2	19	0	26

Table 8: Distribution of evidence-gap types across reasoning paradigms.

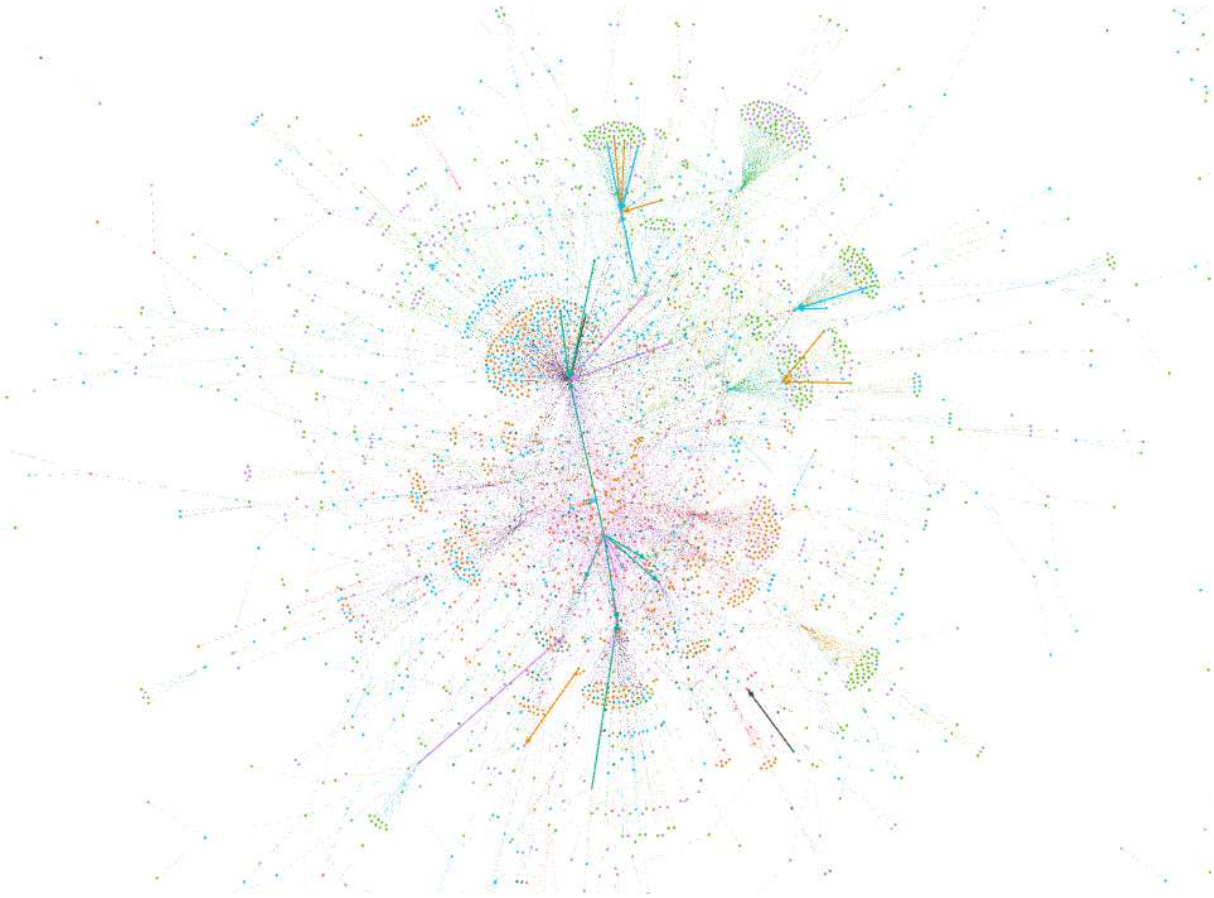

Figure 3: Visualization of the extracted connectome neuroscience KG.

Construction stage	Count
Anchor review papers	6
Expanded literature pool	255 papers
Extracted triplets after filtering	4,371
KG nodes after normalization	3,665
Candidate paths mined	∼40,000
Retained reasoning pathways	51
Final questions	43
Final option-level claims	215

Table 9: Summary of construction counts.

micro-averaged over all 215 option-level claims. Precision measures the fraction of selected options that are correct, while recall measures the fraction of positive options selected by the model. FPR is computed over negative options, and FNR is computed over positive options. AvgSel denotes the average number of selected options per question.

Pass@3 is computed by sampling three model responses under the same prompt format and checking whether at least one response exactly matches the gold option set. When a model produces malformed outputs, we apply a deterministic parser that extracts option letters and discards unsupported text. If no valid option letter is detected, the answer is treated as an empty set.

Risk level	Condition	Action
Green	Auditors agree on support status, evidential warrant, and revised gold label.	Adopt the option label directly.
Yellow	Auditors agree on the core label but disagree on gap type, evidence sufficiency, or revision action.	Send to adjudication and revise if needed.
Red	Auditors disagree on scientific validity, or the option involves high-risk gaps such as causal direction or hub stitching.	Send to strong-model adjudication and apply evidence-card constrained revision.

Table 10: Risk stratification protocol for option-level evidence auditing.

Audit / revision statistic	Value
Final option-level claims	215
Claims requiring adjudication	129
Questions with modified option configurations	40
Rewritten or calibrated options	88
Final retained questions	43

Table 11: Summary of evidence-auditing and revision statistics.

E Additional Experimental Results

E.1 Exploratory Composite Ranking

Figure 4 shows an exploratory composite ranking over closed-book metrics. This figure is not used as the primary evaluation result because the weights are necessarily subjective. We include it only as a compact visualization of the trade-off between exact-match accuracy, option-level quality, and error rates.

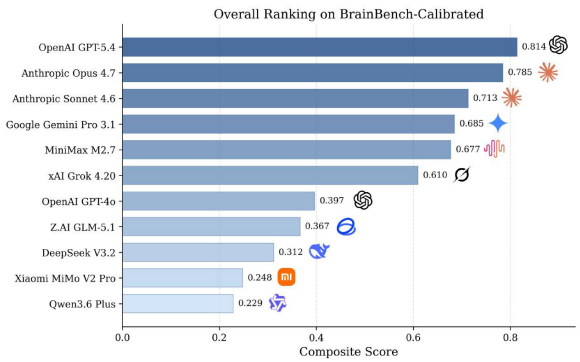

Figure 4: Exploratory composite ranking of closed-book multi-select performance. Metrics are min-max normalized, and lower-is-better metrics such as FPR and FNR are inverted. This ranking is provided only as a supplementary summary.

Let $x_{m,k}$ denote the value of metric k for model m. For higher-is-better metrics, we min-max normalize $x_{m,k}$ so that larger values indicate better performance. For lower-is-better metrics, such as FPR and FNR, we invert the normalized score. The composite score is then computed as a weighted sum, with larger weights assigned to EM and Option-F1. The score is computed as $S_m = 0.35\widetilde{\mathrm{EM}}_m + 0.25\widetilde{\mathrm{OptionF1}}_m + 0.15\widetilde{\mathrm{Precision}}_m + 0.15\widetilde{\mathrm{Recall}}_m + 0.05\widetilde{\mathrm{FPR}}_m + 0.05\widetilde{\mathrm{FNR}}_m$. Because different weighting schemes may produce different rankings, the main paper reports raw metrics rather than relying on this composite score.

E.2 Paradigm-Level Closed-Book Performance

Figure 5 shows the closed-book performance of leading models across the five reasoning paradigms. This visualization is descriptive. Because Cross-path contains only 2 questions, performance on this axis should not be interpreted as a stable estimate.

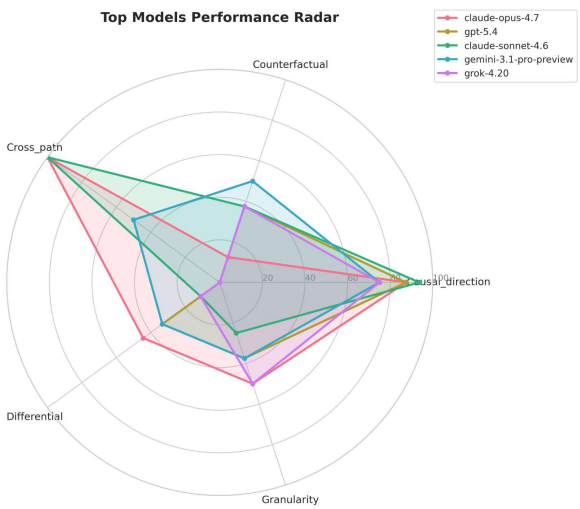

Figure 5: Descriptive performance radar of leading LLMs across five mechanistic reasoning paradigms. The Cross-path axis is based on only 2 questions and should be interpreted cautiously.

E.3 Detailed Evidence-Enhanced Prompting Results

Table 12 reports Pass@1 and Pass@3 before and after evidence-enhanced prompting. Models receive relevant evidence snippets but not gold labels or evidence-gap annotations.

Table 13 reports evidence-enhanced prompting results by reasoning paradigm. The table is intended as a diagnostic breakdown rather than a statistically stable per-paradigm comparison.

F Representative Benchmark Item

This section presents five representative examples, one for each reasoning paradigm in BrainBench-Calibrated. Each example illustrates the transformation from KG-derived candidate relations to calibrated multi-select mechanistic reasoning questions through option-level evidence auditing. The examples are representative benchmark instances with implementation-specific metadata omitted for readability; their stems, option labels, and audit rationales follow the same evidence-calibration procedure as the released items.

F.1 Causal-Direction Example: DMN Connectivity and Alzheimer's Disease Dementia

Table 14 summarizes the evidence card used to construct this item, including the KG seed relation, the supporting evidence excerpt, and the dispute signals that motivate the causal-direction probe. Building on this scaffold, we formulate a multi-select question (stem below) whose options are individually adjudicated against the evidence; the resulting option-level audit, including gap types and rationales, is reported in Table 15.

Question stem. A research group publishes two studies examining the relationship between default mode network (DMN) connectivity and Alzheimer's disease dementia (AD dementia).

Study 1 is cross-sectional. Using resting-state fMRI in AD dementia patients and age-matched controls, the investigators report decreased DMN connectivity in the posterior cingulate cortex and medial temporal lobe, but increased connectivity in medial prefrontal regions. The database label assigned to this observation is `network_altered_in_disease`.

Study 2 is longitudinal. In cognitively normal older adults followed for eight years, baseline amyloid-β burden in posterior cingulate and medial temporal regions predicts later DMN connectivity decline and subsequent conversion to AD dementia. Baseline DMN connectivity, however, does not significantly predict later AD conversion.

Which of the following statements about the mechanistic relationship between DMN dysfunction and AD dementia are well-founded?

A. Study 1's regionally heterogeneous DMN connectivity changes constitute direct evidence that DMN dysfunction causally drives AD dementia progression, because regional specificity rules out a nonspecific disease effect on brain networks.

B. Study 2's finding that baseline amyloid-β burden predicts later DMN connectivity decline, while baseline DMN connectivity does not predict AD conversion, supports the interpretation that AD-related pathology disrupts DMN function rather than DMN dysfunction initiating AD pathology.

C. The label `network_altered_in_disease` is inconsistent with interpreting Study 1 as evidence for a DMN $\rightarrow$ AD causal direction, because the label semantically implies that the disease state modifies the network.

D. Clinical AD dementia onset directly drives the initial preferential deposition of amyloid-β within DMN hubs, which then induces the characteristic metabolic hyperactivity observed across these networks.

E. A cross-sectional study using simultaneous fMRI and tau PET provides stronger evidence for the causal direction of DMN dysfunction in AD than longitudinal amyloid tracking, because tau pathology is more closely linked to neurodegeneration.

Gold labels. The evidence-supported options are **B** and **C**.

F.2 Counterfactual Example: Removing α-Synuclein

The evidence card and option-level audit for this item are shown in Table 16 and Table 17, respectively.

Question stem. Consider the following thought experiment. A research team generates an adult conditional knockout of α-synuclein using

Model	N	Baseline Pass@1 (%)	Evidence Pass@1 (%)	Δ Pass@1	Baseline Pass@3 (%)	Evidence Pass@3 (%)
OpenAI GPT-5.4	43	53.5	72.1	+18.6	60.5	76.7
xAI Grok 4.20	43	46.5	46.5	+0.0	48.8	55.8
Deepseek V3.2	43	34.9	41.9	+7.0	48.8	55.8
Qwen 3.6-Plus	43	34.9	55.8	+20.9	46.5	62.8

Table 12: Performance comparison between baseline prompting and evidence-enhanced prompting.

Paradigm	N	GPT-5.4			Grok-4.20			DeepSeek V3.2			Qwen 3.6-Plus		
		Base	Evid.	Δ	Base	Evid.	Δ	Base	Evid.	Δ	Base	Evid.	Δ
Causal-direction	16	87.5	87.5	+0.0	75.0	62.5	-12.5	56.2	50.0	-6.2	37.5	31.2	-6.2
Counterfactual	8	37.5	50.0	+12.5	37.5	12.5	-25.0	12.5	12.5	+0.0	50.0	50.0	+0.0
Cross-path	2	0.0	0.0	+0.0	0.0	0.0	+0.0	50.0	0.0	-50.0	50.0	100.0	+50.0
Differential	9	33.3	77.8	+44.4	11.1	33.3	+22.2	22.2	44.4	+22.2	33.3	66.7	+33.3
Granularity	8	37.5	75.0	+37.5	50.0	75.0	+25.0	25.0	62.5	+37.5	12.5	87.5	+75.0

Table 13: Pass@1 before and after evidence augmentation across paradigms.

Field	Value
KG seed relation	`default mode network` → `Alzheimer's disease dementia`
Entity types	BrainNetwork → Disease
Extracted relation	`network_altered_in_disease`
Reasoning paradigm	Causal direction
Evidence excerpt	"The pattern of connectivity changes in Alzheimer's disease dementia differ between these subsystem regions, with decreased DMN connectivity in the posterior cingulate and medial temporal lobe and increases found in medial prefrontal regions."
Dispute signals	`correlation_only`; `confounder_possible`
Potential confounders	amyloid-β deposition; posterior cingulate cortex; medial temporal lobe; medial prefrontal cortex
Core ambiguity	The evidence supports altered DMN connectivity in AD dementia, but does not by itself establish whether DMN dysfunction causes AD dementia or whether AD-related pathology disrupts DMN connectivity.

Table 14: Evidence-card summary for a representative causal-direction item. The KG relation is used only as a candidate scaffold; option labels are adjudicated relative to the evidence and the study designs described in the question.

Option	Label	Gap type	Audit rationale
A	Not supported	`association_to_causation`	Study 1 is cross-sectional and establishes association, not causal direction. Regional specificity does not rule out disease-driven network disruption.
B	Supported	–	Study 2 provides temporal precedence: amyloid-β burden predicts later DMN decline, while baseline DMN connectivity does not predict AD conversion. This supports an AD-pathology → DMN-disruption interpretation.
C	Supported	–	The relation label `network_altered_in_disease` semantically implies that the disease modifies the network. This conflicts with using the label as evidence for a DMN → AD direction.
D	Not supported	`causal_direction_error`	The option reverses the pathophysiological timeline by claiming that clinical AD dementia onset drives initial amyloid-β deposition in DMN hubs.
E	Not supported	`causal_direction_error`	Although tau PET is relevant to neurodegeneration, cross-sectional multimodal imaging does not establish temporal precedence and therefore does not provide stronger directional evidence than longitudinal tracking.

Table 15: Option-level evidence audit for the representative causal-direction item. The example tests whether models distinguish cross-sectional association, longitudinal temporal precedence, semantic directionality in relation labels, and biomarker relevance.

a tamoxifen-inducible Cre-lox system. The manipulation achieves more than 95% reduction of

Field	Value
Removed node	α-synuclein
Node type	Mechanism
Counterfactual manipulation	Adult conditional knockout of α-synuclein followed by longitudinal assessment of motor networks, dopaminergic neurons, histopathology, and proteomic changes.
Primary downstream effect	Motor-related resting-state networks
Secondary disease context	Parkinson's disease and parkinsonian neurodegeneration
Candidate alternative pathways	Tau pathology, cortical atrophy, neurodegeneration, synaptic dysfunction, pathology propagation, and retrograde axonal damage.
Reasoning paradigm	Counterfactual reasoning
Calibration note	The raw KG candidates contained duplicate downstream effects with inconsistent dependency ratings. During calibration, we treat KG paths only as candidate scaffolds and adjudicate final option labels relative to evidence about α-synuclein's physiological synaptic role and pathological aggregation role.
Core reasoning challenge	The item tests whether models can distinguish preserved neuron counts from preserved functional connectivity, and whether they can separate the existence of alternative neurodegenerative pathways from their sufficiency to recapitulate classical synuclein-driven Parkinsonian mechanisms.

Table 16: Evidence-card summary for a representative counterfactual item. The KG proposes dependency relations involving α-synuclein, motor-related resting-state networks, and Parkinson's disease. The final labels are determined by evidence-calibrated reasoning rather than raw KG dependency ratings.

α-synuclein protein across the brain within two weeks. The mice are then aged for twelve months and assessed using resting-state fMRI, behavioral motor testing, histopathology, and proteomic analysis.

At follow-up, the knockout mice show largely preserved dopaminergic neuron counts in the substantia nigra pars compacta, with only a small reduction relative to wild-type littermates. However, functional connectivity analysis reveals alterations in several motor-related resting-state networks.

Based on this scenario and mechanistic reasoning about α-synuclein's physiological synaptic role and pathological role in neurodegeneration, which of the following statements are well-founded?

A. Measurable alterations in motor-related resting-state networks would be mechanistically plausible even with largely preserved dopaminergic neuron counts, because α-synuclein contributes to presynaptic vesicle cycling and SNARE-complex regulation at nigrostriatal terminals, and compensation by β- or γ-synuclein cannot be assumed to be complete.

B. The near-complete preservation of dopaminergic neurons confirms that α-synuclein is not necessary for Parkinson's disease pathogenesis, because tau-mediated pathways can independently and fully recapitulate all cardinal features of Parkinson's disease, including the classical pattern of ventrolateral substantia nigra vulnerability.

C. Motor-related resting-state networks should remain largely intact because the absence of α-synuclein triggers a compensatory structural reorganization of SNARE complexes that stabilizes nigrostriatal functional connectivity.

D. In α-synuclein knockout mice crossed with a P301S tau line, motor deficits would emerge because tau oligomers structurally substitute for α-synuclein in SNARE complexes, initiating aberrant nigrostriatal synaptic vesicle release.

E. The cortical atrophy to tau pathology cascade dictates accelerated and uniform dopaminergic cell loss across all substantia nigra tiers within three months, functioning as the primary and sufficient driver of nigral neurodegeneration.

Gold label. The evidence-supported option is **A**.

F.3 Cross-Path Example: DMN Hyperconnectivity and Hypoconnectivity in Alzheimer's Disease

The evidence card and option-level audit for this item are shown in Table 18 and Table 19, respectively.

Question stem. A research consortium conducts a longitudinal study combining amyloid-PET, tau-PET, and resting-state fMRI in cognitively unim-

Option	Label	Gap type	Audit rationale
A	Supported	–	Preserved dopaminergic neuron counts do not guarantee preserved functional connectivity. α-synuclein has a physiological role in presynaptic vesicle cycling and SNARE-complex regulation, so motor-network alterations remain mechanistically plausible even without major neuronal loss.
B	Not supported	`absolute_overclaim`	The option over-interprets neuron preservation as showing that α-synuclein is unnecessary for Parkinson's disease pathogenesis. Alternative tau-related pathways may contribute to parkinsonian phenotypes, but this does not imply that they fully recapitulate classical synuclein-driven Parkinson's disease.
C	Not supported	`unsupported_mechanism`	The proposed compensatory structural reorganization of SNARE complexes is not established by the evidence card. It also assumes that neuron preservation is sufficient for normal resting-state connectivity.
D	Not supported	`unsupported_mechanism`	The option introduces an unsupported molecular mechanism: tau oligomers are not established to structurally substitute for α-synuclein in SNARE complexes to drive nigrostriatal vesicle-release abnormalities.
E	Not supported	`absolute_overclaim`	The option asserts a uniform, accelerated, and sufficient tau-driven mechanism for dopaminergic cell loss. This removes necessary temporal, anatomical, and mechanistic qualifications.

Table 17: Option-level audit for the representative counterfactual item. The item tests whether models can reason about the counterfactual removal of a mechanism while avoiding unsupported claims about compensation, alternative pathways, and pathway sufficiency.

Field	Value
Shared network	Default mode network (DMN)
Disease context	Alzheimer's disease continuum
Path A	Reduced functional connectivity $\rightarrow$ amyloid-β pathology $\rightarrow$ tau pathology $\rightarrow$ DMN $\rightarrow$ Alzheimer's disease dementia
Path B	Hyperconnectivity $\rightarrow$ amyloid-β pathology $\rightarrow$ DMN $\rightarrow$ Alzheimer's disease dementia
Conflict type	Mechanism conflict: opposite DMN connectivity directions in the same disease context
Core reconciliation	A stage-dependent model: early amyloid-positive, tau-negative individuals may show DMN hyperconnectivity, whereas later tau-positive or symptomatic individuals may show DMN hypoconnectivity.
Key moderators	Disease stage, amyloid burden, tau positivity and spread, DMN subregion, cognitive reserve, APOE genotype, and connectivity measurement method.
Reasoning paradigm	Cross-path conflict reasoning
Calibration note	The KG provides two structurally plausible but apparently conflicting paths. The final item tests whether models can reconcile them using biomarker stage rather than collapsing the conflict into a single universal mechanism.

Table 18: Evidence-card summary for a representative cross-path conflict item. The two KG-derived paths encode opposite DMN connectivity patterns in Alzheimer's disease. The calibrated question tests whether models can resolve this conflict through stage-dependent mechanistic reasoning.

paired individuals and individuals with mild cognitive impairment due to Alzheimer's disease. Participants are followed for five years with annual imaging. At baseline, participants are stratified by amyloid status and tau status. Resting-state fMRI analyses focus on within-network functional connectivity of the default mode network, measured using static seed-based correlation from the posterior cingulate cortex.

The study reports the following findings:

- **Finding 1:** At baseline, cognitively unimpaired Aβ^+/tau$^-$ individuals show higher DMN connectivity than cognitively unimpaired Aβ^-/tau$^-$ individuals.
- **Finding 2:** At baseline, MCI Aβ^+/tau$^+$ individuals show lower DMN connectivity than cognitively unimpaired Aβ^-/tau$^-$ individuals.
- **Finding 3:** Over five years, cognitively unimpaired Aβ^+/tau$^-$ individuals who convert to Aβ^+/tau$^+$ status show progressive DMN connectivity decline, with connectivity falling below normal levels after tau positivity is detected.
- **Finding 4:** Among cognitively unimpaired

Aβ^+/tau$^-$ individuals who remain tau-negative, elevated DMN connectivity persists but does not increase further.

Which of the following interpretive statements are best supported by the pattern of findings? Select all that apply.

A. The persistence of elevated DMN connectivity in Aβ^+/tau$^-$ individuals proves that preserved structural integrity directly generates hyperconnectivity and that the hyperconnectivity is a protective compensation against later neurodegeneration.

B. Finding 1 is best explained by a specific astrocytic glutamate-release mechanism through connexin hemichannels in posterior DMN hubs, which elevates local activity before synaptic loss.

C. The reduced DMN connectivity in MCI Aβ^+/tau$^+$ individuals shows that tau pathology is both necessary and sufficient for DMN hypoconnectivity, independent of amyloid burden.

D. The findings support a stage-dependent reconciliation: early Aβ^+/tau$^-$ status is associated with elevated DMN connectivity, whereas later tau positivity is associated with progressive connectivity decline. Persistent hyperconnectivity in tau-negative individuals is compatible with models in which elevated network activity may contribute to pathological vulnerability or later tau seeding.

E. The findings suggest that early reductions in functional connectivity directly initiate amyloid accumulation, which then promotes tau aggregation and cognitive decline.

Gold label. The evidence-supported option is **D**.

F.4 Granularity Example: Region-Specific Cortical Associations in Alzheimer's Disease

This section presents a representative example from the anatomical granularity paradigm. Table 20 summarizes the evidence-card constraints and reasoning risks used to construct the item, while Table 21 provides the option-level audit rationale.

Question stem. A research group conducts a multimodal neuroimaging study in patients with early-stage Alzheimer's disease. Resting-state fMRI identifies reduced functional connectivity within the default mode network. Lesion-network mapping further shows that episodic memory deficits are associated with damage to regions whose lesion maps converge on a cortical network distinct from the default mode network. Structural MRI reveals cortical thinning mainly in dorsolateral association cortex and temporal neocortical regions lateral to the hippocampal formation, while midline parietal structures are relatively preserved.

A separate schizophrenia cohort undergoes the same imaging protocol. In this cohort, the medial parietal region shows default-mode dysconnectivity but no involvement in the lesion-derived episodic memory network. By contrast, the dorsolateral parietal region participates in both the cingulo-opercular network and the lesion-derived episodic memory network.

Based on these findings and known region-specific network and disease associations, select ALL statements that are correctly supported at the appropriate level of anatomical specificity.

A. The lateral temporal cortex is associated with Alzheimer's disease and participates in the lesion-derived episodic memory network, consistent with the convergence of lesion maps onto an episodic memory circuit in the Alzheimer's disease cohort.

B. The medial parietal cortex is associated with Alzheimer's disease because it belongs to the default mode network, which shows reduced connectivity in the Alzheimer's disease cohort.

C. The lateral prefrontal cortex functions as a critical hub within the lesion-derived episodic memory network, which provides a shared mechanistic basis for memory impairments in Alzheimer's disease and Parkinson's disease dementia.

D. The altered functional connectivity observed between the lateral parietal cortex and the cingulo-opercular network in early Alzheimer's disease directly causes the progressive episodic memory decline characteristic of the pathology.

Option	Label	Gap type	Audit rationale
A	Not supported	`causal_direction_error`	The findings show persistent hyperconnectivity in Aβ^+/tau$^-$ individuals, but they do not establish that preserved structural integrity directly generates hyperconnectivity or that hyperconnectivity is protective. The option imposes an unsupported causal and compensatory interpretation.
B	Not supported	`unsupported_mechanism`	Amyloid-associated hyperconnectivity may be related to excitatory dysregulation, but the specific astrocytic connexin-hemichannel mechanism is not established by the described findings. The option introduces an overly specific molecular mechanism not warranted by the evidence.
C	Not supported	`absolute_overclaim`	The data associate tau positivity with reduced DMN connectivity, but they do not prove that tau is necessary and sufficient for hypoconnectivity or that the effect is independent of amyloid burden.
D	Supported	–	This option reconciles the apparently conflicting paths by using disease stage and tau status. It matches the observed pattern: early Aβ^+/tau$^-$ individuals show elevated connectivity, while individuals who become tau-positive show subsequent connectivity decline.
E	Not supported	`causal_direction_error`	This option reverses the temporal interpretation. The described findings show elevated connectivity in early amyloid-positive stages and reduced connectivity later with tau positivity; they do not show that early hypoconnectivity initiates amyloid accumulation.

Table 19: Option-level audit for the representative cross-path conflict item. The item tests whether models can reconcile apparently contradictory DMN hyperconnectivity and hypoconnectivity findings through biomarker stage and disease progression, rather than selecting a single context-free mechanism.

Field	Value
Candidate region group	Lateral parietal cortex; lateral prefrontal cortex; lateral temporal cortex; medial parietal cortex
Reasoning paradigm	Anatomical granularity / region-specific attribution
Core evidence constraint	The regions exhibit partially overlapping network memberships but differ in evidence-supported disease associations and lesion-derived episodic memory network involvement.
Representative region-level constraints	Lateral parietal cortex: Alzheimer's disease, schizophrenia, cingulo-opercular network, lesion-derived episodic memory network; Lateral prefrontal cortex: Alzheimer's disease, Parkinson's disease dementia, default mode network, fronto-parietal network; Lateral temporal cortex: Alzheimer's disease, lesion-derived episodic memory network, fronto-parietal network; Medial parietal cortex: schizophrenia, default mode network.
Key reasoning risks	Invalid network-to-region inheritance, sibling-region transfer, hub stitching, and association-to-causation extrapolation.
Core challenge	Shared network membership may encourage models to incorrectly transfer disease associations or mechanistic roles across anatomically distinct cortical regions.
Gold options	A and E

Table 20: Evidence-card summary for a representative granularity item. The item evaluates whether models preserve evidence-supported region-level constraints without overgeneralizing across shared networks or neighboring cortical regions.

E. Both the lateral parietal cortex and lateral temporal cortex are associated with Alzheimer's disease and share membership in the fronto-parietal network and lesion-derived episodic memory network, but only the lateral parietal cortex is additionally associated with schizophrenia.

Gold labels. The evidence-supported options are **A** and **E**.

F.5 Differential Example: Sensorimotor-Network Involvement Across Neurodegenerative Conditions

This section presents a representative example from the cross-disease differential reason-

Option	Label	Gap type	Audit rationale
A	Supported	–	The lateral temporal cortex is associated with Alzheimer's disease and participates in the lesion-derived episodic memory network. The option preserves the correct region-level attribution.
B	Not supported	`granularity_inheritance_error`	The medial parietal cortex belongs to the default mode network, but this does not license an Alzheimer's disease association. The stem explicitly indicates relative preservation of midline parietal structures in the Alzheimer's disease cohort.
C	Not supported	`hub_stitching`	The option falsely transfers lesion-derived episodic memory network membership to the lateral prefrontal cortex and then stitches this unsupported network role to a shared Alzheimer's/Parkinson's disease dementia mechanism.
D	Not supported	`association_to_causation`	The option turns a network association involving the lateral parietal cortex and cingulo-opercular network into a direct causal explanation for Alzheimer's-related episodic memory decline, which is not supported by the evidence.
E	Supported	–	The option correctly preserves all region-specific constraints: both lateral parietal and lateral temporal cortex are associated with Alzheimer's disease and share fronto-parietal and lesion-derived episodic memory network membership, while schizophrenia is associated only with the lateral parietal cortex among the two.

Table 21: Option-level audit for the representative granularity item. The item tests whether models avoid over-inheriting disease and network associations across anatomically distinct cortical regions.

ing paradigm. The example illustrates how BrainBench-Calibrated evaluates whether models can distinguish disorders that share partially overlapping functional-network involvement while preserving disease-specific clinical, genetic, regional, and mechanistic constraints. Table 22 summarizes the evidence-card constraints and reasoning risks used to construct the item, while Table 23 provides the option-level audit rationale.

Question stem. A 58-year-old patient presents with progressive behavioral disinhibition, apathy, and loss of empathy over 18 months. Neurological examination reveals subtle upper motor neuron signs, including brisk reflexes and mild spasticity in the upper extremities. Structural MRI shows frontal and temporal atrophy with additional involvement of the precentral gyrus bilaterally. Voxel-based morphometry reveals grey matter loss in the supplementary motor area and cerebellum. Genetic testing reveals a hexanucleotide repeat expansion in `C9orf72`. Functional connectivity analysis shows disruption within the sensorimotor network.

Given the clinical, imaging, and genetic findings, and the disease-level constraints summarized in Table 22, which of the following statements about differential diagnosis within the sensorimotor-network context are supported? Select all that apply.

A. The combination of behavioral-variant FTD symptoms, `C9orf72` expansion, upper motor neuron signs, and cerebellar plus motor-cortical involvement supports `C9orf72`-bvFTD within the sensorimotor-network differential, while disorders of consciousness are less supported by this feature pattern.

B. Cerebellar grey matter loss by itself determines a diagnosis of `C9orf72`-associated frontotemporal dementia and excludes all other sensorimotor-network disorders from the differential.

C. In motor neuron disease, supplementary motor area pathology is mediated through selective retrograde axonal transport deficits, which constitute the primary mechanistic link shared with behavioral variant frontotemporal dementia.

D. In behavioral variant frontotemporal dementia, salience-network disruption is mechanistically driven by aberrant astrocytic pruning of dopaminergic synapses within the anterior cingulate cortex, leading to socioemotional deficits.

Field	Value
Shared network context	Sensorimotor network
Disease set	Behavioral variant frontotemporal dementia (bvFTD); C9orf72-bvFTD; disorders of consciousness; motor neuron disease
Reasoning paradigm	Cross-disease differential reasoning
Core evidence constraint	The disorders share partially overlapping sensorimotor-network involvement but differ in clinical presentation, genetic evidence, regional involvement, and mechanism-level support.
Representative disease-level constraints	bvFTD: cerebellum, primary motor cortex, supplementary motor area, network degeneration; C9orf72-bvFTD: cerebellum, primary motor cortex, sensorimotor cortex, supplementary motor area, C9orf72 repeat expansion; Disorders of consciousness: sensorimotor cortex and supplementary motor area involvement without frontotemporal behavioral syndrome; Motor neuron disease: primary motor cortex, sensorimotor cortex, supplementary motor area, upper motor neuron involvement.
Clinical and imaging cues used in the item	Behavioral disinhibition, apathy, loss of empathy, upper motor neuron signs, frontal-temporal atrophy, precentral gyrus involvement, supplementary motor area degeneration, cerebellar atrophy, and C9orf72 repeat expansion.
Key reasoning risks	Overgeneralizing shared sensorimotor-network involvement into a single disease mechanism, subtype inheritance without evidence, unsupported mechanistic insertion, and disease-mechanism stitching across disorders.

Table 22: Evidence-card summary for a representative differential reasoning item. The item evaluates whether models preserve disease-specific constraints despite overlapping sensorimotor-network involvement across disorders.

E. Because all four conditions involve the sensorimotor network, network degeneration should be treated as a shared mechanism across behavioral variant frontotemporal dementia, C9orf72-bvFTD, disorders of consciousness, and motor neuron disease.

Gold label. The evidence-supported option is **A**.

G Prompt Templates

The prompts below are construction-time prompts. Some prompts request multiple candidate questions or a larger option pool to increase diversity. Only items that pass evidence calibration, revision, and pruning are retained in the final five-option benchmark.

G.1 Knowledge Extraction Prompt

```
You are a neuroscience knowledge extraction expert building a structured evidence graph from scientific literature on brain networks and brain disorders.
Extract structured scientific knowledge from the provided text. Focus on: brain regions, brain networks, biological mechanisms, pathological processes, neuroimaging findings, and neurological/psychiatric disorders. Only extract relations clearly supported by the text.

Entity types
BrainRegion          Anatomical structures (e.g. hippocampus, thalamus)
BrainNetwork         Large-scale networks (e.g. DMN, salience network)
Disease              Specific clinical disorders only do not label generic phrases ("neurodegeneration") as Disease
Mechanism            Biological/physiological mechanisms
PathologicalProcess  Propagation/accumulation /progression of pathology
ImagingBiomarker     Neuroimaging findings or measurements

Allowed relation types
part_of_network: BrainRegion → BrainNetwork
network_altered_in_disease: BrainNetwork → Disease
region_implicated_in_disease: BrainRegion → Disease
mechanism_affects_network: Mechanism → BrainNetwork
mechanism_contributes_to_pathology: Mechanism → PathologicalProcess
pathology_implicated_in_disease: PathologicalProcess → Disease
biomarker_detects_mechanism: ImagingBiomarker → Mechanism
biomarker_detects_pathology: ImagingBiomarker → PathologicalProcess

Extraction rules (1) Extract only explicitly supported relations. (2) Use normalized entity names; expand abbreviations (e.g. PCC → posterior cingulate cortex). (3) No duplicate triples per chunk. (4) Each triple requires a 10-30 word evidence span, a 1-2 sentence relation_explanation, and a confidence score
```

Option	Label	Gap type	Audit rationale
A	Supported	–	The option integrates behavioral symptoms, genetic evidence, upper motor neuron signs, and cerebellar plus motor-cortical involvement. The reasoning preserves disease-specific constraints rather than inferring diagnosis solely from shared sensorimotor-network disruption.
B	Not supported	absolute_overclaim	Cerebellar involvement contributes to the differential pattern but does not uniquely determine C9orf72-bvFTD or exclude all other sensorimotor-network disorders.
C	Not supported	unsupported_mechanism	The option introduces a specific retrograde axonal transport mechanism that is not supported by the evidence card.
D	Not supported	unsupported_mechanism	The option introduces an unsupported molecular mechanism involving astrocytic pruning and shifts beyond the evidence-supported sensorimotor-network differential context.
E	Not supported	hub_stitching	The option incorrectly treats shared sensorimotor-network involvement as sufficient evidence for a common disease mechanism across all four disorders.

Table 23: Option-level audit for the representative differential item. The item tests whether models can preserve disease-specific evidence constraints despite overlapping functional-network involvement.

(0-1): 0.9-1.0 explicitly stated; 0.7-0.89 strongly implied; 0.5-0.69 moderately supported; <0.5 do not extract.

Triple ID: {PMID}_{chunk_id}_{index} (index starts at 01 per chunk). Fixed metadata per triple: pmid, paper_title, chunk_id.

Output: JSON with key "triples" containing an array of triple objects (schema below), or {"triples": []} if no valid relations found.

```
{ "triple_id": "...", "head": "...",
  "head_type": "...",
  "relation": "...", "tail": "...",
  "tail_type": "...",
  "evidence": "...",
  "relation_explanation": "...",
  "confidence": 0.0, "pmid": "...",
  "paper_title": "...", "chunk_id": "..." }
```

G.2 Single-Path Semantic Review Prompt

You are a senior neuroscience researcher and knowledge-graph quality auditor. Evaluate whether a reasoning path extracted from a brain-network knowledge graph constitutes a scientifically sound causal chain.

Input (filled at runtime)
Node sequence with types; relationship sequence; step-by-step edge details (evidence span + explanation for each hop).

Scoring dimensions (1-5 each; provide score + rationale)

1. **Causal chain completeness** Is every step directional and causal (not merely correlational)? Can the path be narrated as a coherent mechanistic story?
2. **Hub-stitching detection** Does a generic intermediate node (e.g. DMN, hippocampus) artificially connect two otherwise unrelated topics? Test: would the endpoints still be related if the hub were removed?
3. **Disease-context consistency** Are all steps in the same disease context, or is any cross-disease link directly supported by literature?
4. **Scientific accuracy** Are all causal claims consistent with mainstream neuroscience? Any reversed causality or oversimplification?
5. **Information value** Do intermediate nodes reveal non-obvious mechanisms, or are they near-synonyms of their neighbors?
6. **Question-design potential** What level of expertise is required to reconstruct the intermediate chain from endpoints alone?

Issue tags (select all that apply, or none):
hub_stitching | cross_disease_context | tautological | too_generic | direction_ambiguous | scientifically_questionable | low_information

Output: strict JSON:

```
{ "scores": {
"causal_completeness":{"score":4,"rationale":""},
"hub_stitching":{"score":5,"rationale":""},
"context_consistency":{"score":3,"rationale":""},
"scientific_accuracy":{"score":4,"rationale":""},
"information_value":{"score":3,"rationale":""},
"question_potential":{"score":4,"rationale":""}},
"overall_score": 23,
"issues": ["none"],
"issue_details": "...",
"improvement_suggestions": "...",
"recommended_question_types":[""],
"brief_narrative": "..." }
```

G.3 Cross-Path Question Generation Prompt

You are an expert neuroscience examination designer. Generate **exactly one** extremely challenging multiple-response question from a **cross-path** reasoning pair. The question must require deep integration of two partially conflicting mechanistic chains, not rote recall or heuristic judgment.

Input fields (filled at runtime)
pair_type, conflict_level, shared_network; Path A and Path B (IDs, grades, node sequences); combined hop count; conflict-authenticity,

question-value, and reconciliation-complexity scores with rationales; recommended moderating variables; modification suggestion; overall verdict rationale.

Stem design
Describe a realistic experimental/clinical scenario (e.g. "A longitudinal PET-fMRI study . . . reveals. . . "). Embed enough context for self-containment. Do not name the conflict explicitly the test-taker must infer it.

Distractor requirements
- ≥ 2 near-correct options with subtle errors (wrong stage, anatomy, or directionality).
- ≥ 1 option inverting the moderating-variable logic (e.g. preclinical finding attributed to dementia stage).
- ≥ 1 option with a real literature finding but misattributed mechanistic explanation.
- ≥ 2 "minimal pairs": options identical except for one moderating variable; only one is correct.

Cross-path integration requirement
Familiarity with Path A or Path B alone must be insufficient for a correct answer; correct answering requires understanding the interaction, tension, or stage-dependent relationship between both paths.

Universal question-design rules (all MCQ prompts).
- Format: Multiple-response (select-all-that-apply); correct-answer count not disclosed; ≥ 7 options preferred 8-10; any number of correct answers.
- Forbidden: "All/None of the above"; options eliminable by pure logic without domain knowledge; hedging language as a reliable correctness signal; systematic length differences between correct and distractor options.
- Distractors: Every distractor must be grounded in authentic neuroscience; ≥ 2 "near-correct" options with subtle errors (wrong directionality, stage, anatomy, or unsupported inference); ≥ 1 factually-true-but-irrelevant option.
- Correct options: May use paraphrased or alternative terminology; confidence level must match evidence strength (strong evidence $\Rightarrow$ decisive wording; genuine ambiguity $\Rightarrow$ hedged wording).

Output format. Return valid JSON only with fields: question_type (multiple_response), question_stem, options (keys A-H or more), correct_answers, explanation (with overall_logic and per-option option_analysis), and difficulty_rationale. Metadata keys specific to each prompt type are noted inline. Additional JSON keys: pair_id.

G.4 Granularity Discrimination Question Prompt

You are an expert neuroscience examination designer. Generate **one** benchmark-grade multiple-response question assessing **neuroanatomical granularity discrimination**: the ability to attribute diseases, networks, and mechanisms to the correct anatomical level (parent region vs. subregion).

Input fields (filled at runtime)
group_id; parent region; child regions (list); per-region connection-difference features (diseases, networks, mechanisms, biomarkers); source path IDs.

Targeted LLM failure modes
(1) Conflation of parent with subregion; (2) over-generalising a disease/network association across sibling subregions; (3) upward attribution of a subregion property to the parent; (4) failure to recognise that the same disease may involve multiple subregions via different mechanisms.

Distractor requirements
- ≥ 2 granularity-level errors: correct statement attributed to wrong anatomical level.
- ≥ 1 sibling-swap: disease of one child incorrectly attributed to another child.
- ≥ 1 upward attribution: parent-level claim that is true only for a specific child.
- ≥ 1 network-disease chain error: individual regionnetwork and regiondisease facts both true, but the mechanistic chain is unsupported.
- ≥ 2 granularity minimal pairs: same claim at two different anatomical levels; only one correct.

Stem design
Present a realistic neuroimaging/clinical scenario. Reference the subregion implicitly (via connectivity, cytoarchitectonic, or functional clues) do not name it. Cover ≥ 3 distinct regions from the group across all options.

Universal question-design rules (all MCQ prompts).
- Format: Multiple-response (select-all-that-apply); correct-answer count not disclosed; ≥ 7 options preferred 8-10; any number of correct answers.
- Forbidden: "All/None of the above"; options eliminable by pure logic without domain knowledge; hedging language as a reliable correctness signal; systematic length differences between correct and distractor options.
- Distractors: Every distractor must be grounded in authentic neuroscience; ≥ 2 "near-correct" options with subtle errors (wrong directionality, stage, anatomy, or unsupported inference); ≥ 1 factually-true-but-irrelevant option.
- Correct options: May use paraphrased or alternative terminology; confidence level must match evidence strength (strong evidence $\Rightarrow$ decisive wording; genuine ambiguity $\Rightarrow$ hedged wording).

Output format. Return valid JSON only with fields: question_type (multiple_response), question_stem, options (keys A-H or more), correct_answers, explanation (with overall_logic and per-option option_analysis), and difficulty_rationale. Metadata keys specific to each prompt type are noted inline. Additional JSON key: group_id.

G.5 Differential Diagnostic Question Prompt

You are an expert neuroscience exam designer. Generate **three** benchmark-level multiple-response questions assessing **differential diagnostic reasoning** across diseases sharing the same network substrate. Each question should focus on a different disease subset to ensure full coverage of the disease set.

Input fields (filled at runtime)
overlap_id; shared network/substrate; disease set; question-value score; per-disease differentiating features (brain regions, mechanisms, biomarkers).

Targeted LLM failure modes
(1) Cross-disease feature conflation via shared network; (2) region-mechanism dissociation (knowing both facts but failing to link them correctly); (3) biomarker-mechanism conflation due to semantic similarity; (4) subset-superset disease conflation; (5) assuming identical regional roles across diseases.

Distractor requirements

- ≥ 2 feature-swap options: a feature unique to Disease X wrongly attributed to Disease Y (high semantic similarity required).
- ≥ 2 conjunction-fallacy options: each constituent fact true, but the conjunction is false (e.g. Disease A involves Region R and Mechanism M, but M does not operate via R in Disease A).
- Subset/superset pairs where applicable.
- ≥ 1 mechanism-swap minimal pair: same disease and region, different mechanism.
- ≥ 1 disease-swap minimal pair: identical except disease name; only one correct.
- ≥ 1 option with correct disease, region, and biomarker but wrong directionality (increased vs. decreased).
- ≥ 1 factually true option from a different network than the shared substrate (out-of-scope trap).

Stem design
Brief clinical vignette with imaging findings pointing to the shared network but not naming the diagnosis. At least 2-3 diseases must remain plausible after reading the stem. Embed ≥ 1 subtle differentiating detail between closely related diseases without foregrounding it.

Universal question-design rules (all MCQ prompts).

- Format: Multiple-response (select-all-that-apply); correct-answer count not disclosed; ≥ 7 options preferred 8-10; any number of correct answers.
- Forbidden: "All/None of the above"; options eliminable by pure logic without domain knowledge; hedging language as a reliable correctness signal; systematic length differences between correct and distractor options.
- Distractors: Every distractor must be grounded in authentic neuroscience; ≥ 2 "near-correct" options with subtle errors (wrong directionality, stage, anatomy, or unsupported inference); ≥ 1 factually-true-but-irrelevant option.
- Correct options: May use paraphrased or alternative terminology; confidence level must match evidence strength (strong evidence $\Rightarrow$ decisive wording; genuine ambiguity $\Rightarrow$ hedged wording).

Output format. Return valid JSON only with fields: question_type (multiple_response), question_stem, options (keys A-H or more), correct_answers, explanation (with overall_logic and per-option option_analysis), and difficulty_rationale. Metadata keys specific to each prompt type are noted inline. Additional JSON key: overlap_id.

G.6 Causal Direction Question Prompt

You are an expert neuroscience examination designer. Generate **exactly one** multiple-response question assessing **causal directionality reasoning** between two nodes in a pathological network.

Input fields (filled at runtime)
causal_id; source path ID; Node A and Node B with types; focal relation (original wording); evidence text; explanation; causal-language strength; reverse-edge flag; potential confounders; dispute types; chain context; review scores (ambiguity genuineness, methodological richness, confounder relevance, question value) each /5 with rationale; true direction if known; best study design to resolve; modification suggestion.

Six distractor categories
(a) Direction-evidence mismatch (≥ 2, each different direction/study-type); (b) Correct direction, wrong evidence base; (c) Confounder misdirection (real confounder, wrong directional conclusion); (d) Bidirectionality oversimplification or overcomplication; (e) Study-design precision error (sounds sophisticated, cannot actually resolve this ambiguity); (f) Chain-context exploitation (treats chain-level order as pairwise causal evidence).

Correct option requirements
≥ 1 describing specific evidence supporting a specific direction with appropriate confidence; ≥ 1 identifying a genuine (non-generic) limitation; ≥ 1 making a precise study-design suitability claim grounded in the specific neurobiology; ≥ 1 addressing what the provided evidence text actually warrants directionally.

Minimal pairs required
Evidence-conclusion pair; confounder-impact pair; design-disambiguation pair.

Stem design
Present ≥ 2 pieces of evidence pointing in different directions or carrying different evidential weight. Include ≥ 1 easily-overlooked methodological detail that is critical for evaluating at least one option. Frame as "Which statements about this mechanistic relationship and its supporting evidence are well-founded?" not as a binary A$\rightarrow$B vs. B$\rightarrow$A choice.

Universal question-design rules (all MCQ prompts).

- Format: Multiple-response (select-all-that-apply); correct-answer count not disclosed; ≥ 7 options preferred 8-10; any number of correct answers.
- Forbidden: "All/None of the above"; options eliminable by pure logic without domain knowledge; hedging language as a reliable correctness signal; systematic length differences between correct and distractor options.
- Distractors: Every distractor must be grounded in authentic neuroscience; ≥ 2 "near-correct" options with subtle errors (wrong directionality, stage, anatomy, or unsupported inference); ≥ 1 factually-true-but-irrelevant option.
- Correct options: May use paraphrased or alternative terminology; confidence level must match evidence strength (strong evidence $\Rightarrow$ decisive wording; genuine ambiguity $\Rightarrow$ hedged wording).

Output format. Return valid JSON only with fields: question_type (multiple_response), question_stem, options (keys A-H or more), correct_answers, explanation (with overall_logic and per-option option_analysis), and difficulty_rationale. Metadata keys specific to each prompt type are noted inline. Additional JSON key: causal_id.

G.7 Counterfactual Question Prompt

You are an expert neuroscience exam designer. Generate **one** benchmark-level multiple-response question assessing **counterfactual causal reasoning**: predicting system-level consequences of removing a specific mechanistic node, with rigorous attention to necessity, alternative-pathway sufficiency, and graded dependency.

Input fields (filled at runtime)
candidate_id; removed node and its type; source path IDs; downstream effects with dependency ratings; alternative pathways per effect; review scores (causal validity, counterfactual tractability, alternative-pathway accuracy, question value) each /5 with rationale; known issues; modification suggestion.

Targeted LLM failure modes
(1) Binary necessity thinking (ignoring graded dependency); (2) path-counting heuristic (more paths $\neq$ lower dependency); (3) conflating existence with sufficiency of alternative paths; (4) ignoring emergent/phenotypic consequences; (5) symmetry bias ("removing A must eliminate B").

Anti-heuristic principles: correct options should not all be hedged; use strong wording when dependency is genuinely very strong or negligible; hedge only when the evidence genuinely warrants it. Distractors must match correct options in length and specificity.

Six distractor categories
(a) Valid reasoning structure, erroneous pathway details (wrong intermediate node, inaccessible precondition); (b) Approximately correct conclusion, wrong causal justification; (c) Precise-condition fallacy (conditional framework correct, one element subtly wrong wrong disease subtype, reversed temporal direction); (d) Cross-effect contamination (conflates dependency or pathway analysis of two different downstream effects); (e) Temporal/contextual frame shift (correct for a different removal paradigm than specified in stem); (f) Pathway topology correct, biological capacity wrong (over- or under-estimates compensatory capacity).

Correct option requirements
≥ 1 definitive (non-hedged) prediction where evidence warrants it; ≥ 1 identifying which alternative pathway dominates and its second-order consequence; ≥ 1 precise negative claim about what cannot be inferred; ≥ 1 comparative claim across ≥ 2 downstream effects (e.g. "Effect A more severely disrupted than Effect B").
Alternative-pathway options
≥ 2 options specifying concrete intermediate-node pathways: one with correct pathway but wrong compensatory capacity, one with correct capacity but wrong pathway composition.
Minimal pairs
≥ 2 pairs differing in: a specific intermediate node, a contextual condition, a qualitative downstream consequence, or the direction of a compensatory change (not mere degree modifiers like "partially" vs. "completely").

Stem design
Specify removal method, timing, and observation window. Every contextual detail must determine the correctness of ≥ 1 option (e.g. "adult-onset" rules out developmental compensation claims; "6-month window" rules out multi-year consequences). Present partial preliminary data and ask which additional downstream consequences are warranted.
Universal question-design rules (all MCQ prompts).
- Format: Multiple-response (select-all-that-apply); correct-answer count not disclosed; ≥ 7 options preferred 8-10; any number of correct answers.
- Forbidden: "All/None of the above"; options eliminable by pure logic without domain knowledge; hedging language as a reliable correctness signal; systematic length differences between correct and distractor options.
- Distractors: Every distractor must be grounded in authentic neuroscience; ≥ 2 "near-correct" options with subtle errors (wrong directionality, stage, anatomy, or unsupported inference); ≥ 1 factually-true-but-irrelevant option.
- Correct options: May use paraphrased or alternative terminology; confidence level must match evidence strength (strong evidence $\Rightarrow$ decisive wording; genuine ambiguity $\Rightarrow$ hedged wording).

Output format. Return valid JSON only with fields: question_type (multiple_response), question_stem, options (keys A-H or more), correct_answers, explanation (with

overall_logic and per-option option_analysis), and difficulty_rationale. Metadata keys specific to each prompt type are noted inline. Additional JSON key: candidate_id.

G.8 Option Auditing Prompt

You are a neuroscience evidence auditor. Your task is **not** to judge plausibility but to decide whether a claim is **well-founded by the provided evidence card**. Distinguish strictly between textual support (what the text says) and evidential warrant (what the evidence logically supports).

Input fields (filled at runtime)
question_id, stage, question_stem; option_id, claim_text, original_gold; evidence card: source_id, source_path_ids, chain_context, evidence_text, dispute_types, hierarchy_or_diff_features, modality_or_design_notes, input_quality_issues.
Evidence priority: (1) evidence_text; (2) dispute_types; (3) hierarchy_or_diff_features; (4) modality_or_design_notes. If input_quality_issues $\neq$ none, be conservative: set risk_signal $\geq$ medium.

Action policy
Decide revised_gold first, then apply:

original_gold	revised_gold	Default action
correct	well_found	keep
correct	not_well_found	rewrite
incorrect	well_found	keep
incorrect	not_well_found	rewrite

Hard constraint: never drop a well_found option unless it is a duplicate.

Output: JSON only:

```
{ "question_id": "...", "option_id": "...",
"textual_support":"entailed|partially_supported
    |not_stated|contradicted",
"evidential_warrant": "adequate|limited|invalid",
"revised_gold": "well_found|not_well_found",
"gap_type":"causal_direction_error|
association_to_causation|hub_stitching|
granularity_inheritance_error|
unsupported_mechanism|
absolute_overclaim|none",
  "risk_signal": "low|medium|high",
  "action": "keep|rewrite|drop",
  "evidence_ids": ["..."],
  "rationale": "2-4 sentences",
  "confidence": 0.0 }
```

G.9 Option Rewriting Prompt

You are an expert neuroscience exam designer. Rewrite **one option** to meet a specified target criterion.

Input fields (filled at runtime)
Target definition; question stem; original option text; chain context; evidence text; dispute types; modality/design notes; hierarchy/diff features.

Rewriting constraints

1. Return exactly one sentence in rigorous scientific style; no bullet points.
2. Length: 180-250 characters (prevents length-based guessing).
3. Banned words: completely, entirely, always, purely, solely, perfectly, abolished.
4. Use nuanced academic phrasing (e.g. significantly reduced, partially attenuated, modulated).
5. No giveaway adverbs (obviously, definitely, clearly).
6. Preserve the original option's thematic meaning; ensure grammatical coherence after the question stem.

Output: JSON only:

```
{ "rewritten_claim": "...",
  "rewrite_rationale": "1-2 sentences." }
```