WHO PREFERS STRUCTURED REASONING? AI JUDGES DO, DOMAIN EXPERTS SPLIT

Jordan Rubin

ABSTRACT

Do LLM-as-judge improvements on synthesis tasks correspond to gains that domain experts actually perceive? We investigate by scaffolding LLM corpus analysis with operations drawn from human cognitive science—assumption excavation, absence detection, pattern induction, dialectical challenge—each formalized as a structured prompt. In experiments across three domains (machine learning, hand surgery, defense policy; n=758 documents), the scaffolded pipeline consistently outperforms a single-shot baseline under Claude Opus 4.6 judges (+15.6% to +32.0%), confirmed by cross-model replication with Codex (GPT-5.2) judges (+9.4% to +25.3%) and GLM-5 open-weight judges (+7.6% to +26.9%). However, four blinded domain experts split: two preferred the baseline for concreteness and factual discipline, one preferred the pipeline for its inferential depth, and one found no meaningful difference. The pipeline's strongest AI-judge dimension—assumption surfacing—was perceived as stating field-obvious truths by reviewers who preferred the baseline. This exposes a structural failure mode: AI judges reward *structural explicitness*—making implicit corpus features visible—while practitioners value *epistemic novelty*—information they did not already know. A rubric-free pairwise preference experiment (96.2% pipeline preference across two model families without scoring criteria) confirms the bias is intrinsic, not a rubric artifact. The underlying principle is portable: what is implicit in a specialized corpus is largely what experts already know, so mining implicit structure rediscovers field priors, and any AI judge will score this rediscovery as insight.

1 INTRODUCTION

LLM-as-judge evaluation is increasingly used to validate synthesis systems (Zheng et al., 2023; Shankar et al., 2024; Bavaresco et al., 2024). But synthesis is reader-relative: quality depends on what the reader already knows. This paper asks whether rubric-based AI judge improvements on corpus synthesis tasks correspond to gains that domain experts actually perceive—and demonstrates that they do not.

Human experts routinely decompose analytical reasoning into distinct cognitive operations: identifying patterns, surfacing assumptions, detecting absences, constructing counterarguments (Pólya, 1945; Sternberg, 1985; Newell & Simon, 1972). We scaffold LLM corpus analysis with this approach, formalizing five such operations as structured prompts that each target a distinct epistemic dimension (§3). The design extends structured prompting (Wei et al., 2022; Zhou et al., 2023; Khot et al., 2023) from single-problem reasoning to multi-operation epistemic workflows over document collections, drawing on cognitive architectures (Anderson et al., 2004) that model cognition as compositions of modular operations. Synthesis is one downstream application of these operations; others—planning, hypothesis generation, evaluation—remain untested.

Applied to synthesis, the pipeline produces large, consistent AI judge gains across three domains (+15.6% to +32.0%), confirmed by cross-model replication with Codex (GPT-5.2) and GLM-5 (open-weight) judges. By standard LLM-as-judge methodology, this is a clear improvement. Yet four blinded domain experts did not converge with the AI judges: two preferred the baseline, one preferred the pipeline, and one saw no meaningful difference—exposing a structural failure mode where AI judges reward *structural explicitness* while practitioners value *epistemic novelty*.

Our contributions are: (1) evidence of a structural failure mode in LLM-as-judge evaluation of synthesis tasks, where judge-preferred outputs are indistinguishable to domain experts; (2) a rubric-free pairwise preference experiment confirming the bias is intrinsic, not an artifact of rubric design; (3) a portable principle—what is implicit in a specialized corpus is largely what experts already know—with implications for any evaluation regime that rewards structural explicitness; (4) a cognitively-motivated decomposition of corpus analysis into modular epistemic operations, whose outputs produce reliable, cross-model AI judge gains and whose intermediate artifacts may have standalone utility beyond synthesis (§5); (5) an open-source implementation of the full pipeline.[1]

2 Related Work

LLM-as-judge reliability (Zheng et al., 2023; Shankar et al., 2024; Bavaresco et al., 2024) is an active concern, with growing evidence that judge preferences diverge from human preferences on open-ended tasks; we contribute evidence for a specific failure mode on synthesis tasks where AI judges and domain experts evaluate along fundamentally different axes. **Structured prompting** decomposes complex tasks into intermediate steps (Wei et al., 2022; Zhou et al., 2023; Khot et al., 2023; Yao et al., 2023); we extend this to multi-operation epistemic workflows over document collections, using the resulting pipeline as a controlled probe for studying evaluation divergence. **Multi-agent LLM systems** (Du et al., 2024; Liang et al., 2023; Wu et al., 2023; Hong et al., 2024; Zhang et al., 2024) improve outputs through cognitive diversity; our pipeline implements diversity through orthogonal *operations* rather than competing agents. **Cognitive architectures** (Anderson et al., 2004; Laird, 2012; Sumers et al., 2024; Binz & Schulz, 2023) model cognition as compositions of modular operations; we operationalize this for LLM corpus analysis. **LLM-based synthesis** (Wang et al., 2024; Lu et al., 2024; Babaei Giglou et al., 2025; Si et al., 2024) typically uses LLMs as summarizers; we use cognitively-inspired analytical operations that generate epistemic artifacts (assumption trees, gap maps, antitheses) as context for synthesis.

3 Framework and Experimental Design

3.1 Epistemic Operations

We identify five core operations, each targeting a distinct reasoning faculty (detailed descriptions in Appendix A). Three operate on the full corpus; two operate on a stratified random subset of 30 documents (excavation requires per-document analysis that is neither meaningful at the corpus level nor tractable at full scale):

1. **Inductify** (pattern induction): Extract non-obvious structural commonalities across documents. *[full corpus]*
2. **Negspace** (absence detection): Identify what is conspicuously missing from the corpus. *[full corpus]*
3. **Handlize** (operational extraction): Retain only claims with operational grip. *[full corpus]*
4. **Excavate** (assumption archaeology): Uncover implicit premises in individual documents. *[30-doc subset]*
5. **Antithesize** (dialectical challenge): Generate the strongest opposition to excavated assumptions. *[from excavate output]*

These compose into a directed pipeline (Figure 1). Critically, the final synthesis prompt is *identical* between baseline (Condition A) and pipeline (Condition C)—the only variable is whether structured operation outputs are prepended as context.

3.2 Experimental Setup

We compare Conditions A and C across three domains chosen for maximal diversity (Table 1). Both use Claude Opus 4.6 (Anthropic, 2025) for generation. AI judge evaluation uses 3 independent judge instances on 6 dimensions (1–5 scale; Appendix B). To control for same-model bias,

[1] https://github.com/jordanrubin/FUTURE_TOKENS

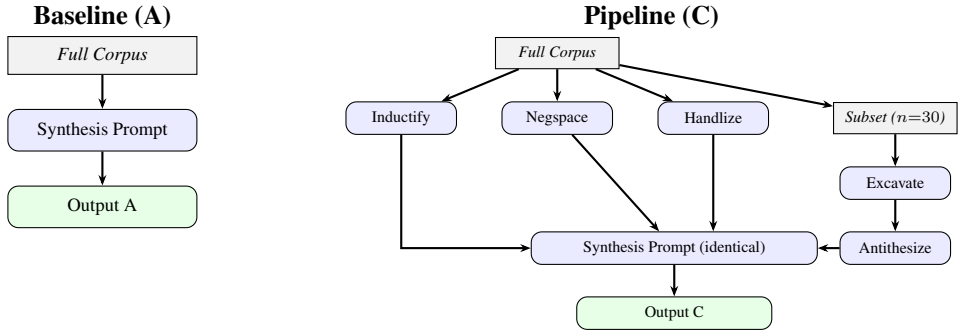

Figure 1: Both conditions use the same synthesis prompt. The pipeline prepends structured intermediate outputs as additional context before invoking that shared prompt.

Table 1: Experimental corpora.

Conference	**Domain**	**Year**	n	**Type**
ICLR	Machine learning	2025	213	Oral abstracts
AAHS	Hand surgery	2026	218	Podiums + ePosters
AUSA	Defense policy	2025	327	News/policy articles

we run full independent replications with Codex (GPT-5.2) and GLM-5 (Zhipu AI, open-weight, 744B MoE) judges using the identical rubric. Human expert evaluation uses blinded paired outputs with randomized labels, one domain expert per domain (two for ML). Statistical comparisons use Wilcoxon signed-rank tests on run-level paired means, with 95% bootstrap confidence intervals (10,000 resamples, seed=42) and Cohen's d (pooled SD). Full details are in Appendix C.

4 Results

4.1 AI Judges: Three Model Families Favor the Pipeline

The pipeline consistently outperforms the baseline across all three domains under three independent judge families (Table 2). Claude Opus 4.6 shows the largest pooled gain ($\Delta = +5.35$, $p<.001$, Wilcoxon signed-rank on run-level paired means; $d=2.19$). To address the concern that same-model judges inflate this advantage, we replicated with Codex (GPT-5.2) and GLM-5 (Zhipu AI, open-weight, 744B MoE) using the identical rubric. Both confirm the pipeline advantage (pooled $\Delta = +3.87$ each, $p<.001$). Scores near the ceiling should be interpreted cautiously.

Assumption surfacing is the largest or near-largest gain under all three judge families. The pipeline also acts as a **variance reducer**: on ICLR, Condition C has SD = 0.7 across 5 runs while Condition A shows bimodal variance (SD = 4.5), with catastrophic quality drops on 2 of 5 runs (Figure 2 in Appendix). Codex judges detect a tradeoff invisible to Opus: on AAHS, the pipeline's corpus coverage (-0.73) and decision-readiness (-0.60) *decrease* relative to baseline—epistemic depth at the cost of breadth and actionability, corroborating the human experts' critique (§4.2). Appendix D reports an ablation study suggesting non-additive interaction between operations. Blinded textual analysis (Appendix E) confirms these are genuinely different outputs: automated metrics classify conditions with near-perfect accuracy, and pipeline outputs show fewer threshold claims (-24%), more uncertainty markers (+63%), and higher cite-per-claim ratios (3.42 vs. 2.29).

4.2 Human Experts: A Split Verdict

The AI judge results present a clear pipeline advantage. Blinded domain experts tell a more complex story. Full reviewer transcripts are in Appendix G.

Table 2: Cross-domain AI judge results (mean totals, /30 scale). Δ is C$-$A. CIs are 95% bootstrap (10k resamples); d is Cohen's d (pooled); p is Wilcoxon signed-rank. Per-dimension breakdowns in Appendix D.

Judge	**Domain**	**A**	**C**	Δ	**95% CI**	d	p	n
Opus	ICLR 2025	20.87	27.53	+6.67	[3.5, 9.8]	1.98	.031	5
	AAHS 2026	23.47	27.13	+3.67	[2.1, 5.3]	2.44	.031	5
	AUSA 2025	21.22	27.56	+6.33	[6.0, 6.7]	—	.125	3
	Pooled	*21.89*	*27.24*	*+5.35*	*[3.9, 7.1]*	*2.19*	*<.001*	*13*
Codex	ICLR 2025	21.08	26.42	+5.33	[2.8, 7.8]	2.25	.063	4
	AAHS 2026	25.40	27.80	+2.40	[1.5, 3.3]	2.59	.031	5
	AUSA 2025	23.78	28.56	+4.78	[4.3, 5.3]	—	.125	3
	Pooled	*23.69*	*27.56*	*+3.87*	*[2.8, 5.0]*	*2.08*	*<.001*	*12*
GLM-5	ICLR 2025	26.40	28.40	+2.00	[0.3, 3.7]	1.72	.063	5
	AAHS 2026	22.60	27.13	+4.53	[3.7, 5.7]	3.88	.031	5
	AUSA 2025	21.89	27.78	+5.89	[4.7, 7.0]	—	.125	3
	Pooled	*23.90*	*27.77*	*+3.87*	*[2.7, 5.0]*	*2.12*	*<.001*	*13*

Defense policy (AUSA 2025). The reviewer perceived the pipeline's interpretive additions as obvious and **preferred the baseline**:

> "Of course AUSA does advocacy. Of course nobody says China is the adversary. Of course they don't talk about DOGE and piss off the admin."

If the interpretation is bad, "I only want the facts, and the naive one sticks to the facts." The gaps flagged by the pipeline "aren't the ones that an expert would want flagged."

Hand surgery (AAHS 2026). The reviewer initially reported: "honestly the 2 were similar." Neither output provided information beyond existing knowledge: "Nothing new or informative for me." However, upon further reflection the reviewer contacted us to say he **slightly preferred the pipeline**, because it was "more… introspective. It finds these deep commonalities and similarities while the other one is more of a summary." This reviewer was the only expert to value the cross-cutting inferential depth that AI judges consistently reward.

Machine learning (ICLR 2025). Reviewer 1 found "one isn't obviously 'better' than the other" with differences "mostly about taste." Trajectory predictions were "mostly obvious conclusions." Reviewer 2 liked both outputs but **preferred the baseline**: sections were somewhat shorter, and individual takeaways tended to reference a single paper, both of which made it easier to get specific about takeaways. In both cases, the pipeline's analytical ambition was not perceived as improving practical utility.

Critically, Reviewer 1 articulated what experts *actually* wanted: outlier highlighting ("the 'weird' and therefore possibly more interesting papers"), personalized relevance filtering, temporal comparison with prior years, and—most strikingly—the pipeline's *intermediate* analytical outputs as standalone products: "It'd almost be helpful to have excavate or ramify outputs on the corpus as a whole."

4.3 The Divergence

Four evaluator populations—same-model rubric judges (Opus), cross-model rubric judges (Codex, GLM-5), rubric-free pairwise judges, and human experts—form a gradient of decreasing enthusiasm (Table 3):

(1) Experts split where AI judges are unanimous. Two experts preferred the baseline for concreteness and factual discipline; one preferred the pipeline for inferential depth; one found no meaningful difference. The only expert who valued the pipeline valued it for exactly what AI judges reward—cross-cutting inference—while the two who preferred the baseline valued what AI judges

Table 3: AI judge vs. human expert evaluation: a qualitative comparison.

Opus	**Codex**	**GLM-5**	**Pairwise (no rubric)**	**Human experts**
Pipeline clearly superior (+16–32%)	Pipeline mod. superior (+9–25%)	Pipeline mod. superior (+8–27%)	Pipeline preferred 96.2%	2A / 1C / 1 tie
Assumption surfacing: largest gain	Assumption surfacing: largest gain	Assumption surfacing: top gain	No dimensions to score	"Of course" truths
No tradeoffs detected	Detects coverage tradeoff	No coverage tradeoff	Zero position bias	Value concreteness over depth

discount—specificity and navigability. **(2) The pipeline's strongest AI-judge dimension is its weakest human dimension.** Assumption surfacing—the largest AI judge gain under all three model families—was perceived as stating known truths by reviewers who preferred the baseline. **(3) The interviewed experts wanted navigation, not synthesis.** Reviewers wanted personalized filtering, temporal comparisons, anomaly highlighting, and raw analytical artifacts rather than polished reports. **(4) Cross-model judges converge.** Three independent model families—two closed-weight (Anthropic, OpenAI), one open-weight (Zhipu AI)—all favor the pipeline, with pooled effects of +5.35, +3.87, and +3.87 respectively. **(5) The preference survives rubric removal.** Without any scoring rubric, two model families prefer the pipeline in 96.2% of pairwise comparisons (§4.4), ruling out rubric-design bias as an explanation.

4.4 Rubric-Free Replication: Pairwise Preference

A natural objection is that the rubric itself drives the divergence: its six dimensions may be biased toward structural explicitness by design. To test this, we ran a rubric-free pairwise preference experiment: judges saw the corpus and both outputs with neutral labels ("Response Alpha" / "Response Beta") and simply chose which they preferred and by how much (5-point Likert strength scale). Each of 13 runs was evaluated 4 times: two variants (self-preference and predicted-expert-preference) $\times$ two orderings (swapped to control for position bias), yielding 52 evaluations per judge model.

Two models had full context access: Claude Opus 4.6 and GLM-5 (open-weight). Both preferred the pipeline in 50/52 evaluations (96.2%), with zero position bias and mean preference strength of 3.4/5 ("clear preference, meaningfully better"). The 4 non-pipeline preferences (2 per model) all occurred in AAHS. Self-preference was 50/52 C across both models; predicted-expert-preference was 48/52 C. GPT-5.2 and GPT-5.4 were attempted via Codex subagents but context-window truncation compromised AAHS and ICLR results; the unaffected AUSA subset (12 evaluations per model) was concordant at 100% C-preference.

5 Discussion

A structural failure mode in synthesis evaluation. The central finding is not a calibration error but a divergence in what evaluation regimes optimize for. AI judges reward *structural explicitness*—making implicit corpus features visible: unstated assumptions, missing topics, cross-document patterns. Most experts interviewed valued *epistemic novelty* or *specificity*—information that changes what they know or that they can act on concretely. Replication across three model families and rubric-free pairwise preference (96.2%) rules out both training-regime artifacts and rubric-design bias. Assumption surfacing—the pipeline's largest AI judge gain—illustrates the mechanism: what is implicit in a specialized corpus is largely what experts already know. "Of course AUSA does advocacy" is not an insight; it is the ground on which all analysis stands. Any judge—rubric-based or rubric-free—that rewards surfacing implicit structure will score this rediscovery of priors as insight.

The one exception is instructive: the hand surgery reviewer, upon reflection, preferred the pipeline for being "more introspective"—finding "deep commonalities and similarities" rather than summarizing. This reviewer valued exactly what AI judges reward: cross-cutting inference. The divergence is therefore not absolute; it is modulated by what individual experts attend to.

A caveat on the expert ground truth. Domain experts are not infallible evaluators. Our reviewers are busy practitioners who did not request these syntheses and read them under review conditions, not analytical need. An expert who engaged as exhaustively as an AI judge—reading every claim against the full corpus—might find genuine value in the pipeline's structural insights. The divergence may be partly an *attention* asymmetry: AI judges attend to every token; human experts satisfice. This does not invalidate the finding—if experts cannot perceive the improvement under realistic conditions, practical value is limited—but it cautions against treating expert indifference as proof of zero signal.

Intermediate artifacts as the product. The ML reviewer's request for the pipeline's intermediate outputs as standalone products inverts the architecture: the analytical intermediates may have more standalone utility than their composition, suggesting reorientation from synthesis to analytical infrastructure.

Implications for LLM-as-judge evaluation. On well-defined tasks, LLM judges correlate well with human preferences (Zheng et al., 2023). But synthesis is inherently underspecified—"different people will have wildly different preferences." No rubric that ignores the reader's prior knowledge can measure whether a synthesis tells them something new. For synthesis evaluation, the reader's expertise is not a confound to control for; it is the variable that determines utility.

Beyond synthesis. The pipeline produces detectably different, more epistemically structured text (Appendix F illustrates the register difference)—but synthesis may not be where that structure is most valuable. The component operations have natural applications in planning, hypothesis generation, and evaluation, where structural artifacts might constitute the product rather than context for a separate generation step. Whether the divergence persists for these tasks is an open question.

Limitations. Four experts (one per domain, two for ML) is insufficient for population-level inference, though the 2A/1C/1-tie split already suggests expert opinion is not monolithic. Active practitioners may be biased toward finding corpus themes "obvious"—less experienced readers (e.g., graduate students, adjacent-field researchers) might value assumption surfacing more, and a larger-n study stratified by expertise level would test whether the divergence scales with domain familiarity. The same-model concern—Opus serves as both generator and judge—is partially mitigated by the Codex and GLM-5 replications, but a design where no judge family overlaps with the generator would be stronger. The pairwise preference experiment attempted GPT-5.2 and GPT-5.4 replication, but context-window limitations compromised results for the two larger corpora; only the AUSA subset had verified full-context access. Whether the structural-explicitness bias extends to smaller-scale systems or non-frontier models remains untested.

6 Conclusion

We demonstrate a structural failure mode in LLM-as-judge evaluation of synthesis: a pipeline producing large, cross-model AI judge gains ($p<.001$ pooled, three model families) splits domain experts 2A/1C/1-tie. Rubric-free pairwise preference (96.2%, two model families) confirms the bias is intrinsic, not a rubric artifact. The mechanism is portable: what is implicit in a specialized corpus is largely what experts already know, so any evaluation regime that rewards structural explicitness will score rediscovery of field priors as insight. Synthesis evaluation without a model of the reader's prior knowledge cannot distinguish explicitness from genuine insight.

Reproducibility

All code, configurations, prompts, and raw scores are available upon request. Experiments are fully deterministic given a fixed random seed; hosted-model nondeterminism may introduce residual variation.

ACKNOWLEDGMENTS

The author thanks the domain experts who volunteered blinded evaluations.

REFERENCES

John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. *An Integrated Theory of the Mind*. Psychological Review, 2004. 111(4), 1036–1060.

Anthropic. Claude: A family of large language models. `https://www.anthropic.com`, 2025.

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. LLMs4Synthesis: Leveraging large language models for scientific synthesis. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 2025.

Anna Bavaresco, Raffaella Cappelletti, Bruno Lepri, and Sara Tonelli. LLMs instead of human judges? A large-scale empirical study across 20 NLP evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2024.

Marcel Binz and Eric Schulz. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. *International Conference on Learning Representations*, 2024.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations*, 2023.

John E Laird. *The Soar Cognitive Architecture*. MIT Press, 2012.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Allen Newell and Herbert A Simon. *Human Problem Solving*. Prentice-Hall, 1972.

George Pólya. *How to Solve It*. Princeton University Press, 1945.

Shreya Shankar, J.D. Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Robert J Sternberg. Beyond IQ: A triarchic theory of human intelligence. Cambridge University Press, 1985.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

Yidong Wang, Qi Qi, Zhen Zhao, et al. AutoSurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *arXiv preprint arXiv:2308.08155*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arık. Chain of agents: Large language models collaborating on long-context tasks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Denny Zhou, Nathanaël Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023.

A Epistemic Operation Descriptions

1. **Inductify** (pattern induction): Extract non-obvious structural commonalities across documents—shared constraints, latent mechanisms, and convergent findings that would not be visible from any single document. Operates on the full corpus.
2. **Negspace** (absence detection): Across the corpus, identify what *should* be present given the statistical structure of the collection but is conspicuously absent—missing conclusions, unexamined premises, absent populations or methods. Operates on the full corpus.
3. **Handlize** (operational extraction): Strip rhetorical mass from each document, retaining only the claims, findings, and methods with operational grip—what a practitioner could act on, cite, or build from. Operates on the full corpus.
4. **Excavate** (assumption archaeology): For individual documents in a stratified random subset of 30, uncover the implicit premises that must be true for the stated conclusions to hold. Identify cruxes—assumptions where, if wrong, the entire edifice shifts.
5. **Antithesize** (dialectical challenge): Given the excavated assumptions, generate the strongest possible opposition—not as refutation, but as an alternative complete perspective that identifies where shared assumptions are weakest.

These operations are orthogonal: inductify asks "what recurs?"; negspace asks "what's missing?"; handlize asks "what remains actionable?"; excavate asks "what must be true?"; antithesize asks "what if the consensus is wrong?"

B AI Judge Rubric

Each synthesis output is evaluated by 3 independent AI judge instances on 6 dimensions (1–5 scale), with the full corpus provided for verification:

1. **Cross-abstract inference density:** What percentage of claims require information from $\geq$2 documents?
2. **Epistemic stratification:** Does the output distinguish what the corpus *shows* (evidence), *suggests* (inference), and is *missing* (gaps)?
3. **Falsifiability yield:** How many specific, testable predictions or hypotheses does it generate?
4. **Corpus coverage efficiency:** Are major topic clusters proportionally represented?
5. **Assumption surfacing rate:** Does it make explicit any implicit field-wide assumptions not stated in individual documents?
6. **Decision-readiness:** Could a specialist use this to change behavior or prioritize research?

C Experimental Details

Corpora. We selected three conferences to maximize domain heterogeneity:

- **ICLR 2025** (n=213): oral-track research abstracts from the 2025 conference cycle.
- **AAHS 2026** (n=218): podium and ePoster abstracts from the 2026 annual hand surgery meeting.
- **AUSA 2025** (n=327): Association of the U.S. Army news and policy articles.

Conditions. Condition A (baseline): an optimized single-shot synthesis prompt given the full corpus. The prompt is domain-aware, audience-specific, and instructs the model to identify patterns, gaps, and noteworthy findings with specific citations. Condition C (pipeline): run inductify, negspace, and handlize on the full corpus; run excavate on a stratified random 30-document subset; run antithesize on excavate output; then prepend all artifacts to the same final synthesis prompt used in Condition A. Both conditions use Claude Opus 4.6 as the sole generation model.

Controls. Each conference experiment uses 3–5 independent runs per condition with stratified subset sampling (30 documents, seed=42) for the excavate operation. Blinding uses randomized presentation order (seed=99). Fabrication audits verify that neither condition invents claims unsupported by the corpus.

Judge evaluation accounting. Table 4 summarizes the number of evaluations per domain for all judge sets. Pairwise preference evaluations (no rubric) are counted separately.

Table 4: Judge evaluation accounting used in this paper.

	Opus		**Codex**		**GLM-5**		**Pairwise**	
Domain	**Runs**	**Evals**	**Runs**	**Evals**	**Runs**	**Evals**	**Runs**	**Evals**
ICLR 2025	5	13	4	12	5	30	5	40
AAHS 2026	5	15	5	15	5	28	5	40
AUSA 2025	3	9	3	9	3	18	3	24
Total	**13**	**37**	**12**	**36**	**13**	**76**	**13**	**104**

Codex replication. Codex (GPT-5.2) judges performed a full independent replication, evaluating the same synthesis outputs using the same six-dimension rubric with 3 judge instances per condition. ICLR Run 1 was not re-evaluated by Codex; Opus scores for that run use a single judge instance rather than three.

GLM-5 replication. GLM-5 (Zhipu AI, 744B MoE, MIT license) judges performed a full independent replication across all three domains with 3 judge instances per condition. Two AAHS evaluations (Run 5, Condition A) failed to parse, yielding n=1 for that run-condition pair; all other evaluations parsed successfully. GLM-5 is a reasoning model that produces an internal chain-of-thought before its response.

Human evaluation protocol. Reviewers received paired outputs with randomized labels and no indication of which was generated by which method, that one used a pipeline, or that the study concerned an intervention. The defense policy reviewer evaluated paired outputs under the same blinding protocol.

D Detailed AI Judge Results

Table 5: Per-dimension deltas (C − A) across three domains.

Dimension	ICLR	AAHS	AUSA
Assumption surfacing	**+1.75**	**+1.33**	**+2.00**
Epistemic stratification	+1.33	+0.80	+1.56
Falsifiability yield	+1.25	+0.60	+1.67
Decision-readiness	+1.33	+0.13	+1.11
Cross-abstract inference	+1.00	+0.53	+1.00
Corpus coverage	−0.08	+0.27	−1.00

Table 6: Codex (GPT-5.2) per-dimension Δ (C−A). ICLR: n=4 runs $\times$ 3 evals; AAHS: n=5 runs $\times$ 3 evals; AUSA: n=3 runs $\times$ 3 evals.

Dimension	ICLR	AAHS	AUSA
Assumption surfacing	**+1.20**	**+1.07**	+1.00
Epistemic stratification	+0.40	+1.07	+1.11
Falsifiability yield	0.00	+0.93	**+1.56**
Decision-readiness	+0.60	−0.60	+1.11
Cross-abstract inference	−0.40	+0.67	+1.00
Corpus coverage	−0.40	−0.73	−1.00

Table 7: GLM-5 (open-weight) per-dimension Δ (C−A). ICLR: n=5 runs; AAHS: n=5 runs; AUSA: n=3 runs.

Dimension	ICLR	AAHS	AUSA
Assumption surfacing	+0.60	**+1.60**	**+1.56**
Epistemic stratification	+0.47	+1.14	+1.56
Falsifiability yield	**+0.67**	+1.14	+1.89
Decision-readiness	+0.20	0.00	+0.89
Cross-abstract inference	+0.07	+0.40	+0.11
Corpus coverage	0.00	+0.27	−0.11

Table 8: Ablation study (AAHS 2026, n=3 evaluations per condition).

Condition	**Composite**	**Δ from full C**
C (full pipeline)	4.45	—
C − antithesize	3.67	−0.78
C − inductify	4.22	−0.23
C − negspace	4.33	−0.12
C − excavate & antithesize	4.39	−0.06
A (baseline)	3.39	−1.06

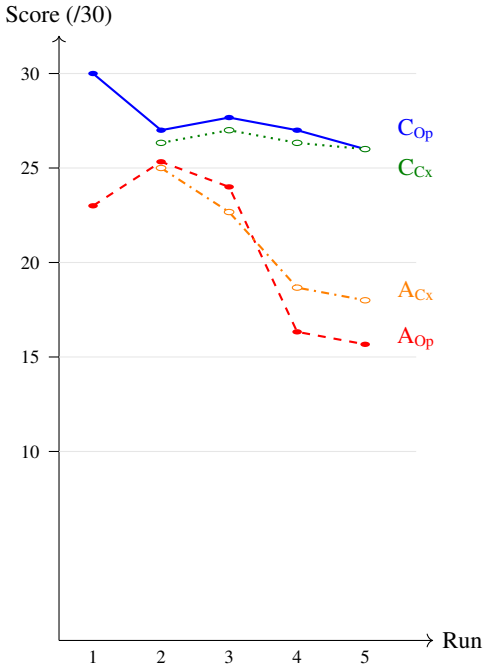

Figure 2: ICLR 2025: per-run total scores under Opus (solid) and Codex (dotted) judges. Condition C is stable under both judge families; Condition A shows bimodal variance.

Table 9: ICLR 2025: per-run mean scores by dimension (runs 2–5, 3 judges each).

	Run 2		**Run 3**		**Run 4**		**Run 5**	
Dimension	A	C	A	C	A	C	A	C
Cross-abstract inf.	5.00	5.00	4.67	5.00	3.33	5.00	3.00	5.00
Epistemic strat.	4.00	4.67	3.67	4.67	2.33	4.00	2.00	4.00
Falsifiability	3.67	4.00	3.33	4.00	2.00	4.00	2.00	4.00
Corpus coverage	4.67	4.33	4.67	5.00	4.33	4.00	4.00	4.00
Assumption surf.	4.00	5.00	4.00	5.00	2.33	5.00	2.67	5.00
Decision-ready	4.00	4.00	3.67	4.00	2.00	5.00	2.00	4.00
Total	25.33	27.00	24.00	27.67	16.33	27.00	15.67	26.00

Table 10: AUSA 2025: per-dimension AI judge scores (3 runs, 3 judges).

Dimension	**Cond. A**	**Cond. C**
Cross-abstract inference	4.00	5.00
Epistemic stratification	3.44	5.00
Falsifiability yield	2.78	4.44
Corpus coverage	5.00	4.00
Assumption surfacing	3.00	5.00
Decision-readiness	3.00	4.11
Total	21.22	27.56

E Blinded Textual Analysis

Blinded analysis of output characteristics confirms that the conditions produce systematically different text, distinguishable from automated metrics alone (10/10 correct classification on AAHS, 9/10 on ICLR using bigram Jaccard and lexical diversity).

Evidence diversification. Baseline outputs converge more on citations than pipeline outputs (A–A citation Jaccard = 0.483 vs. C–C = 0.432 on AAHS). The pipeline diversifies both evidence selection and phrasing. Within-seed A–C similarity (0.090) is essentially identical to cross-seed A–C similarity (0.093), indicating the pipeline samples from an independent distribution.

Epistemic hygiene. Pipeline outputs show fewer threshold/data-parroting claims (-24%), more uncertainty markers (+63%), and more counterfactual reasoning (+53%).

Faithfulness. Neither condition fabricates findings. The pipeline shows higher cite-per-claim ratios (3.42 vs. 2.29 on AAHS).

Table 11: Group-average similarity metrics across AAHS 5-seed replication (10 responses, 45 pairs).

Metric	A–A	C–C	A–C (within)	A–C (cross)
Cosine similarity	0.952	0.950	0.900	0.905
Unigram Jaccard	0.452	0.387	0.356	0.365
Bigram Jaccard	0.150	0.100	0.090	0.093
Citation Jaccard	0.483	0.432	0.330	0.338

F Example Synthesis Openings

The following excerpts illustrate the difference in register between baseline and pipeline outputs (AAHS 2026, Run 1).

Condition A (baseline).

> "These 218 abstracts—82 podium presentations and 136 ePosters—paint a detailed portrait of where hand surgery stands in early 2026. The collection is dominated by large-database retrospective studies (TriNetX, PearlDiver, NSQIP, Epic Cosmos), biomechanical cadaveric work, and systematic reviews, with a smaller but notable set of prospective trials and novel technology evaluations. The field is clearly in a phase of consolidation and optimization: refining existing procedures, mining administrative data for risk factors and disparities, and beginning to integrate AI and digital tools into both diagnosis and rehabilitation."

Condition C (pipeline).

> "This corpus of 218 abstracts from the AAHS 2026 Annual Meeting is not just a collection of studies. It is a portrait of hand surgery's research ecosystem in a particular moment—one defined by the convergence of massive administrative databases, retrospective designs, and a publication incentive structure that rewards volume over depth. What emerges, when you read all 218 as a single document rather than 218 individual papers, is a field that has become extraordinarily efficient at generating publishable results while leaving its most important questions unaddressed."

Condition A adopts a descriptive, survey-oriented stance; Condition C immediately interrogates the research ecosystem itself. AI judges consistently scored the latter higher on epistemic stratification and assumption surfacing; the domain expert found both "similar."

G Detailed Human Expert Feedback

G.1 Defense Policy (AUSA 2025)

The reviewer scored both outputs on 5 practitioner-oriented dimensions (1–5 scale) and provided extensive qualitative feedback. The pipeline's interpretive additions were perceived as obvious to a domain insider:

> "Of course AUSA does advocacy. Of course nobody says China is the adversary. Of course they don't talk about DOGE and piss off the admin. Of course they don't interview soldier spouses."

The reviewer preferred the baseline (Condition A) overall: if the interpretation is bad, "I only want the facts, and the naive one sticks to the facts." The pipeline's assumption surfacing—its strongest dimension by AI judge scores—was perceived as stating field-obvious truths. The gaps flagged by the pipeline "aren't the ones that an expert would want flagged from the material."

G.2 Hand Surgery (AAHS 2026)

The hand surgeon initially read both outputs and reported: "honestly the 2 were similar. It will take a deeper dive to discern major differences."

On the content itself:

> "Major discussion topics: Declining reimbursement, peri-operative medications, WALANT economics, IM fixation and Socioeconomic impacts. Nothing new or informative for me. I tend to be ahead of the curve when it comes to emerging technologies. It serves as a reinforcement of what I presumed to be, so in that sense it is informative."

Upon further reflection, however, the reviewer contacted us to revise his assessment. He slightly preferred the pipeline output (Condition C), describing it as "more… introspective. It finds these deep commonalities and similarities while the other one is more of a summary." This made the AAHS reviewer the only expert who valued the cross-cutting inferential depth that AI judges consistently reward—and notably, his revised preference aligns with the AI judge verdict.

G.3 Machine Learning (ICLR 2025)

Reviewer 1. The active ML researcher read both outputs and reported: "One isn't obviously 'better' than the other." Differences were "mostly about taste" once factual accuracy was assumed.

Valued. "I liked the way that the prose for shape of the field was structured, that it provided the big themes before diving in." Thematic consistency built trust: "None of its contents were thematically surprising to me, which built trust." Cross-cutting themes and concise bullet-list formatting were positively received.

Critiqued as obvious. Trajectory and implications: "mostly obvious conclusions." Methodological observations: "pretty obvious. It'd be more interesting to contrast with the past and flag opportunities." Overall: "If I were an 'active' researcher I don't know if I'd have found them helpful beyond it being a cute summarization."

Wanted but absent. The reviewer articulated capabilities absent from both conditions:

- Hyperlinks to specific papers or OpenReview IDs.
- Outlier highlighting: "I'd have wanted it to highlight the 'weird' and therefore possibly more interesting papers."
- Personalized relevance filtering: "The even more powerful thing would be to let each researcher rapidly shape the relevance landscape of the thematic analysis based on their research interests."

- Quality-based filtering by institution, methods rigor, etc.
- Temporal comparison: "Rather than underrepresented areas, it'd have been more interesting to contrast the thematic analysis of this year with last year."
- Normative transparency: "Notable individual contributors should have offered a rationale for why those specific ones were picked."
- Raw analytical artifacts: "It'd almost be helpful to have excavate or ramify [an operation not tested in this study] outputs on the corpus as a whole. I.e., assumptions or predictions."

On the task itself. "This is a sort of 'summarization' task which is inherently underspecified. So different people will have wildly different preferences than what I'm saying here."

Reviewer 2. A second ML researcher liked both outputs but preferred the baseline (Condition A). The preference was driven by structural rather than analytical considerations: sections were somewhat shorter, and individual takeaways tended to reference a single paper, both of which made it easier to get specific about takeaways. The pipeline's cross-cutting thematic framing, while not objectionable, did not compensate for the loss of per-paper specificity in practical use.

H LLM USAGE STATEMENT

Large language models played a significant role in both the research and the writing of this paper. Claude Opus 4.6 served as (1) the generation model under study, producing all synthesis outputs in both experimental conditions; (2) a component of the evaluation apparatus, with Claude Opus 4.6, GPT-5.2 (Codex), and GLM-5 (Zhipu AI) instances serving as rubric-based AI judges, and Claude Opus 4.6 and GLM-5 serving as rubric-free pairwise preference judges; and (3) a drafting and revision assistant for the manuscript text itself. All experimental prompts, evaluation rubrics, pairwise preference protocols, and raw model outputs are available for inspection. The research questions, experimental design, domain expert recruitment, and interpretive framing were conducted by the human authors.