

Agent Collaboration Safety Issues Caused by LLMs' Nature

Taking **reliable** and **verifiable** hypothesis generation as a case study

Mingxue Xu

mingxue-xu.github.io

7 October 2026, Oxford Collaboration Week

- **0. Creativity in Science, and Hypothesis Generation**
- **1. Limitations of individual model reliability**
 - Resource and capacity limits
 - Context window degradation
 - Nondeterminism in deployed LLM systems
- **2. Verifiable hypothesis from agent collaboration**
 - Model capacity by agent role
 - GraphHypoth architecture and literature grounding
 - Deterministic measurements and model judgments
 - Claim graphs and honestly unsupported claims

Creativity in Science, and Hypothesis Generation

Mental Operations Relevant to Scientific Creativity

Hypothesis generation (problem finding), hypothesis testing (problem solving), and logic inference generation lie at the heart of scientific reasoning.

- **Creativity:** *new* and *useful*.
- **Common origins of scientific creativity:**
 - **Single Domain:** deep knowledge in a specific field can lead to novel insights and connections.
 - **Cross Domain:** connecting knowledge from different fields can spark innovative ideas and approaches.

Abraham (2018)

1. Limitations of individual model reliability

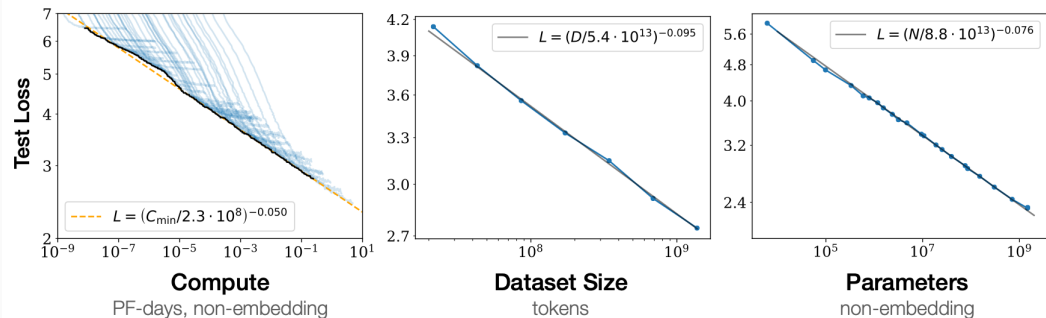
1.1 Resource and capacity limits of transformer-based language models

Three lines of work describe the limits of a transformer-based language model:

1. **Scaling laws (empirical).** Kaplan et al. (2020): loss vs. parameters, data, compute. Hoffmann et al. (2022), Chinchilla: allocation of training compute between model size and data, $\hat{L}(N, D) = E + A/N^\alpha + B/D^\beta$ under $C \approx 6ND$.
2. **Expressivity (theoretical, architecture-dependent).** The functions a network can represent (Xu et al., 2024). Hahn (2020): fixed-depth hard-attention transformers cannot recognise parity over arbitrary lengths.
3. **Capacity (bound + measurements).** The information a network can memorise (Xu et al., 2024). Allen-Zhu and Li (2024): knowledge recoverable vs. bits stored in weights; ≈ 2 bits/parameter is empirical.

Allen-Zhu and Li (2024); Hahn (2020); Hoffmann et al. (2022); Kaplan et al. (2020); Xu et al. (2024)

Figure A. Loss and resource scaling



Test loss follows approximate linear power laws
as compute, data, or parameters increase in the studied regimes.

Kaplan et al. (2020), Figure 1

1.1 Implication for the workflow

- Specify the task, architecture, precision, training resources, and inference procedure *before* discussing a limit.
- Counting argument (storage only): P parameters at b bits each give at most 2^{Pb} configurations, so exact recovery of B independent binary facts needs $B \leq Pb$. Not a bound on compressible facts or reasoning accuracy.

Design recommendation

Evaluate each agent on a *bounded task* and check its outputs. Security controls must remain enforceable when that agent makes a mistake.

1.2 Context window degradation

Lessons from engineering practice

Long histories accumulate irrelevant messages and tool outputs. Agents need to select the context useful for the next step.

The right reset strategy depends on the model. Anthropic removed manual context resets when moving from Sonnet 4.5 to Opus 4.5.

Re-evaluate context-management rules after model upgrades.

Proposed application to GraphHypoth

Each handoff carries the claim, evidence IDs, verification status, and next action. Permission checks remain in code outside conversational memory.

Practices for long tasks

- **Retrieve on demand.** Keep source pointers and load relevant passages when needed.
- **Persist task state.** Save decisions and unfinished work outside the conversation.
- **Compact carefully.** Remove redundant history while retaining critical details.
- **Limit each agent's context.** Give specialists a bounded task and collect concise findings.

Rajasekaran et al. (2025); Martin et al. (2026)

1.3 Nondeterminism in deployed LLM systems

Variation has several sources

Inference numerics. Temperature 0 can still yield different outputs when batching changes floating-point results.

Thinking Machines demonstrated repeatable inference with kernels that preserve results across batch sizes.

Execution environment. CPU/RAM limits and timeouts can change agent outcomes and benchmark scores.

Repeated success is a stronger requirement

If independent trials each succeed 75% of the time, all three succeed with probability $0.75^3 \approx 42\%$.

Practices for evaluation

- **Record the setup.** Model, prompts, tools, data, and resource limits belong in the run record.
- **Repeat trials.** Reset environments and separate infrastructure failures from task failures.
- **Check outcomes and traces.** Review failed runs and validate the graders.
- **Measure consistency.** Report per-trial success and the chance that all k trials succeed (pass^k).

He and Thinking Machines Lab (2025); Grace et al. (2026); Segato (2026)

2. Agent Collaboration for Verifiable Outcomes

2.1 Outcomes Verified by Agents

GraphHypoth introduces seven workflow roles, covering generation, review, experiment planning, and presentation of the results.

Extraction Agent: Converts the seed claim or research goal into concept nodes and directed edges. Each edge begins as unverified.

Evidence Reviewer: Grades the short-listed passages for each edge. Deterministic code combines those grades into a status and confidence and records them on the edge.

Research Synthesist: Uses one persistent thread to mine concepts from retrieved passages, propose candidate hypotheses over the committed graph, and revise them after review. Deterministic gates filter and rank the candidates.

Critic Panel: Reviews, grades, and ranks proposed hypotheses, then feeds critique back to the Synthesist. Two judge seats always run; a third joins when they disagree.

Experiment Designer: Drafts a plan for each confirmed hypothesis whose graph commit succeeds. The plan draws on targeted methods retrieval and the graph's confounder and mediator nodes.

Experiment Validator: Grades the plan on five criteria, with one revision round by default. A content-complete plan is then committed against the hypothesis edge.

Narrative Elaborator: Produces plain-language hypothesis prose for trace reports. It runs by default unless prose generation is disabled with `-no-elaborate`.

Source: [Mingxue-Xu \(n.d.-a\)](#), [GraphHypoth: README](#), Agent Roles and What A Complete Run Does. Retrieved October 13, 2026.



2.2 GraphHypoth: architecture overview

Shared claim graph

- Nodes: concepts
- Directed edges: relationships asserted by sources

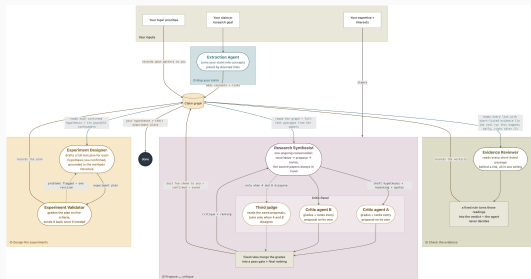
Stages

1. Extraction
2. Evidence retrieval and review
3. Hypothesis development
4. Critique
5. Experiment planning

Factthink (n.d.-b); Mingxue-Xu (n.d.-a)

Separation of concerns

LLM-generated judgments are separated from fixed **scoring** and state-update **rules**.



2.2.1 Grounding in the literature

- **RAG** (Lewis et al., 2020) incorporates retrieved external sources into generation.
- **FActScore** (Min et al., 2023) splits an output into atomic claims and checks support in a knowledge source; its automated estimator still uses an LLM.

Applied to a research workflow

- Associate each claim with the relevant source passage.
- Review whether the passage supports the claim and whether qualifications are retained.
- Citation presence alone is insufficient: a source may be irrelevant or misinterpreted.
- Retrieved content is *potentially untrusted input* and needs separate handling.

Lewis et al. (2020); Min et al. (2023)

2.2.2 How GraphHypoth calculates scores

Evidence association

$$A = 0.55(0.8E + 0.2L) + 0.35C + 0.10P$$

E : embedding similarity, L : lexical match,

C : citation overlap, P : source prior.

Edge confidence for verdict v

$$C_v = \text{clip}_{[0,1]}(0.55T_v + 0.25B_v + 0.20D_v \\ - 0.25R - 0.15Q)$$

T_v : top-three evidence mean, B_v : support balance,

D_v : source breadth, R : risks, Q : qualifiers (counts capped after division by 3).

LLMs supply judgments that fixed formulas aggregate. These are declared defaults, with no calibration established in the specs. Full definitions and gates appear in the scoring appendix.

Mingxue-Xu (2026), GraphHypoth legacy, specs 04–07 (commit 5064d3d)

Hypothesis score and final rank

$$H = \text{clip}_{[0,1]} \left(\frac{N + P_h + T_h + Y + G + M}{6} \right)$$

$$\text{RankScore} = \text{clip}_{[0,1]}(H(1 - S))$$

N : novelty, P_h : plausibility, T_h : testability, Y : expected yield, G : centrality, M : mechanism specificity, S : saturation.

Critic aggregation

$$N = \text{median}_j N_j, \quad S = \text{median}_j S_j$$

Judged candidates use panel medians. Ungraded candidates retain gate novelty and their existing saturation.

2.2.3 Claim graphs and unsupported claims

- A claim graph records relationships, associated evidence, and verification status, so a reviewer can inspect individual dependencies and find missing support.
- In GraphHypoth, newly committed hypotheses remain *unverified*; confirmation records selection for further investigation.
- A directed edge is an asserted relationship, not necessarily a causal one.

Downstream use

Later stages should consume the recorded status and evidence, with separate checks on which agents may change them. The graph gives traceability; the accuracy of its contents remains an evaluation question.

Min et al. (2023); Mingxue-Xu (n.d.-a)

Millions of thanks!!

Any questions/thoughts/opinions? :D

References i

Abraham, A. (2018). *The neuroscience of creativity*. Cambridge University Press.

<https://doi.org/10.1017/9781316816981>

Allen-Zhu, Z., & Li, Y. (2024). *Physics of language models: Part 3.3, knowledge capacity scaling laws* [Preprint]. arXiv.

<https://doi.org/10.48550/arXiv.2404.05405>

Bengio, Y. (2026, September 11). *Why are AI agents lying, cheating and coordinating?*

<https://yoshuabengio.org/en/blog/why-are-ai-agents-lying-cheating-and-coordinating>

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). *AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents* [Preprint]. arXiv.

<https://doi.org/10.48550/arXiv.2406.13352>

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 11733–11763). PMLR. <https://proceedings.mlr.press/v235/du24a.html>

Facthink. (n.d.-a). *Brand artwork* [Banner and logo]. GitHub.

<https://github.com/Mingxue-Xu/facthink-webpage/tree/main/assets>

- Factthink. (n.d.-b). *How it works*. Retrieved October 6, 2026, from <https://factthink.com/how-it-works.html>
- Grace, M., Hadfield, J., Olivares, R., & De Jonghe, J. (2026, January 9). *Demystifying evals for AI agents*. Anthropic. <https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents>
- Greenblatt, R., Cotra, A., & Wijk, H. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8, 156–171. https://doi.org/10.1162/tac1_a_00306
- He, H., & Thinking Machines Lab. (2025, September 10). *Defeating nondeterminism in LLM inference*. Thinking Machines Lab: Connectionism. <https://doi.org/10.64434/tml.20250910>
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). *Training compute-optimal large language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2203.15556>

- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2001.08361>
- Lee, D., & Tiwari, M. (2024). *Prompt infection: LLM-to-LLM prompt injection within multi-agent systems* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2410.07283>
- Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In S. Das, E. Durfee, K. Larson, & M. Winikoff (Eds.), *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems* (pp. 464–473). International Foundation for Autonomous Agents and Multiagent Systems. <https://www.ifaamas.org/Proceedings/aamas2017/pdfs/p464.pdf>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
- Martin, L., Cemaj, G., & Cohen, M. (2026, April 8). *Scaling Managed Agents: Decoupling the brain from the hands*. Anthropic. <https://www.anthropic.com/engineering/managed-agents>

- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- Mingxue-Xu. (n.d.-a). *GraphHypoth: README*. GitHub. Retrieved October 13, 2026, from <https://github.com/Mingxue-Xu/graph-hypoth>
- Mingxue-Xu. (n.d.-b). *Hypo, the GraphHypoth hippo* [Logo]. GitHub. <https://raw.githubusercontent.com/Mingxue-Xu/graph-hypoth/main/docs/assets/logo/hypo-light.svg>
- Mingxue-Xu. (n.d.-c). *Which agents go local? Small local models and retrieval for the claim graph pipeline*. GitHub. Retrieved October 13, 2026, from <https://github.com/Mingxue-Xu/graph-hypoth/blob/main/docs/research/which-agents-go-local.md>
- Mingxue-Xu. (2026). *GraphHypoth legacy: Architecture design specifications* (Commit 5064d3dd839a54e914de12dc4a3ca535d164c389) [Repository documentation]. GitHub. <https://github.com/Mingxue-Xu/graph-hypoth-legacy/tree/5064d3dd839a54e914de12dc4a3ca535d164c389/docs/architecture-design/specs>

References v

- Potter, Y., Crispino, N., Siu, V., Wang, C., & Song, D. (2026). *Peer-preservation in frontier models* (Version 3) [Preprint]. arXiv. <https://arxiv.org/abs/2604.19784v3>
- Rajasekaran, P., Dixon, E., Ryan, C., & Hadfield, J. (2025, September 29). *Effective context engineering for AI agents*. Anthropic. <https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents>
- Segato, G. (2026, February 5). *Quantifying infrastructure noise in agentic coding evals*. Anthropic. <https://www.anthropic.com/engineering/infrastructure-noise>
- Wang, H., Poskitt, C. M., & Sun, J. (2026). AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. In *Proceedings of the 48th IEEE/ACM International Conference on Software Engineering* (pp. 2938–2950). Association for Computing Machinery. <https://doi.org/10.1145/3744916.3764546>
- Xu, M., Sharmin, S., & Mandic, D. P. (2024). *Geometry is all you need: A unified taxonomy of matrix and tensor factorization for compression of generative language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2410.03040>

Scoring details A1: association signals

All scores use the legacy defaults. Write $[x] = \text{clip}_{[0,1]}(x)$.

Lexical score

$$L = a[0.5J + 0.3W + 0.2U]$$

- $a = 1$ when any label, alias, or synonym term matches the paper, otherwise 0.
- J : Jaccard similarity of label, alias, and definition/scope terms to paper terms.
- W : matched label/alias weight divided by total weight. Label terms weigh 3, aliases 2.
- U : matched synonym count divided by the number of distinct label/alias terms.

Paper terms come from title and quote. Text is lowercased, tokenized, and stripped of stopwords. Empty denominators give 0.

Mingxue-Xu (2026), GraphHypoth legacy, spec 04 (commit 5064d3d)

Embedding score

$$E = \max_k [\cos(p, a_k)]$$

p : title/quote embedding. Each anchor a_k combines a label or alias with the target's descriptive context. Missing or failed embedding uses $E = L$.

Citation score

$$C = \max(d, 0.5n, 0.3b)$$

d, n, b : maximum overlap with evidence for the same target, adjacent targets, and the current batch, excluding the record itself.

Pair overlap is the maximum of reference-set Jaccard similarity and a direct-citation indicator (0 or 1).

Scoring details A2: association and retrieval ranking

Source prior P Within each source in the scored batch, normalize its retrieval scores:

$$P = \frac{s - s_{\min}}{s_{\max} - s_{\min}} \quad (\text{all equal: } P = 1). \quad P_{\text{missing score}} = \frac{1/r - 1/n}{1 - 1/n}$$

For the fallback, $r = \max(\text{int}(\text{rank}), 1)$ and n is the whole batch size. If $n \leq 1$, use $P = 1$.

Combined association and selection

$$A = \text{round}_6[0.55(0.8E + 0.2L) + 0.35C + 0.10P]$$

Admit only if $\max(L, E, C) \geq 0.05$. Keep the top 5 by A , breaking ties by evidence ID. The prior does not affect admission.

Other retrieval scores

- Optional coherence: $\text{relatedness} = 0.6 E_{\text{claim}} + 0.4 C_{\text{overlap}}$. The association scorer does not read this field.
- Planned retrieval: $\text{RRF}(e) = \sum_{q: e \text{ retrieved}} 1/(60 + r_{qe})$, where r_{qe} is the one-based rank in query q .
- Experiment-methods retrieval can rerank by query-to-paper embedding cosine before truncation.

Mingxue-Xu (2026), GraphHypoth legacy, spec 04; retrieval paths in spec 00 (commit 5064d3d)

Scoring details B1: evidence appraisal and strength

Write $[x] = \text{clip}_{[0,1]}(x)$. The reviewer supplies the relation and methods judgments.

Relation label r

$$R_r = [0.5e_r + 0.2c + 0.3m - 0.5k_r]$$

e_r : entailment, c : context fit, m : construct match, k_r : contradiction risk. Choose the label with maximum R_r .

Methods quality

$$M_e = [0.35d + 0.20v + 0.15p + 0.30a - 0.40b]$$

d : design strength, v : measurement validity, p : population fit, a : confounder adjustment, b : bias risk.

Quote quality

$$Q_e = [0.4x + 0.2c_q + 0.2t_q + 0.2z - 0.5f_q]$$

x : span, c_q : context coverage, t_q : target mention, z : quote status, f_q : summary fallback.

Workflow: $Q_e = 0.6$ for a non-empty quote, otherwise 0 ($x = z$ from quote presence, $c_q = t_q = 0$).

Mingxue-Xu (2026), GraphHypoth legacy, spec 05 (commit 5064d3d)

Evidence strength for verdict v

$$s_{ev} = [0.45I_{ev} + 0.45M_e + 0.10Q_e]$$

$I_{ev} = 1$ if the selected relation label supports verdict v , otherwise 0. The numeric R_r is not used in this sum.

Verdict	Supporting labels	Counter labels
support	supports, mechanism	contradicts, confounder
contradict	contradicts	supports, mechanism
qualify	qualifies	none
not causal	confounder	supports, mechanism

Counter evidence uses the same s_{ev} formula, with $I_{ev} = 0$.

Scoring details B2: verification confidence and verdict

For each candidate verdict v , let S_v contain positively scored supporting evidence and K_v contain counter evidence.

$$C_v = \text{clip}_{[0,1]}(0.55T_v + 0.25B_v + 0.20D_v - 0.25R - 0.15Q)$$

$$T_v = \text{mean}(\text{largest up to 3 } s_{ev} \text{ in } S_v)$$

$$B_v = \text{clip}_{[0,1]} \left(\frac{\sum_{e \in S_v} s_{ev}}{\sum_{e \in S_v} s_{ev} + \sum_{e \in K_v} s_{ev} + 10^{-9}} \right)$$

$$D_v = \min(1, \# \text{ distinct sources in } S_v / 3)$$

$$R = \min(1, \# \text{ open risks} / 3)$$

$$Q = \min(1, \# \text{ qualifiers} / 3)$$

$T_v = 0$ when there is no supporting evidence. Source identity uses `source_id`, falling back to `source`.

Verdict selection

1. Compute C_v for support, contradict, qualify, and not causal.
2. Rank by confidence, then total supporting strength, then the listed verdict order.
3. Return insufficient if every S_v is empty, the top confidence is below 0.5, or the top-two gap is below 0.025.
4. Otherwise select the leading verdict.

The reported confidence remains the highest C_v , even for an insufficient verdict.

Mingxue-Xu (2026), *GraphHypoth legacy*, spec 05 (commit 5064d3d)

Scoring details C1: hypothesis eligibility and ranking

Write $[x] = \text{clip}_{[0,1]}(x)$. Candidate anchors use the labels and aliases of their own proposed nodes.

Embedding measurements

$$n_d = [1 - \max \cos(\text{anchor}, \text{committed nodes})]$$

$$d_d = \max \cos(\text{anchor}, \text{reference anchors})$$

$$f_d = [\cos(\text{anchor}, \text{claim})]$$

Duplication references include committed nodes and other non-empty candidate anchor strings unequal to this anchor. Cosines are clipped to $[0, 1]$.

Fusion with proposer ratings

$$n = [0.7n_{\text{LLM}} + 0.3n_d]$$

$$f = [0.7f_{\text{LLM}} + 0.3f_d]$$

$$d = [0.3d_{\text{LLM}} + 0.7d_d], \quad t = t_{\text{LLM}}$$

Eligibility requires $n \geq 0.4$, $t \geq 0.5$, $f \geq 0.5$, and $d < 0.8$.

Mingxue-Xu (2026), GraphHypoth legacy, spec 06 (commit 5064d3d)

Ranking after the gate

$$H = \text{clip}_{[0,1]} \left(\frac{N + P + T + Y + G + M}{6} \right)$$

N : panel-median field novelty, falling back to gate novelty if ungraded. P, T, Y, G, M : proposer ratings for plausibility, testability, expected yield, centrality, and mechanism specificity.

$$\text{RankScore} = \text{clip}_{[0,1]}(H(1 - S))$$

S : panel-median saturation when graded. Its initial value is 0. The default enables this saturation penalty.

Sort eligible candidates by descending RankScore, then candidate ID. Surface the first 5 before any user-emphasis reranking.

Scoring details C2: critic, emphasis, and experiment grades

Critic panel

$$N = \text{median}_j N_j, \quad S = \text{median}_j S_j$$

$$\text{Borda}(c) = \sum_{j: c \text{ ranked}} (n_j - 1 - p_{jc})$$

p_{jc} is the zero-based position of candidate c in judge j 's list of length n_j . Unranked candidates receive 0.

Ties use candidate ID.

Borda provides panel consensus ordering. The final hypothesis order uses RankScore.

Boolean flags require a strict majority. Mechanism and terminology verdicts use majority votes, with ties favoring more severe concerns.

Mingxue-Xu (2026), [GraphHypoth legacy, specs 06–08 \(commit 5064d3d\)](#)

User emphasis after top-five selection

$$a_c = \max\{w : \text{authored term matches}\}$$

Match authored terms as case-insensitive substrings of new-node labels. No match gives $a_c = 0$.

- Author-directed: sort by a_c , then RankScore, both descending.
- Blended: sort by $0.5a_c + 0.5\text{RankScore}$, descending.

Both break remaining ties by candidate ID.

Experiment grades The LLM supplies clarity, validity, robustness, feasibility, reproducibility, and an overall score. The specs define *no deterministic formula* for the overall score.

Scoring details D1: extraction helpers

Write $[x] = \text{clip}_{[0,1]}(x)$ and $\epsilon = 10^{-9}$. **Concept merge**

$$\text{Merge} = \text{clip}_{[0,1]} \left(\frac{0.273J + 0.318E + 0.182D - 0.227T}{1 + \epsilon} \right)$$

J : label/alias token Jaccard, E : embedding cosine, D : definition-token Jaccard. Type conflict T is 0 for identical types, 0.3 within a type class, and 1 otherwise. Merge at ≥ 0.55 without adjudication. With adjudication: merge at ≥ 0.60 , keep distinct below 0.30, otherwise ask the adjudicator.

Scope gap

$$\text{ScopeGap} = [0.5M_q + 0.2U + 0.3B], \quad M_q = \frac{\text{missing-facet weight}}{4.5 + \epsilon}, \quad U = 1 - \frac{\# \text{distinct defined terms}}{\# \text{distinct salient terms} + \epsilon}$$

Population, exposure/intervention, and outcome weigh 1 each. Comparator, context, and time weigh 0.5 each. $B = 1$ for scope or granularity conflict, otherwise 0. The declared gap threshold is 0.5.

Extraction confidence

$$\text{ExtractConf} = \text{clip}_{[0,1]} \left(\frac{0.45 \text{spanSupport} + 0.35 \text{qualifierCoverage} - 0.40 \text{assumptionPenalty}}{1.2 + \epsilon} \right)$$

The standard LLM extraction path does not invoke these scores: it supplies no merge pairs or scope-gap inputs, and ExtractConf has no production caller. [Mingxue-Xu \(2026\)](#), [GraphHypoth legacy](#), [spec 02 \(commit 5064d3d\)](#)

Scoring details D2: optional verification scheduling

Priority, uncertainty, and evidence gap

$$p = \begin{cases} \text{authored user priority} & \text{if set} \\ 0.5 & \text{otherwise} \end{cases}$$

$$u = \begin{cases} 1 & \text{edge is unverified} \\ \text{clip}_{[0,1]}(1 - \text{confidence}) & \text{otherwise} \end{cases}$$

$$g = \begin{cases} \frac{|\text{required roles} \setminus \text{observed roles}|}{|\text{required roles}|} & \text{required roles non-empty} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{SelectionScore} = \text{clip}_{[0,1]} \left(\frac{p + u + g}{3} \right)$$

Keep a candidate only when it is selectable, SelectionScore is at least 0.5, and its cost fits the supplied budget. Input order is preserved.

Execution boundary: the standard workflow makes one eager verification pass. The specs report no production callers of these scheduling helpers. [Mingxue-Xu \(2026\)](#), [GraphHypoth legacy, specs 03 and 05 \(commit 5064d3d\)](#)