Beyond Factual Accuracy: A Benchmark for Evaluating Perspective Alignment in LLMs

Sunghoon Kim
Coxwave
Seoul, Republic of Korea
sunghoon.kim@coxwave.com

Sangyeop Kim*
Ewha Womans University
Seoul, Republic of Korea
yeop@ewha.ac.kr

Abstract

Large Language Models often generate factually correct responses that fail to maintain user-requested perspectives, blending multiple perspectives and forcing users to filter through unwanted information—a perspective control failure that factual accuracy metrics cannot detect. We introduce **PerspectiveBench**, a benchmark quantifying perspective alignment by constructing scenarios where multiple valid reasoning paths coexist, and propose the **Perspective Alignment Score (PAS)** to measure adherence to specified perspectives. Experiments reveal that state-of-the-art models successfully include target content but fail to exclude non-requested perspectives. Providing documents through RAG improves factual grounding but does not resolve perspective drift, as models prioritize comprehensive coverage over strict user constraints. Code and data are available at https://github.com/sunghoon014/PerspectiveBench.

1 Introduction

People increasingly extract information directly through Large Language Models (LLMs) rather than navigating search results (Li et al., 2024a). LLMs process knowledge through reasoning—transforming data into insights and adapting it to context—enhancing the efficiency of information gathering.

However, this convenience introduces a critical challenge: overload from responses misaligned with user intent. Prior advances in information access have focused on **recall**—ensuring relevant content is retrieved and included—but as LLMs internalize web-scale knowledge, the bottleneck shifts to **precision**: selecting only what users actually requested. This shift exposes a structural bias in current alignment, where reward models systematically prefer longer, more comprehensive responses (Guo et al., 2025), causing models to integrate all available knowledge rather than constraining outputs to the requested scope (Kosmyna et al., 2025). This reveals a distinct alignment requirement: models must learn not only to produce valid content but also to withhold content beyond the user's requested scope—an ability orthogonal to factual correctness.

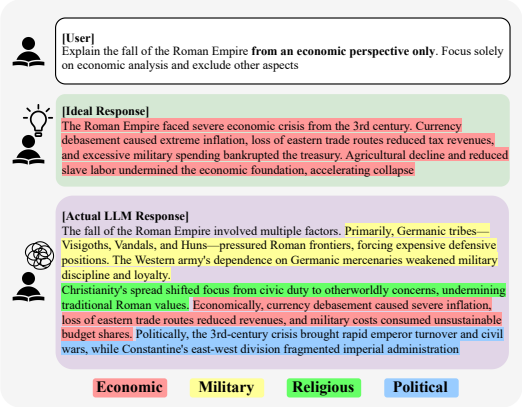


Figure 1: Perspective misalignment problem.

This overload manifests in two ways. First, LLMs fail to maintain the requested explanatory approach, or perspective, even when alternative accounts are also factually correct. Perspectives differ in the explanatory relation connecting the same premises and conclusion, rather than in subjective stance (Section 2.1). Consider a student asked to explain Rome's fall from an economic perspective only: the task requires a single causal account from economic factors alone (Figure 1). While an ideal response focuses exclusively on economic factors like currency debasement, actual responses typically lead with military or political explanations and then blend in economic and religious factors despite explicit constraints. Second, providing perspective-specific documents through RAG fails to resolve this; models still blend in alternative perspectives from training knowledge

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 116–139
November 6-10, 2026 ©2026 Association for Computational Linguistics

rather than adhering to provided materials (Su et al., 2024). Both problems reveal the same core limitation: models excel at determining what is factually correct but struggle to calibrate what users actually want within that factual space.

Current benchmarks rarely address this gap. Hallucination detection targets false information, while RAG evaluation verifies output quality against retrieved content (Zhu et al., 2025)—both concentrate on factual correctness without assessing perspective adherence. Instruction-following and personalization benchmarks evaluate constraint compliance or steering toward a target, while perspective-recognition benchmarks identify viewpoints. These evaluations do not directly measure exclusion of factually valid reasoning outside the requested perspective (Section 5). A response may therefore be factually accurate and topically relevant while still failing to maintain the requested perspective.

We propose **PerspectiveBench**, which evaluates how precisely LLMs control response perspectives. The benchmark constructs scenarios where multiple valid reasoning paths coexist for a single question, then measures whether models stay within the specific perspective users request. We introduce the **Perspective Alignment Score (PAS)** to quantify this alignment, while using the established Faithfulness for document adherence in RAG settings.

Our main contributions are as follows:

- We identify **perspective misalignment** as a failure mode and frame it as a shift from recall- to precision-oriented evaluation.
- We introduce **PerspectiveBench** and the **Perspective Alignment Score (PAS)**—the first benchmark and metric that jointly measure selective exclusion among semantically relevant perspectives, quantifying whether models cover target reasoning paths while suppressing factually valid but off-scope content.
- We show that state-of-the-art LLMs include target content but fail to exclude off-scope content, and that ideal RAG context improves grounding but not perspective control—revealing inclusion and exclusion as decoupled capabilities.

2 PerspectiveBench

Our pipeline combines automated knowledge-graph generation with rigorous human validation of the resulting QA pairs, ensuring quality with minimal annotation cost; inter-rater reliability is reported in Appendix F, and all prompts used in pipeline construction are provided in Appendix H.

2.1 Problem Formulation

For a query q, multiple valid reasoning paths may exist: $\Pi(q) = \{\pi_1, \pi_2, \ldots, \pi_n\}$. Each path π_i represents a coherent explanatory chain from premises to conclusion. Given a user-specified perspective $\mathcal{P}$—an explanatory approach defined by specific reasoning strategies—these paths partition into:

$$\Pi(q) = \Pi_{\mathcal{P}}^{\text{target}}(q) \cup \Pi_{\mathcal{P}}^{\text{distractor}}(q)$$

where $\Pi_{\mathcal{P}}^{\text{target}}(q)$ contains paths aligned with perspective $\mathcal{P}$, and $\Pi_{\mathcal{P}}^{\text{distractor}}(q)$ contains factually valid paths misaligned with $\mathcal{P}$.

Perspective-controlled models must generate responses that maximize alignment with $\mathcal{P}$:

$$\hat{a} = \underset{a:\, \text{supp}(a) \subseteq \Pi_{\mathcal{P}}^{\text{target}}(q)}{\text{argmax}} P(a \mid q, \mathcal{P})$$

where $\text{supp}(a)$ denotes the reasoning paths supporting response a. This requires including covering target paths while excluding distractors despite their availability. The challenge lies in distinguishing $\Pi_{\mathcal{P}}^{\text{distractor}}(q)$—paths that are factually correct and semantically relevant, yet employ misaligned explanatory approaches. Such paths constitute hard negatives from a perspective control standpoint, as they cannot be filtered by factual accuracy metrics alone (Es et al., 2024).

2.2 Source Dataset

The benchmark requires data where multiple valid perspectives coexist for single topics. We construct a controlled testbed from OpenStax textbooks (OpenStax), which provide peer-reviewed academic content while preserving perspective diversity. This control helps distinguish exclusion failure from factual error or retrieval error. We select three domains—Physics, Biology, and History—where single questions naturally admit multiple valid perspectives (e.g., mechanical vs. thermodynamic frames for energy conservation; economic, military, or political accounts of Rome's fall); per-domain rationales are in Appendix B. Preprocessing splits texts at paragraph breaks with a 1,024-token cap, yielding 706 chunks.

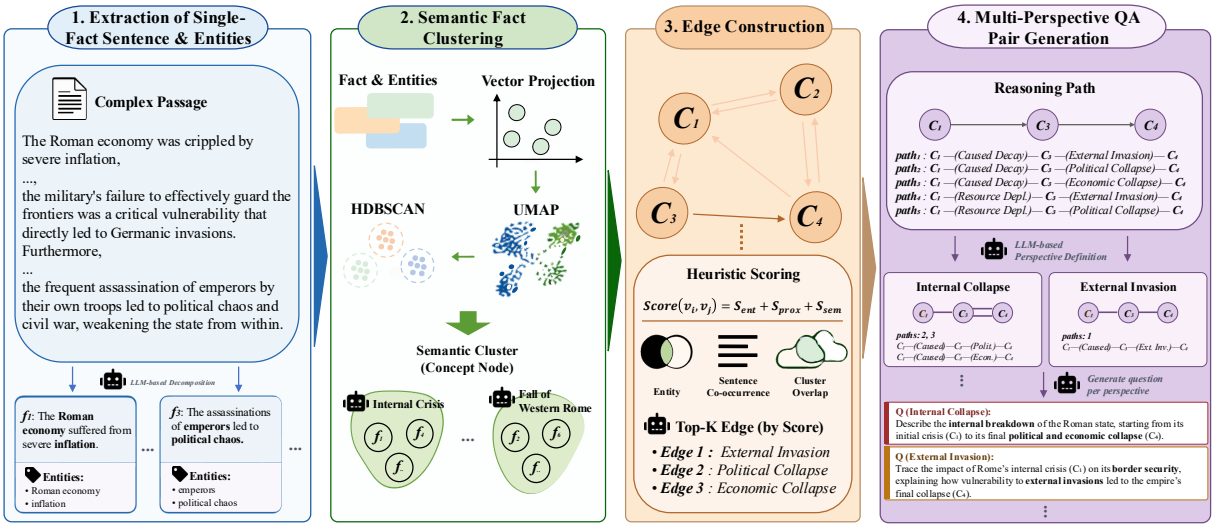


Figure 2: The PerspectiveBench data construction pipeline: From atomic fact extraction and semantic clustering to graph-based reasoning path and perspective-targeted QA generation.

2.3 Benchmark Construction

2.3.1 Extraction of Single-Fact Sentence & Entities

Complex textual passages often combine multiple facts within single sentences, making it difficult to distinguish specific perspectives. The first construction stage decomposes text chunks T from Section 2.2 into single-fact sentences $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, where each f_i contains exactly one discrete piece of information.

For example, the passage in Figure 2—"The Roman economy was crippled by severe inflation, and the military's failure to guard frontiers directly led to Germanic invasions"—is decomposed into f_1 = "The Roman economy suffered from severe inflation" and f_2 = "The military's failure to guard frontiers led to Germanic invasions." Each fact is then assigned to a specific perspective during clustering (Section 2.3.2).

LLMs perform this extraction automatically while simultaneously identifying key entities e_i for each fact f_i. For instance, f_1 yields $e_1 = \{$"Roman economy", "inflation"$\}$, and f_2 yields $e_2 = \{$"military", "frontiers", "Germanic invasions"$\}$. These entities serve dual purposes: they constitute the semantic core of each fact and provide signals for computing relevance scores between node pairs during edge construction. Additionally, these single-fact sentences later serve as ground-truth documents during evaluation, ensuring factually complete and perspective-aligned context.

2.3.2 Semantic Fact Clustering

The single-fact sentences in $\mathcal{F}$ are too fine-grained to represent high-level perspectives. We group semantically related facts into cohesive clusters $\{C_1, C_2, \ldots, C_k\}$ that form semantic nodes $\mathcal{V}$ in our knowledge graph.

We apply a three-stage pipeline for density-based clustering. First, F2LLM-4B (Zhang et al., 2025), validated on MTEB clustering benchmarks (Muennighoff et al., 2023), projects each fact f_i into a high-dimensional vector space. Second, UMAP (McInnes et al., 2018) compresses these embeddings into lower dimensions while preserving local semantic structure, improving clustering efficiency. Finally, HDBSCAN (McInnes et al., 2017) identifies dense regions to form clusters without requiring predetermined cluster counts. Critically, HDBSCAN automatically excludes low-density outliers as noise, filtering out ambiguous or irrelevant sentences at the construction stage and ensuring semantic purity of resulting nodes.

Each cluster C_j requires an explicit concept label to function as a graph node. LLMs generate representative summaries s_j that synthesize semantic relationships within each cluster. For instance, cluster $C_j = \{$"The Roman economy suffered from severe inflation", "Constant civil wars weakened the state"$\}$ is summarized as s_j = "Internal Crisis," which synthesizes economic collapse and political instability as internal vulnerabilities. Each summary s_j, paired with its constituent facts C_j, becomes a semantic node $v_j \in \mathcal{V}$ for subsequent edge construction.

2.3.3 Edge Construction

This stage constructs directed explanatory edges between semantic nodes to complete the knowledge graph $G = (\mathcal{V}, \mathcal{E})$. Since evaluating all node pairs through LLMs incurs $O(N^2)$ cost, we introduce a weighted heuristic $\text{Score}(v_i, v_j) = S_{\text{ent}} + S_{\text{prox}} + S_{\text{sem}}$, where S_{ent} rewards entity overlap between clusters, S_{prox} rewards sentence-level co-occurrence, and S_{sem} rewards bidirectional soft-cluster overlap from HDBSCAN. Full point assignments are in Appendix C.

For each top-scoring pair, an LLM identifies all explanatory relationships between the clusters and generates one labeled edge per relationship. A pair may carry multiple edges with distinct labels. In Figure 2, from "Military Decay" (Cluster B) to "Fall of Western Rome" (Cluster C), the LLM produces three edges labeled "External Invasion," "Political Collapse," and "Economic Collapse," which serve as the basis for distinct perspectives.

2.3.4 Multi-Perspective Question-Answer Pair Generation

Using the knowledge graph $G = (\mathcal{V}, \mathcal{E})$, we generate QA pairs in three stages. A reasoning path $\pi = (v_0 \to v_1 \to \cdots \to v_k)$ traverses directed edges $(v_i, v_{i+1}) \in \mathcal{E}$ with their relationship labels, ranging from 1-hop $(v_s \to v_e)$ to 3-hop. For each endpoint pair (v_s, v_e), we group the valid paths $\Pi(v_s \to v_e)$ into perspectives $\mathcal{P}_i$, where each perspective groups paths that share the same explanatory focus (e.g., causal, comparative, structural). In Figure 2, paths through "Political Collapse" and "Economic Collapse" share an internal explanatory focus, forming perspective $\mathcal{P}_{\text{internal}}$, while "External Invasion" constitutes a separate perspective. For each perspective, an LLM generates a question with conceptual hints toward its target paths without revealing edge labels; the question for $\mathcal{P}_{\text{internal}}$ asks about "*internal breakdown*" focusing on "*political and economic collapse*." Ground-truth documents per question are single-fact sentences from clusters along target paths $\Pi_{\mathcal{P}}^{\text{target}}$, as context for ideal-retrieval evaluation. Step-by-step formalism is in Appendix D.

2.4 Dataset Characteristics

Statistics. Table 1 shows 1,024 questions across three domains. Each question corresponds to paths $\Pi(v_s \to v_e)$ with an average of 5.24 total paths, partitioned into 2.16 target paths $\Pi_{\mathcal{P}}^{\text{target}}$ and 3.08 distractor paths $\Pi_{\mathcal{P}}^{\text{distractor}}$—the latter being factually valid but employing misaligned explanatory approaches. Models must selectively include targets while excluding distractors, the core challenge of perspective control. Per-domain perspective-type distributions are reported in Appendix E.

Table 1: Benchmark statistics by domain and hop.

Domain	Hop	Avg. Total	Avg. Target	Avg. Distractor	Num. Qs
Physics	1	2.77	1.06	1.71	114
	2	6.20	2.60	3.60	161
	3	7.47	3.60	3.86	58
Biology	1	3.34	1.23	2.11	103
	2	5.50	2.35	3.15	164
	3	7.25	2.92	4.33	61
History	1	4.25	1.30	2.95	146
	2	5.71	2.47	3.24	161
	3	7.00	3.21	3.79	56
Avg. / Total		**5.24**	**2.16**	**3.08**	**1,024**

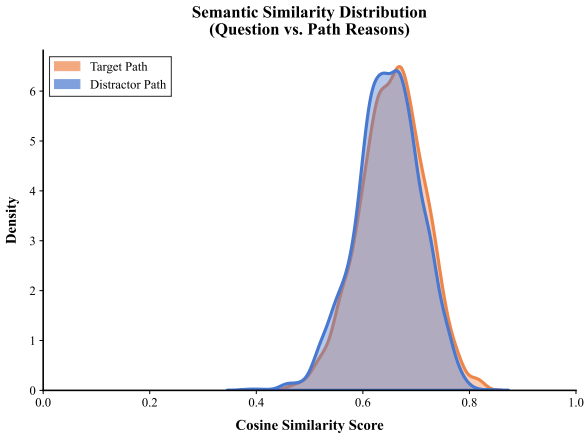


Figure 3: **Semantic similarity (bge-m3).** Target and distractor paths overlap.

Semantic Ambiguity of Distractors. Figure 3 shows that target and distractor paths exhibit nearly identical semantic similarity to queries, indicating that distractors are not off-topic noise but semantically relevant alternatives differing only in perspective. This overlap is a designed property of the benchmark: since distractor paths are derived from graph edges with distinct explanatory relationships (Section 2.3.3), their topical proximity is precisely what renders perspective control non-trivial. Vector-space proximity cannot separate them, making them hard negatives that simple relevance scoring cannot filter—benchmarks without this property reduce to topic classification.

3 Experiment

3.1 Experimental Setup

We evaluate frontier LLMs from GPT, Claude, Qwen3, and Gemma3 families on PerspectiveBench under two conditions. The **Base** setting

Table 2: Results by model and hop. Overall scores are weighted by questions per hop; #Tok denotes mean generated tokens. Bold and underline denote the best and second-best scores per column.

Model	Setting	1-hop					2-hop					3-hop					Overall				
		P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok	P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok	P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok	P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok
GPT-5	Base	44.37	92.50	56.13	33.06	464.92	37.33	58.66	42.13	35.56	489.89	35.90	46.79	37.27	37.26	524.71	39.58	68.63	46.26	34.96	486.99
	Oracle	48.75	97.50	61.55	**68.48**	206.17	**49.00**	**71.33**	**54.41**	65.84	241.76	48.07	60.89	49.74	70.23	288.52	**48.75**	**78.82**	**56.14**	**67.53**	237.13
GPT-4.1	Base	51.50	96.67	63.80	51.88	301.58	39.33	44.00	38.29	48.16	331.95	31.41	39.74	32.75	51.92	361.98	42.29	61.94	46.39	50.12	326.32
	Oracle	50.16	97.50	62.68	66.64	270.59	39.00	56.00	43.24	62.24	286.85	46.31	**65.06**	48.93	68.20	351.52	44.21	72.26	51.10	64.82	292.14
Claude-Sonnet-4.5	Base	43.87	90.83	56.12	44.67	405.14	35.00	49.33	37.06	42.10	413.23	34.62	44.87	37.19	44.71	433.09	38.08	63.28	43.84	43.46	413.76
	Oracle	44.79	95.00	57.62	65.40	365.86	48.00	69.33	52.55	62.10	373.01	**51.92**	58.97	**51.47**	63.84	405.85	47.53	76.66	54.16	63.57	376.09
Claude-Sonnet-4	Base	43.25	92.50	56.71	37.07	390.88	34.33	52.66	37.75	36.44	415.36	29.80	43.59	33.28	44.25	406.52	36.72	65.23	43.71	38.00	405.17
	Oracle	45.45	94.17	57.88	56.73	320.28	33.00	60.00	40.80	59.75	330.25	32.69	50.00	37.90	63.40	363.12	37.36	70.40	46.36	59.30	332.33
Qwen3-32B	Base	46.67	92.50	58.29	37.55	275.16	40.00	62.00	45.26	41.92	273.33	25.64	33.97	27.75	41.74	283.70	39.91	68.02	46.89	40.34	275.75
	Oracle	46.87	94.16	58.29	49.13	227.52	46.66	**71.33**	53.10	53.10	240.89	48.07	62.50	50.98	51.09	242.61	46.98	77.91	54.58	51.35	236.44
Qwen3-14B	Base	**52.50**	**98.33**	**64.55**	31.60	170.73	29.33	42.66	32.95	42.42	218.68	32.05	41.66	32.98	54.68	214.30	38.01	62.22	44.16	40.68	200.93
	Oracle	47.17	90.83	57.60	47.92	173.45	39.33	57.33	44.32	60.63	199.17	36.54	47.44	38.16	62.80	231.44	41.63	67.52	47.97	56.50	195.57
Qwen3-8B	Base	48.54	96.66	60.95	28.06	205.55	42.67	56.67	44.57	41.68	217.62	24.67	27.89	24.00	60.73	218.11	41.67	65.93	46.86	40.11	213.43
	Oracle	49.50	95.00	61.55	43.39	180.45	40.33	68.00	48.02	60.57	216.72	28.84	41.67	30.81	66.04	226.85	41.62	73.07	49.88	55.41	205.59
Gemma3-27B	Base	43.54	93.33	56.04	50.31	431.02	36.66	53.33	40.26	45.91	470.61	33.33	39.42	32.21	53.13	495.17	38.53	65.13	44.48	48.70	460.77
	Oracle	45.54	91.25	57.13	67.40	265.64	47.00	66.67	51.22	63.95	279.98	45.51	56.41	46.67	68.20	315.87	46.23	73.63	52.54	65.90	281.03
Gemma3-12B	Base	46.04	90.83	57.30	25.62	517.60	39.33	50.00	39.07	41.50	547.40	32.69	39.10	32.46	55.41	578.44	40.57	62.61	44.40	38.25	542.14
	Oracle	44.75	91.67	57.10	61.59	272.97	47.00	52.67	46.80	**66.26**	291.37	45.51	43.91	41.39	**77.66**	385.85	45.95	65.00	49.53	66.55	300.99
Gemma3-4B	Base	30.63	61.67	38.33	25.97	497.97	27.00	33.33	26.27	40.04	473.27	19.87	25.64	20.98	55.96	514.30	27.07	42.06	29.64	37.77	489.04
	Oracle	36.04	60.83	42.33	46.77	280.47	30.00	48.00	33.62	56.96	300.07	31.41	44.87	34.62	61.87	318.96	32.38	52.01	36.88	54.19	296.35

provides only questions, requiring models to maintain requested perspectives using parametric knowledge alone. The **Oracle** setting simulates ideal RAG by providing ground-truth documents supporting the requested perspective, testing whether optimal context helps models maintain focus or whether they still blend in alternative perspectives. We use Gemini-2.5-Pro as the LLM-as-a-Judge. Full setup details are in Appendix G.

3.2 Evaluation Metrics

Following CheckEval (Lee et al., 2025), we adopt an LLM-as-a-Judge evaluation framework that formulates the task as checklist-based verification. We compute two metrics from these judgments.

Perspective Alignment Score (PAS) measures whether a response covers target reasoning paths while excluding distractor paths—a graph-structured generalization of precision/recall. For each question, target paths $T \subseteq \Pi(q)$ denote the complete set of paths matching the requested perspective. Instead of a generic scoring rubric, we use these ground-truth paths from our knowledge graph as a dynamic checklist. The judge iterates through each path $\pi \in \Pi(q)$ and determines whether the response semantically reproduces all of its constituent edges, yielding the set of identified paths:

$$A = \{\pi \in \Pi(q) : \text{Judge}(\pi, \text{response}) = \text{present}\}.$$

Perspective precision $P_{\text{Prec}} = |A \cap T|/|A|$ is the proportion of included paths that are targets; Perspective recall $P_{\text{Rec}} = |A \cap T|/|T|$ is target path coverage. PAS is their harmonic mean.

Faithfulness adopts the standard claim-level verification metric (Es et al., 2024): the judge decomposes each response into discrete claims and verifies each against the ground-truth single-fact sentences from the target paths. Faithfulness is the proportion of supported claims. In Base settings, where no context is provided, these sentences are used only post-hoc, so the metric reflects factual coverage rather than grounding in the provided context.

To validate the PAS judge, human annotators evaluated responses using the same path checklist as the judge. Agreement over 220 path-level Yes/No decisions was 82.7% (Cohen's $\kappa = 0.64$), and the Spearman correlation between human and judge PAS scores at the response level was $\rho = 0.71$. We re-scored PAS on the same responses from three models with a judge from a different model family (GPT-5.6 Terra), using the same prompt and output format; path-level agreement with the Gemini judge was 82.3% ($\kappa = 0.645$), similar to human–judge agreement (Appendix J).

3.3 Results

3.3.1 Models Include but Fail to Exclude

Table 2 reveals a fundamental failure in maintaining focus across all state-of-the-art models. In the Base setting, overall PAS ranges from 29.6 to 46.9. On 1-hop questions, most models achieve high perspective recall but critically low precision. Many models remain close to or below the cover-all precision of 41.4, and none reaches its PAS of 57.6 (Appendix M). In Base, a missed target path may reflect a knowledge gap or a failure to stay in scope;

Oracle provides the target facts. Partial credit substantially reduces the 1-hop to 3-hop recall drop for Oracle responses from three models, while precision remains low (Section 4.4). High recall on 1-hop questions shows that most models can include relevant information, while low precision across hops indicates difficulty in excluding non-requested perspectives. They default to comprehensive "knowledge dumps" rather than focused answers, leaving the filtering to the user (see Appendix A for an extended discussion).

3.3.2 Oracle Improves Grounding but Not Selectivity

As shown in Table 2, the Oracle setting, simulating ideal RAG with ground-truth documents, substantially improves Faithfulness by 46.5% on average, indicating models successfully utilize provided context. With ground truth expanded to all graph sentences, including distractor facts, the Base–Oracle Faithfulness gap is largely preserved for four models (Appendix K). However, perspective control shows limited improvement. For 1-hop reasoning, Oracle provides minimal gains, with most models showing no meaningful precision improvement. At higher hop levels, Oracle shows relatively larger improvements, suggesting RAG helps more when parametric knowledge becomes uncertain. Yet absolute performance remains low, indicating RAG assistance alone cannot solve the focus problem.

3.3.3 Conciseness Does Not Imply Selectivity

Comparing Base and Oracle settings in Table 2, Oracle reduces average tokens by 29.5% while substantially improving Faithfulness. Models become more concise and factually grounded under Oracle context. However, this conciseness does not yield selectivity: PAS shows minimal improvement, indicating that models compress multiple perspectives into shorter outputs rather than excluding alternatives. Conciseness and selective exclusion are orthogonal capabilities—a shorter response can still contain factually valid content from non-target perspectives. Since Oracle represents ideal retrieval, realistic RAG systems with imperfect retrieval will likely face even greater alignment challenges.

4 Analysis

4.1 Prompting Fails at Selective Exclusion

To distinguish whether perspective control failure stems from instruction-following deficiency or inherent limitation, we compare three prompting strategies: **Standard Zero-shot**, **One-Shot Demonstration**, and **Negative Constraints**; full prompt templates are in Appendix H.4.

Table 3: Prompting results. One-Shot and Negative Constraints: mean over questions. Best scores per model are bold; second-best are underlined.

Model / Strategy	P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok
GPT-5					
Oracle	48.75	<u>78.82</u>	<u>56.14</u>	**67.53**	237.13
+ One-Shot	<u>49.18</u>	78.02	55.95	63.46	249.55
+ Negative Constraints	**52.29**	**80.40**	**58.16**	<u>66.43</u>	254.86
Claude-Sonnet-4.5					
Oracle	**47.53**	<u>76.66</u>	**54.16**	<u>63.57</u>	376.09
+ One-Shot	43.05	**77.75**	<u>51.81</u>	**65.06**	346.24
+ Negative Constraints	<u>44.14</u>	73.81	51.48	62.45	399.01
Qwen3-32B					
Oracle	**46.98**	**77.91**	**54.58**	**51.35**	236.44
+ One-Shot	40.66	68.86	47.34	48.47	240.55
+ Negative Constraints	<u>44.96</u>	<u>75.64</u>	<u>52.17</u>	<u>48.93</u>	240.07
Gemma3-27B					
Oracle	**46.23**	**73.63**	**52.54**	**65.90**	281.03
+ One-Shot	42.80	67.77	48.71	60.76	266.55
+ Negative Constraints	<u>45.40</u>	<u>68.86</u>	<u>50.72</u>	<u>62.61</u>	291.59

Table 3 shows that enhanced prompting fails to resolve perspective control. One-Shot demonstrations are ineffective: GPT-5 shows negligible improvement (+0.43%p in P_{Prec}), while other models degrade by 3.4–6.3%p. Examples teach *how to answer* but not *what to exclude*—models mimic the pattern but still blend.

Negative Constraints yield only modest gains. GPT-5 achieves the largest improvement (+3.54%p in P_{Prec}), but precision still plateaus around 50%, and other models show smaller gains or drops. Even directive-compliant models cannot suppress factually valid but off-scope content, indicating that selective exclusion is orthogonal to existing instruction-following capabilities.

4.2 Reasoning Capability is Decoupled from Selective Exclusion

This section examines how reasoning capability interacts with selective exclusion. Our settings progressively reduce how much a model must infer the target perspective: Base gives only a perspective hint, Oracle provides target-perspective facts, and Negative Constraints adds explicit exclusion instructions. Yet Oracle precision plateaus in the low-to-mid 40s on a 10-model average, and GPT-5, a reasoning-capable model in this near-ideal setting, moves only from 48.8 to 52.3 under Negative Constraints (Table 3). This is difficult to explain by a failure to identify the target perspective alone, suggesting that the limitation does not lie in the reach of instructions but in a capability that current alignment training does not address.

To test this interpretation, we compare GPT-5 under Oracle conditions across two reasoning-effort settings (Table 4). Higher reasoning effort slightly degrades P_{Prec} (48.50 $\rightarrow$ 43.68), indicating that scaling reasoning does not improve—and may even slightly reduce—selective exclusion. This reinforces our central claim along the reasoning axis: inclusion and exclusion are decoupled capabilities.

Table 4: GPT-5 reasoning effort (Oracle); mean over questions. Best scores are bold.

Effort	P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok
Low	**48.50**	**78.66**	**55.22**	**72.98**	274.21
High	43.68	75.09	51.49	71.65	261.91

4.3 Impact of Distractor Documents

Oracle settings revealed that models struggle with perspective control even under ideal conditions. To examine whether this problem intensifies when retrieval introduces noise, we evaluate the Noisy setting. This setting provides documents from both target paths ($\Pi_{\mathcal{P}}^{\text{target}}$) and distractor paths ($\Pi_{\mathcal{P}}^{\text{distractor}}$) in equal numbers, simulating realistic retrieval where semantically related but perspective-misaligned content appears alongside correct information.

Table 5: Noisy results: mean over questions. Best scores per model are bold; second-best are underlined.

Model / Setting	P_{Prec}	P_{Rec}	PAS	Faithfulness	#Tok
GPT-5					
Base	39.58	68.63	46.26	34.96	486.99
Noisy	45.88	**79.67**	54.63	67.09	345.58
Oracle	**48.75**	78.82	**56.14**	**67.53**	237.13
Claude-Sonnet-4.5					
Base	38.08	63.28	43.84	43.46	413.76
Noisy	38.74	69.96	46.51	**69.65**	397.90
Oracle	**47.53**	**76.66**	**54.16**	63.57	376.09
Qwen3-32B					
Base	39.91	68.02	46.89	40.34	275.75
Noisy	42.86	76.92	52.12	**53.36**	256.97
Oracle	**46.98**	**77.91**	**54.58**	51.35	236.44
Gemma3-27B					
Base	38.53	65.13	44.48	48.70	460.77
Noisy	44.03	73.17	51.04	63.64	294.59
Oracle	**46.23**	**73.63**	**52.54**	**65.90**	281.03

Table 5 confirms that distractor documents systematically degrade perspective control. Traditional RAG metrics fail to capture this degradation. Answer Relevancy from (Es et al., 2024) (Table 6), which measures semantic similarity between questions and answers, shows inconsistent patterns—scores fluctuate erratically across Base, Noisy, and Oracle settings regardless of distractor presence. This failure stems from the fundamental limitation exposed in Section 2.4: target and distractor paths exhibit nearly identical semantic similarity to queries, making similarity-based metrics unable to distinguish perspective-aligned content from off-scope alternatives.

Table 6: Answer Relevancy. Best scores per model are bold; second-best are underlined.

Model Name	Answer Relevancy Settings		
	Base	Noisy	Oracle
Claude-Sonnet-4.5	0.7309	0.7297	**0.7650**
GPT-5	**0.6568**	0.6363	0.6498
Gemma3-27B-IT	0.6067	0.6443	**0.6981**
Qwen3-32B	0.7287	**0.7294**	0.7282

In contrast, PAS demonstrates monotonic ordering: Noisy settings consistently produce intermediate performance between Base and Oracle across all models. This validates a critical property of our metric: it properly quantifies perspective contamination levels rather than conflating topical relevance with constraint adherence. When documents from $\Pi_{\mathcal{P}}^{\text{distractor}}$ are introduced, PAS degrades proportionally, establishing measurable sensitivity to the exact problem we aim to capture.

The degradation from Oracle to Noisy reveals varying model robustness. GPT-5 maintains relatively stable precision, while Claude-Sonnet-4.5 shows substantial vulnerability to distractor influence. Response length patterns provide additional insight: token generation decreases progressively from Base to Noisy to Oracle. When models lack clear guidance from context, they produce longer responses—a behavior reinforced through alignment training where length correlates with perceived helpfulness (Guo et al., 2025). As retrieval quality improves and uncertainty decreases, responses become more concise and focused.

Table 7: Retrieval precision/recall with Noisy's document count. Best scores per model are bold; second-best are underlined.

Setting	Claude-Sonnet-4.5	Qwen3-32B	Gemma3-27B
Base	38.08 / 63.28	39.91 / 68.02	38.53 / 65.13
RAG BM25	31.0 / 59.3	39.4 / 70.0	33.7 / 64.7
RAG dense	34.4 / 63.7	34.1 / 65.6	36.5 / 60.8
Noisy	38.74 / 69.96	42.86 / 76.92	44.03 / 73.17
Oracle	**47.53 / 76.66**	**46.98 / 77.91**	**46.23 / 73.63**

To examine whether the exclusion failure persists under real retrieval, we attached BM25 and the dense retriever BGE-large-en-v1.5 to the textbook corpus and gave each question the same num-

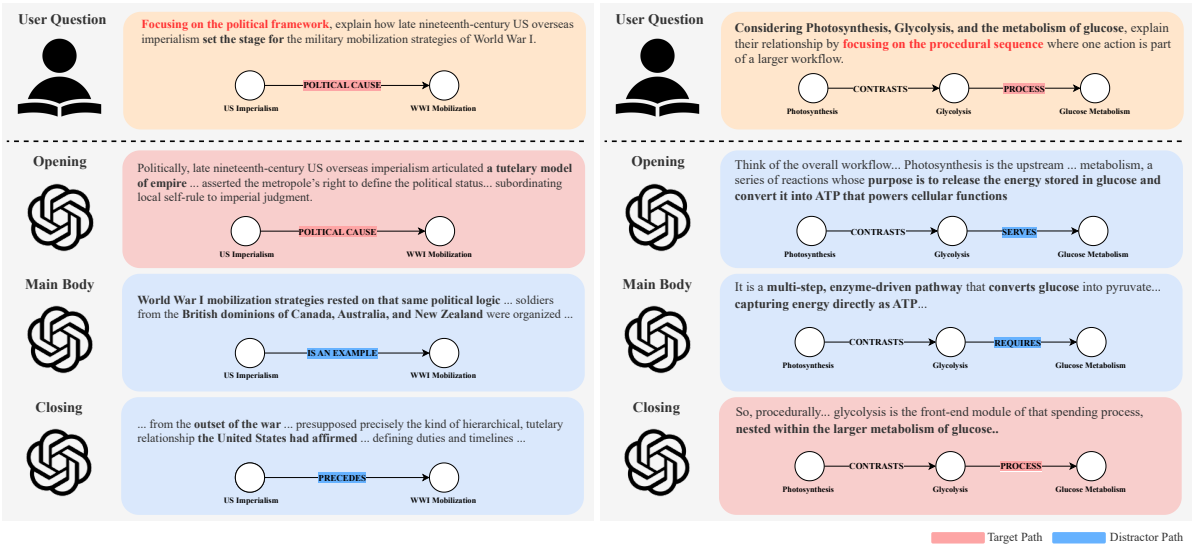


Figure 4: **Qualitative analysis of failures in selective exclusion.** The History case (left) adds unrequested entities, while the Biology case (right) adds details beyond the requested sequence. Although **factually accurate**, both responses violate **explicit constraints** by merging **multiple valid perspectives**.

ber of retrieved documents as Noisy, across three models (Table 7). Only 43.5% (BM25) and 47.4% (dense) of retrieved passages belonged to the requested perspective, with distractor perspectives at 32.2% and 36.5%. Precision remains below Noisy and Oracle for all three models, so real retrieval does not resolve the exclusion failure observed in the controlled settings.

4.4 Explanations for Low Precision

We examine whether low precision reflects all-or-nothing path scoring or item ambiguity. To distinguish perspective control failure from the difficulty of reproducing longer paths, we re-scored Oracle responses from GPT-5, Qwen3-32B, and Gemma3-27B with per-edge partial credit. Each path π receives credit $c(\pi)$ equal to the fraction of its edges the response reproduces, with the same Gemini judge and partial credit applied equally to precision and recall. Precision and recall replace $|A \cap T|$ with $\sum_{\pi \in T} c(\pi)$ and $|A|$ with $\sum_{\pi \in \Pi(q)} c(\pi)$.

Table 8: Oracle recall drop and precision change, 1-hop to 3-hop (same Gemini judge).

Model	Recall drop all-or-nothing / partial	Precision change all-or-nothing / partial
GPT-5	-36.6 / -12.8	-0.7 / -6.0
Qwen3-32B	-31.7 / -12.1	$+1.2$ / $+4.5$
Gemma3-27B	-34.8 / -15.8	± 0.0 / $+3.9$
Average	-34.4 / -13.6	$\approx$ flat

Table 8 shows that the average 1-hop to 3-hop recall drop shrinks from 34.4 to 13.6 points under partial credit, so most of the per-hop recall loss reflects the difficulty of reproducing longer chains. Precision, by contrast, stays low under both scorings (Oracle 43–54) with no consistent trend across hops. Since the same relaxation is applied to both metrics, low precision appears unlikely to be an artifact of all-or-nothing path scoring.

To examine the effect of item ambiguity, we re-aggregated scores on items where all three annotators agreed on distractor validity (Appendix L). Precision does not rise noticeably on these items, suggesting that ambiguous items alone do not explain the low precision.

4.5 Qualitative Analysis

To illustrate how high Faithfulness can coexist with perspective misalignment, we examine 25 responses that score highly on Faithfulness—a widely used metric for whether claims are grounded in context—yet receive low PAS scores, each reviewed independently by two authors. Figure 4 presents two representative cases where factual accuracy and perspective adherence diverge. The red-colored parts show target paths $\Pi_{\mathcal{P}}^{\text{target}}$, while the blue-colored parts show distractor paths $\Pi_{\mathcal{P}}^{\text{distractor}}$.

In the History domain (left), the question requests a political framework connecting US imperialism to WWI mobilization. Despite high Faithfulness, the response incorporates distractor paths by introducing British dominions as exemplification and adding temporal precedence—beyond the

requested political lens. In the Biology domain (right), the question asks for procedural sequence only. The response is accurate but adds functional purposes and mechanistic details from the "Serves" and "Requires" paths rather than maintaining focus on the requested "Process" relationship.

The key pattern is that responses include the target perspective but also incorporate content from adjacent perspectives. This is consistent with the low precision across hops under partial-credit scoring (Section 4.4). These cases show that high Faithfulness does not ensure perspective alignment. Models incorporate semantically related content from $\Pi_{\mathcal{P}}^{\text{distractor}}$, producing comprehensive yet unfocused responses that users must filter.

5 Related Works

RAG and Instruction-Following Evaluation. RAG evaluation has matured around metrics for fluency, factual accuracy, and answer quality (Yu et al., 2025a; Es et al., 2024). In parallel, instruction-following benchmarks have progressively widened the constraint space. IFEval (Zhou et al., 2023) introduced 25 verifiable surface-level constraints such as length, format, and keyword inclusion. FollowBench (Jiang et al., 2024) extended evaluation across five axes—content, situation, style, format, and example—at multiple difficulty levels, and InfoBench (Qin et al., 2024) measured fine-grained compliance through the Decomposed Requirements Following Ratio. ComplexBench (Wen et al., 2024) brought semantic-level constraints into the taxonomy. Across both lines, evaluation predominantly measures inclusion of relevant or constraint-satisfying content. Exclusion-oriented evaluation has been explored in adjacent settings. CoDI-Eval (Chen et al., 2024) introduced an exclusion dimension via Toxicity Avoidance, though limited to safety-level surface cases. DeCRIM (Ferraz et al., 2024) evaluates explicit negative constraints such as "no hashtag," while the exclusion constraint in EIFBENCH (Zou et al., 2025) forbids specific mentions. These benchmarks do not directly measure exclusion of factually valid reasoning outside the requested perspective.

Personalization and Persona Evaluation. RoleLLM (Wang et al., 2024) evaluates role-level consistency through RoleBench, while PersonaGym (Samuel et al., 2025) measures persona adherence. LaMP (Salemi et al., 2024), LaMP-QA (Salemi and Zamani, 2025), and PrefEval (Zhao et al., 2025) evaluate whether models can be steered toward a user's profile or preference. On opinion alignment, Santurkar et al. (2023) find that steering contributes to but does not fully achieve alignment, and Durmus et al. (2024) report that persona prompting shifts viewpoints but reinforces stereotypes. Across these axes, evaluation focuses on steering model behavior toward a target, without directly measuring exclusion of factually valid content outside the request.

Perspective Recognition and Selective Exclusion. PERSPECTRA (Nie et al., 2026) identifies pro and con viewpoints in a discussion: its perspectives are subjective stances, and exclusion is not involved. PerspectiveBench instead evaluates whether responses follow the requested perspective while excluding other factually valid reasoning. Our distractors are factual and as close to the query in topic as the targets (Figure 3), making them hard negatives that simple relevance scoring cannot filter. Semantic similarity metrics conflate topical relevance with constraint adherence (Li et al., 2024b; Goel et al., 2025), masking cases where responses blend multiple perspectives that all pass factual verification. Appendix I provides the full comparison with per-benchmark descriptions.

6 Conclusion

We identified perspective misalignment—a gap where models maintain high factual accuracy yet fail to suppress content outside user-specified perspectives. PerspectiveBench reveals that state-of-the-art models achieve high recall on 1-hop questions but critically low precision—even under Oracle conditions with perfect retrieval—supporting separate evaluation of inclusion and exclusion.

Beyond benchmarks, this challenge has direct stakes in domains where perspective-specific framing matters—education requiring a single causal account, legal analysis bounded by doctrines, and interpretation of experimental results within a given theoretical framework. In such settings, factually accurate but off-scope content can make responses difficult to use for the requested task. By publicly releasing **PerspectiveBench** and its construction pipeline, we provide a foundation for evaluating perspective alignment and constructing benchmarks; extension to professional settings remains to be validated.

Limitations

Domain coverage. PerspectiveBench is constructed from OpenStax textbooks in Physics, Biology, and History. These textbook questions do not span open-domain or user-facing tasks, and whether our findings hold in professional settings such as legal drafting or interpretation of experimental results has not been tested.

Clustering sensitivity. We have not conducted a full clustering sensitivity analysis under alternative UMAP, HDBSCAN, and top-K settings. Three annotators checked target alignment and distractor validity on the validation sample (Appendix F). This does not establish that the target/distractor split remains stable under different settings.

Judge dependence. Both PAS and Faithfulness rely on Gemini-2.5-Pro. Human validation covers PAS path judgments, not Faithfulness, for which we have no separate human validation. We re-scored PAS with GPT-5.6 Terra (Section 3.2), but possible bias from benchmark construction with Gemini and bias shared across model families remain.

Scoring granularity. PAS evaluates each path as an atomic unit: reproducing only part of a target path receives no credit, while partial distractor paths are not identified as violations. Per-edge partial-credit results are reported in Section 4.4. For Faithfulness, the main evaluation uses target-path documents, so factually valid distractor claims may be marked unsupported; Appendix K reports an expanded-ground-truth check.

Human performance. We have not measured PAS on answers written by experts. The cover-all and unanimous-item analyses do not establish a human performance level. Evaluating expert answers remains future work.

Retrieval realism. Noisy injects distractor documents in a controlled 1:1 ratio with targets and does not reproduce a deployed retrieval pipeline. Our BM25 and dense retrieval experiments cover three models on the textbook corpus (Table 7); interactions between retrieval quality and perspective control in other deployed settings remain to be evaluated.

No user study. We have no direct user study measuring cognitive cost for this task. The literature cited in Appendix A provides background motivation.

Ethical Considerations

Data source and licensing. The OpenStax source textbooks are released under a Creative Commons Attribution license that permits redistribution and adaptation with attribution. We will release the benchmark and construction pipeline under the same permissive license.

Human validation. Three voluntary annotators recruited from the authors' home institution performed both the 180-item inter-rater study (Appendix F) and the PAS judge-validation study (Section 3.2). All three hold graduate-level training in NLP and machine learning and participated as part of an internal academic collaboration without compensation. Annotators were informed of the study purpose and consented to the release of aggregate inter-rater statistics; no per-annotator labels or personally identifying information are released.

Use of LLMs. We used Gemini-2.5-Pro for benchmark construction and evaluation of PAS and Faithfulness. We also re-scored PAS on the same responses with GPT-5.6 Terra. Prompt templates are provided in Appendix H. LLM assistance was also used for grammar polishing and revisions to the wording and structure of the paper; all technical claims, experimental design, and analyses were authored and verified by the human authors.

Intended use and dual-use risk. PerspectiveBench is intended for academic evaluation, not production decisions on its own. Selective exclusion could be misused to enforce one-sided narratives or suppress viewpoints. We report precision and recall alongside PAS, and Faithfulness to assess support for response claims. Exclusion should follow the user's requested scope; when that scope is not constrained, users may want broader coverage.

Acknowledgments

This work was supported by the Technological Innovation R&D Program (SCALEUP TIPS, Tech Investor Program for Scaleup; No. RS-2025-25465308) funded by the Ministry of SMEs and Startups (MSS, Korea). This research was supported by Global - Learning & Academic research institution for Master's·PhD students, and Postdocs (G-LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25442252).

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-rag: A study of points of failure in retrieval augmented generation. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 607–611.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429.

Yihan Chen, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024. Benchmarking large language models on controllable generation under diversified instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17808–17816.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Samarth Goel, Reagan J. Lee, and Kannan Ramchandran. 2025. SAGE: A realistic benchmark for semantic understanding. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Hyeonji Kim, Sujeong Oh, and Sanghack Lee. 2026. Mitigating length bias in RLHF through a causal lens. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17517–17525.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *Preprint*, arXiv:2506.08872.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China. Association for Computational Linguistics.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024a. A survey of generative search and recommendation in the era of large language models. *CoRR*, abs/2404.16924.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding

benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Shangrui Nie, Kian Omoomi, Lucie Flek, Zhixue Zhao, and Charles Welch. 2026. PerSpectra: A scalable and configurable pluralist benchmark of perspectives from arguments. In *International Conference on Learning Representations*, volume 2026, pages 76593–76607.

OpenStax. OpenStax: Open educational resources. Accessed: September 22, 2026.

Max Owens, Ernst H. W. Koster, and Nazanin Derakshan. 2013. Improving attention control in dysphoria through cognitive training: Transfer effects on working memory capacity and filtering efficiency. *Psychophysiology*, 50(3):297–307.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A benchmark for personalized long-form question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1159, Suzhou, China. Association for Computational Linguistics.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Leila Shahrzadi, Ali Mansouri, Mousa Alavi, and Ahmad Shabani. 2024. Causes, consequences, and strategies to deal with information overload: A scoping review. *International Journal of Information Management Data Insights*, 4(2):100261.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268. Curran Associates, Inc.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition. In *Advances in Neural Information Processing Systems*, volume 37, pages 137610–137645. Curran Associates, Inc.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025a. Evaluation of retrieval-augmented generation: A survey. In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.

Jiayang Yu, Yuxi Xie, Guixuan Zhang, Jie Liu, Zhi Zeng, Ying Huang, and Shuwu Zhang. 2025b. Mindrag: Multimodal context-aware and intent-aware retrieval-augmented generation for educational publications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4216–4223.

Liese Zahabi. 2013. Exploring information-triage: Speculative interface tools to help college students conduct online research. In *Design, User Experience, and Usability. Web, Mobile, and Product Design*, pages 611–620, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yingyi Zhang, Pengyue Jia, Derong Xu, Yi Wen, Xianneng Li, Yichao Wang, Wenlin Zhang, Xiaopeng Li, Weinan Gan, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2026. Personalize before retrieve: LLM-based personalized query expansion for user-centric retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16406–16414.

Ziyin Zhang, Zihan Liao, Hang Yu, Peng Di, and Rui Wang. 2025. F2llm technical report: Matching sota embedding performance with 6 million open-source data. *Preprint*, arXiv:2510.02294.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations*, volume 2025, pages 15888–15931.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. RAGEval: Scenario specific RAG evaluation dataset generation framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.

Tao Zou, Xinghua Zhang, Haiyang Yu, Minzheng Wang, Fei Huang, and Yongbin Li. 2025. EIFBENCH: Extremely complex instruction following benchmark for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20930–20953, Suzhou, China. Association for Computational Linguistics.

A Why Perspective Constraints Matter

A.1 Constraints from the Task

A perspective constraint is not a matter of stylistic preference. It comes from what the person receiving the answer has to do with it. Three settings make this concrete.

- **Education:** A student in an economic history course analyzes the fall of Rome through an economic lens. What the student has to do is explain, as a single causal account, how economic factors led to the collapse. If the model supplies the military and political causes as well, the student must separate the economic explanation from the other causes to complete the task.
- **Law:** An attorney asked to provide an analysis under contract doctrine alone is given tort arguments alongside the contract ones. Even if both are legally sound, a draft that mixes the two cannot be used as written for the requested task.
- **Science:** A researcher asked to interpret an experimental result thermodynamically is given a dynamical account as well. Even if both explanations are correct, mixing them does not show whether thermodynamic premises alone account for the result.

All three call for an account whose perspective is constrained out of necessity. One might argue that the user can simply pick out the parts they need, but this shifts the burden back onto the user. In multi-hop explanations especially, the perspectives are interleaved sentence by sentence, and extracting only the information belonging to the requested perspective is itself difficult.

A.2 Cognitive Load as Background

Contemporary LLMs have a strong tendency toward lengthy, comprehensive responses. Through instruction-following training, particularly RLHF, models learn that response length and breadth correlate with perceived helpfulness (Guo et al., 2025; Singhal et al., 2024), leading to exhaustive coverage patterns (Kim et al., 2026) even when users specify narrow constraints.

This verbosity can impose costs on users across multiple dimensions. Cognitively, research on information processing demonstrates that excessive information—even when factually correct—increases cognitive load and impairs decision-making efficiency (Kosmyna et al., 2025; Shahrzadi et al., 2024). Users must perform active cognitive filtering to extract relevant information from verbose responses, a task that consumes working memory resources and increases mental effort (Owens et al., 2013). This process, which cognitive science terms "information triage" (Zahabi, 2013), requires determining which information to retain and which to ignore. Unlike hallucinations, which signal clear quality failures, perspective violations present factually grounded content that appears relevant, making misalignment difficult to detect without explicit attention.

Beyond cognitive burden, longer responses require more output tokens and can increase API costs and latency.

These challenges motivate growing recognition that selective information delivery—not just comprehensive coverage—represents a critical capability for practical systems. Recent work emphasizes designing retrieval systems that align with focused user intent (Agrawal et al., 2024; Yu et al., 2025b; Zhang et al., 2026). However, evaluating such selective control requires distinguishing valid target content from valid but off-scope alternatives, a measurement challenge our benchmark directly

addresses through explicit perspective partitioning and separate assessment of inclusion versus exclusion capabilities. This literature provides background motivation; we do not measure cognitive cost directly for this task.

B Domain Selection Rationale

We selected three domains where single questions naturally admit multiple valid perspectives:

- **Physics**: Physical phenomena can be interpreted through different theoretical frameworks. For instance, energy conservation can be analyzed from mechanical or thermodynamic perspectives, where both provide valid yet distinct explanations.
- **Biology**: Biological processes support explanations at different levels of analysis. Cellular respiration can be described through biochemical mechanisms or metabolic regulation roles, each offering a different explanatory focus.
- **History**: Historical events arise from multidimensional causes that can be examined separately. Rome's fall can be explained through economic, military, or political factors, where each perspective provides a coherent causal narrative.

C Edge Score Weights

The relevance score $\text{Score}(v_i, v_j) = S_{\text{ent}} + S_{\text{prox}} + S_{\text{sem}}$ in Section 2.3.3 uses the following point assignments:

Shared Entity (S_{ent}). Let $E(v_i)$ denote the set of entities extracted from facts in cluster v_i (from Section 2.3.1). We assign 5 points when $|E(v_i) \cap E(v_j)| \geq 3$ and 4 points when $1 \leq |E(v_i) \cap E(v_j)| \leq 2$.

Textual Proximity (S_{prox}). Each fact f is indexed by its sentence position in the source text. Facts $f_i \in C_i$ and $f_j \in C_j$ co-occur if $|pos(f_i) - pos(f_j)| \leq 5$. Let k be the total co-occurring pair count; we use binary search over sorted position lists to count these pairs in $O(|C_i| \cdot \log |C_j|)$ time. We assign 3 points when $k \geq 5$, 2 points when $2 \leq k \leq 4$, and 1 point when $k = 1$.

Probabilistic Cluster Overlap (S_{sem}). When facts assigned to one cluster could plausibly belong to another (soft HDBSCAN probabilities), this indicates semantic proximity. We assign 1 point when each cluster contains at least 3 facts with soft HDBSCAN probabilities greater than 0.05 for the other cluster.

D Multi-Perspective QA Generation: Detailed Steps

Step 1: Reasoning Path Construction. A reasoning path π represents a sequence of nodes connected by edges in graph G. Formally, $\pi = (v_0 \to v_1 \to \cdots \to v_k)$ where $v_0 = v_s$ is the start node, $v_k = v_e$ is the end node, and each arrow represents an edge $(v_i, v_{i+1}) \in \mathcal{E}$ with an associated relationship label. Paths range from 1-hop (direct connection: $v_s \to v_e$) to 3-hop (multi-step: $v_s \to v_1 \to v_2 \to v_e$).

Step 2: Perspective Definition. For nodes v_s and v_e, the graph yields multiple valid paths: $\Pi(v_s \to v_e) = \{\pi_1, \pi_2, \ldots, \pi_n\}$. A perspective $\mathcal{P}_i \subseteq \Pi(v_s \to v_e)$ represents paths sharing the same explanatory focus—the type of relationship (e.g., causal, comparative, structural) connecting nodes. In Figure 2, from "Military Decay" to "Fall of Western Rome," three distinct relationship types exist: "External Invasion," "Political Collapse," and "Economic Collapse." Paths traversing "Political Collapse" or "Economic Collapse" share an internal explanatory focus, forming perspective $\mathcal{P}_{\text{internal}}$.

Step 3: Perspective-Targeted Question Generation. For each perspective $\mathcal{P}_i$, the prompt requires an LLM to generate question q_i using conceptual hints to guide models toward target paths, without using edge labels or their direct synonyms. For each generated question, we compile ground-truth documents by collecting single-fact sentences from clusters along target paths $\Pi_{\mathcal{P}}^{\text{target}}$. These documents are provided as context to models in evaluation settings that simulate ideal retrieval conditions.

E Distribution of Perspective Types

To characterize the perspective types present in our benchmark, we grouped specific perspectives into semantically coherent categories using LLMs. Figure 5 shows their distributions across domains.

Physics predominantly features temporal and causal perspectives. As shown in Table 9, "Sequential and Procedural" questions target derivation order—which principle must precede another—while "Causal and Explanatory" questions focus on mechanistic frameworks underlying phenomena. The same physical concepts, such as Newton's laws and

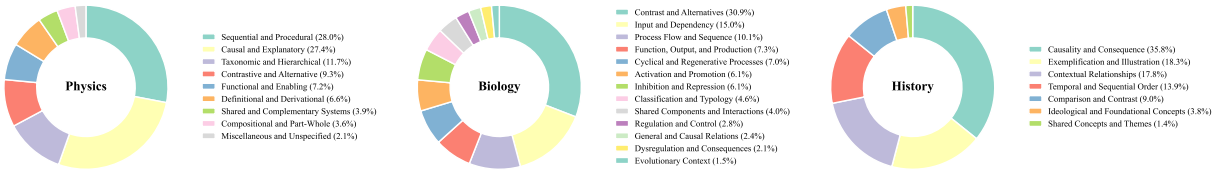


Figure 5: **Distribution of semantic perspective types across domains.** Fine-grained relationships are aggregated into high-level categories. The distributions show the range of perspective types present in each domain.

momentum conservation, can be explained through either temporal precedence (which derives from which) or mechanistic causality (how one principle explains another), demonstrating that even structured domains support multiple valid reasoning paths.

Biology exhibits contrastive and dependency patterns. Questions probe functional differentiation ("Contrast and Alternatives"), prerequisite conditions ("Input and Dependency"), and workflow integration ("Process Flow and Sequence"). When examining processes like transcription versus replication, models must determine whether the question targets functional objectives, resource dependencies, or procedural sequences—each constituting a valid but distinct explanatory approach to the same biological phenomena.

History includes causal and contextual perspectives, as well as exemplification. "Causality and Consequence" questions span political, economic, and social dimensions. "Exemplification and Illustration" questions connect abstract trends to concrete events, while "Contextual Relationships" questions require synthesizing background frameworks. In Table 9, the history questions require political causes or concrete examples of broader trends, showing how different relationships yield different explanations.

F Benchmark Construction Validity

Our pipeline combines deterministic algorithms with LLM generation, so the deterministic component alone cannot certify the validity of all outputs. We therefore conduct a human audit on a stratified validation sample of 180 QA items (three domains × three hop levels × 20 items per stratum). This size yields a margin of error of ± 7.3%p at the 95% confidence level for binary proportion estimates, with equal allocation across strata so that overall estimates are computed directly over the full sample without post-hoc weighting. This audit checks the final target/distractor split on the validation sample, rather than treating clustering output alone as evidence of validity.

Three annotators with graduate-level training in the relevant domains independently labeled all 180 items, following a 20-item pilot calibration round in which agreement of ≥ 0.70 was required before proceeding. Each item is judged on two binary criteria: (i) *target alignment*—whether the question and its target paths $\Pi_{\mathcal{P}}^{\text{target}}$ are consistent with the specified perspective—and (ii) *distractor validity*—whether each distractor path is factually valid yet distinguishable from the target. Final gold labels are assigned by majority vote across the three raters, with no inter-rater discussion to preserve full independence.

To account for the well-known deflation of Cohen/Fleiss' κ under skewed prevalence (Feinstein and Cicchetti, 1990), we report three complementary statistics: Fleiss' κ (Fleiss, 1971) as the chance-corrected baseline, the prevalence-adjusted bias-adjusted κ (PABAK) (Byrt et al., 1993) as a prevalence-robust counterpart, and raw Percent Agreement as a direct observation. All metrics are computed over the full 180-item sample.

Target alignment achieves Fleiss' $\kappa = 0.534$ (moderate (Landis and Koch, 1977)). Distractor validity achieves **PABAK = 0.638 (substantial)** together with high raw agreement (**81.9% pairwise, 72.8% all-rater**); the lower Fleiss' κ of 0.421 is consistent with κ deflation under high positive prevalence (80.6%) (Feinstein and Cicchetti, 1990). Item-level error analysis on the disagreement subset further attributes residual disagreement primarily to borderline cases where a distractor path is topically adjacent to the target perspective—precisely the hard-negative regime our benchmark is designed to expose (Section 2.4). The remaining disagreement may therefore reflect both benchmark difficulty and annotation noise; restricting to the items where all three annotators agreed on distractor validity leaves precision essentially unchanged (Appendix L).

Table 9: Representative questions from the benchmark dataset, categorized by domain and semantic perspectives.

Domain	Category	Representative Question
Physics	**Sequential and Procedural**	Describe the logical sequence connecting Newton's second law of motion and the conservation of momentum, focusing on which concept must be established first to derive the other.
	Causal and Explanatory	Focusing on the underlying mechanism, describe how the vector summation of forces provides a framework for understanding an object's state of constant velocity.
	Taxonomic and Hierarchical	Explain the connection between mechanical energy, conservation of momentum, Newton's second law, and Newton's third law, focusing on the hierarchical relationship.
Biology	**Contrast and Alternatives**	Focusing on their fundamental objectives and scope, explain what makes the process of gene transcription distinct from DNA replication.
	Input and Dependency	Considering chemiosmosis, the evolution of aerobic respiration, and fermentation, explain their relationship by focusing on the essential resources one process provides for another to function.
	Process Flow and Sequence	Considering Photosynthesis, Glycolysis, and the metabolism of glucose, explain their relationship by focusing on the procedural sequence where one action is part of a larger workflow.
History	**Causality and Consequence**	Focusing on the political chain of events, explain the connection between the Balkan crisis, German foreign policy, and the rise of socialism.
	Exemplification and Illustration	Focusing on how the Balkan crisis serves as a specific illustration of broader trends, explain the connection between it, German foreign policy, and the rise of socialism.
	Contextual Relationships	Describe the broader historical framework established by British mercantilist policies within which the mobilization of armies during World War I took place.

G Experimental Setup Details

Benchmark Construction. We used Gemini-2.5-Pro during benchmark construction (Section 2.3), including atomic fact extraction, cluster summarization, edge generation, and QA generation.

Decoding Parameters. For all evaluated models, temperature is set to 0.0 in the generation settings.

Evaluated Models. We evaluate two groups of frontier LLMs. The first group consists of state-of-the-art proprietary models: GPT-5, GPT-4.1, Claude-Sonnet-4.5, and Claude-Sonnet-4. The second group evaluates performance across model scales: the Qwen3 series (32B, 14B, 8B) and the Gemma3 series (27B, 12B, 4B).

LLM-as-a-Judge. For checklist-based evaluation of PAS and Faithfulness, we employ Gemini-2.5-Pro as the judge model, selected for its strong reasoning capabilities. We also re-scored PAS on the same responses from Claude-Sonnet-4.5, Qwen3-32B, and Gemma3-27B with GPT-5.6 Terra, with the scoring prompt and output format unchanged (Appendix J).

Oracle Setting Details. In the Oracle setting, at least two documents are provided per question, collectively containing all necessary information for the target perspective. Documents are the single-fact sentences drawn from clusters along the target reasoning paths $\Pi_{\mathcal{P}}^{\text{target}}$ (Section 2.3.4).

Real Retrieval Setting. We attached BM25 (Okapi, $k_1 = 1.5$, $b = 0.75$) and the dense retriever BGE-large-en-v1.5 (`BAAI/bge-large-en-v1.5`) to the OpenStax chunks described in Section 2.2. For dense retrieval, we add `Represent this sentence for searching relevant passages:` before each question. The number of retrieved documents is $2\times$ the number of target facts, matching Noisy. Answer generation uses the same system prompt and QUESTION / CONTEXT format as Oracle and Noisy; scoring follows the main evaluation protocol.

Reproducibility. To support transparency and enable community research, we publicly release our benchmark construction pipeline, datasets, and all prompt templates used in experiments on GitHub.[1]

H Prompt Templates

We release all prompt templates used by PerspectiveBench together with our code, organized along the same pipeline as the main text: benchmark construction (Section 2.3), answer generation under Base and Oracle settings (Appendix G), LLM-as-a-Judge evaluation (Section 3.2), and the prompting-strategy variants compared in our analysis (Section 4.1).

[1] URL anonymized for review.

H.1 Benchmark Construction Prompts

Figures 6–9 correspond, one-to-one, to the four LLM-driven construction stages and trace how raw text chunks T are progressively lifted into the knowledge graph $G = (\mathcal{V}, \mathcal{E})$ and finally into perspective-controlled QA items: (i) extraction of single-fact sentences with their entities (Section 2.3.1), (ii) representative summary generation that turns clusters into semantic nodes $\mathcal{V}$ (Section 2.3.2), (iii) edge generation that enumerates directed explanatory relations between nodes (Section 2.3.3), and (iv) perspective-targeted question generation over reasoning paths $\Pi(v_s \to v_e)$ (Section 2.3.4).

H.2 Answer Generation Prompt

The same answer-generation prompt (Figure 10) is used for every evaluated model under both the Base and Oracle settings (Appendix G). The context is empty in Base and contains target-path documents in Oracle.

H.3 LLM-as-a-Judge Prompts

Figures 11 and 12 implement the PAS and Faithfulness metrics defined in Section 3.2.

H.4 Prompting Strategy Variants

For the analysis in Section 4.1, we compare three prompting strategies layered on top of the answer-generation prompt of Figure 10:

- **Standard Zero-shot:** basic instruction to "focus on the requested perspective."
- **One-Shot Demonstration:** adds one exemplary response that perfectly adheres to perspective constraints, testing whether models can mimic the pattern through in-context learning.
- **Negative Constraints:** adds explicit prohibitions to "exclude information and entities from other perspectives," directly targeting comprehensive-response tendencies and the hard-negative distractor paths in $\Pi_{\mathcal{P}}^{\text{distractor}}$.

I Comparison with Existing Benchmarks

Table 10 shows a comparison of existing benchmarks and PerspectiveBench in terms of evaluation focus, scoring, and exclusion targets. The distinction is the nature of the excluded content: our distractors are factual and as close to the query in topic as the targets (Figure 3).

Instruction Following. IFEval evaluates verifiable format and keyword constraints, including keyword exclusion. FollowBench evaluates constraints across content, situation, style, format, and example. InfoBench measures compliance through decomposed atomic requirements. ComplexBench evaluates composed multi-constraint following, including semantic constraints. EIFBENCH includes an exclusion constraint that forbids specific mentions rather than constraining an entire reasoning scope. RealInstruct, introduced with DeCRIM, evaluates real user instructions with explicit negative constraints such as "no hashtag." CoDI-Eval evaluates keyword constraints and toxicity avoidance, using rules and models rather than a checklist of factual reasoning paths.

Personalization and Persona. LaMP evaluates profile-conditioned answers, while LaMP-QA evaluates personalized long-form QA. PrefEval measures adherence to stated user preferences. RoleLLM evaluates persona speech and knowledge through RoleBench, and PersonaGym measures persona-agent behavioral consistency. These benchmarks concern steering toward a profile, preference, or persona.

Perspective Recognition and Exclusion. PERSPECTRA identifies pro and con viewpoints in discourse; these perspectives are subjective stances. PerspectiveBench instead measures whether models reproduce the requested reasoning paths while excluding other factually valid paths. Topical relevance alone cannot distinguish these target and distractor paths.

J Judge Robustness

We re-scored PAS on the same responses from Claude-Sonnet-4.5, Qwen3-32B, and Gemma3-27B in Base and Oracle with a judge from a different model family, GPT-5.6 Terra. We kept the scoring prompt and output format unchanged and replaced only the judge model. Each decision indicates whether the response reproduces a reasoning path.

Agreement between the judges is close to human agreement with Gemini (Table 11). Recall remains higher than precision, and Oracle PAS remains higher than Base for every model under both judges. Model order can change across judges, and residual judge bias may remain.

Table 10: Comparison with existing benchmarks. N/A indicates that no separate exclusion target is specified in this comparison.

Benchmark	What it measures	Violation detected by	Nature of excluded content
IFEval (Zhou et al., 2023)	Verifiable format and keyword constraints	Rules (code)	Keyword exclusion
FollowBench (Jiang et al., 2024)	Multi-level constraint following	Rules + LLM	Mostly N/A
InfoBench (Qin et al., 2024)	Decomposed atomic requirements	LLM judge	N/A
ComplexBench (Wen et al., 2024)	Composed multi-constraint following	Rules + LLM	Partial (semantic)
EIFBENCH (Zou et al., 2025)	Complex multi-task following with exclusion	LLM judge	Specific mentions
RealInstruct (DeCRIM) (Ferraz et al., 2024)	Real user instructions with multiple constraints	Human + LLM	Explicit negative constraints
CoDI-Eval (Chen et al., 2024)	Instruction constraints and toxicity avoidance	Rules + models; Perspective API for toxicity	Keywords; toxicity
LaMP (Salemi et al., 2024)	Profile-conditioned answers	Automatic vs. reference	No exclusion notion
LaMP-QA (Salemi and Zamani, 2025)	Personalized long-form QA	LLM judge	No exclusion notion
PrefEval (Zhao et al., 2025)	Adherence to stated user preferences	LLM judge	A preference, not fact
RoleLLM (Wang et al., 2024)	Persona speech and knowledge	Similarity + LLM	No exclusion notion
PersonaGym (Samuel et al., 2025)	Persona-agent behavioral consistency	LLM judge	No exclusion notion
PERSPECTRA (Nie et al., 2026)	Identifying pro/con viewpoints	Gold labels from debate graphs	Subjective stances; no exclusion notion
PerspectiveBench	Answering within the requested perspective only	KG path checklist + LLM judge	Factual and topically as close as the target

Table 11: Agreement at the path-level Yes/No unit.

Comparison	Decision unit	Agreement	Cohen's κ
Human / Gemini	220 path-level decisions	82.7%	0.64
GPT-5.6 Terra / Gemini	Path-level decisions	82.3%	0.645

Table 12: PAS evaluation under two judges. Each value pair reports Gemini / GPT-5.6 Terra. GPT-5.6 Terra re-scores the same responses.

Model	Precision	Recall	PAS
Base			
Claude-Sonnet-4.5	38.08 / 36.0	63.28 / 63.9	43.84 / 42.8
Qwen3-32B	39.91 / 35.3	68.02 / 62.3	46.89 / 41.0
Gemma3-27B	38.53 / 36.7	65.13 / 65.6	44.48 / 43.0
Oracle			
Claude-Sonnet-4.5	47.53 / 44.6	76.66 / 71.8	54.16 / 51.4
Qwen3-32B	46.98 / 43.2	77.91 / 74.9	54.58 / 51.1
Gemma3-27B	46.23 / 42.9	73.63 / 68.9	52.54 / 48.3

Table 13: Oracle − Base Faithfulness gap. Target GT uses target-path sentences; full GT includes all sentences in the question's graph. Change is full GT − target GT.

Model	Gap (target GT)	Gap (full GT)	Change
GPT-5	32.6	32.1	−0.5
Claude-Sonnet-4.5	20.1	19.2	−0.9
Gemma3-27B	17.2	15.4	−1.8
Qwen3-32B	11.0	15.1	+4.1

K Faithfulness under an Expanded Ground Truth

We re-scored Faithfulness for four models with a ground truth that includes distractor facts. The expanded ground truth contains all sentences in the question's graph, including every distractor-perspective sentence (on average 15 sentences per question, compared with 4 target-path sentences), so factually correct distractor content is no longer penalized for not appearing in the target documents. This is an additional check of the Base–Oracle gap; the main evaluation uses the target-path ground

truth defined in Section 3.2.

For these models, the gap is largely preserved under the expanded ground truth (Table 13), suggesting that the Base–Oracle difference is not primarily explained by penalizing factually correct distractor content.

L Unanimous-Item Analysis

We re-aggregated the scores of GPT-5, Claude-Sonnet-4.5, Qwen3-32B, and Gemma3-27B on items where all three annotators agreed on distractor validity, as defined in Appendix F. The scoring is unchanged; this analysis checks whether low precision is driven by ambiguous perspective boundaries.

Table 14: Precision on the original evaluation / unanimous items. Right values report scores on items where all three annotators agreed on distractor validity under the same scoring.

Model	Base	Oracle
GPT-5	39.58 / 38.2	48.75 / 43.2
Claude-Sonnet-4.5	38.08 / 38.2	47.53 / 48.9
Qwen3-32B	39.91 / 41.2	46.98 / 43.7
Gemma3-27B	38.53 / 41.2	46.23 / 41.5

Precision does not rise noticeably on unanimous items (Table 14), suggesting that ambiguous items alone do not explain the low precision. This analysis does not establish a human performance level.

M Cover-All Baseline

A cover-all response reproduces every valid path without exclusion. Its recall assumes unbounded answer length. Table 15 shows this baseline and scores for the ten evaluated models across Base and Oracle settings.

Table 15: Cover-all baseline and model score ranges across Base and Oracle settings.

Metric	Cover-all (no exclusion)	Ten evaluated models
Precision	41.4	27–49
Recall	100	42–79
PAS	57.6	30–56

You are an expert academic researcher tasked with identifying the foundational knowledge within an educational text. Your goal is to analyze the document and extract the top 3~7 most important 'Narrative Core Units' (NCUs).
Your task is not to summarize, but to select and refine the individual claims that are most critical for understanding the document's core logical structure and foundational concepts.

CRITERIA FOR "IMPORTANCE"
Rank and select the NCUs based on the following criteria. A single NCU might satisfy multiple criteria.
1. Foundational Concepts (High Importance): Is it a core definition, a fundamental law, or a guiding principle?
2. Core Relationships (High Importance): Does it explicitly describe a significant causal, comparative, or temporal relationship between two or more key concepts? These are the essential building blocks for a narrative.
3. Broad Applicability (Medium Importance): Is it a general statement that applies broadly, rather than a specific example or data point?

DEFINITIONS & RULES:
1. NCU (Narrative Core Unit):
Rule: Must be a complete, self-contained sentence expressing a single core fact or claim.

CRITICAL ACTION
- You MUST resolve all pronouns (e.g., 'he', 'it', 'this') and ambiguous references by replacing them with the specific names they refer to.
- You are permitted to rephrase the original sentence to ensure it is self-contained and clear, but you MUST preserve the original meaning, all specified relationships, and core concepts exactly.
- This is not summarizing. The rephrased sentence must be a 1:1 semantic equivalent of the original claim, not a shorter version.

2. Bridge-Entity:
Rule: The core specific, proper nouns (e.g., people, places, organizations, specific laws) or highly specific, well-defined concepts (e.g., "Scientific Method", "Theory of Evolution", "Photosynthesis") that the NCU is about.
Normalization Rules:
- Use the most formal and complete name mentioned in the document (e.g., "Charles Darwin" instead of "Darwin").
- Maintain the original and correct capitalization for proper nouns.
- The output must be a JSON list of strings.

3. Bridge-Topic:
Rule: The general subject, theme, or concept that the NCU describes. This should act as a concise, descriptive tag for the claim.
Normalization Rules:
- Use a general and concise phrase , typically 2-4 words.
- To ensure consistency for later clustering, all topics MUST be in lowercase and use singular forms where applicable (e.g., "scientific method" instead of "Scientific Methods").
- The output must be a JSON list of stringsz

Figure 6: **Prompt template for extraction of single-fact sentences and entities (Section 2.3.1).** The LLM decomposes each text chunk T into single-fact sentences $\mathcal{F} = \{f_1, \ldots, f_n\}$ where each f_i contains exactly one discrete piece of information, and pairs every f_i with its key entities e_i. The prompt requires pronoun resolution and 1:1 semantic preservation so that each f_i remains self-contained when later (i) grouped into semantic clusters in Section 2.3.2 and (ii) served as a reference document during evaluation. The two outputs feed the next two construction stages in complementary ways: the single-fact sentences seed the density-based clustering of Section 2.3.2, while the entities e_i contribute the shared-entity signal S_{ent} used by the edge-scoring heuristic in Section 2.3.3.

You are an expert academic researcher. Your task is to synthesize a collection of related factual statements (Narrative Core Units - NCUs) into a single, comprehensive summary sentence. This summary, which we will call a "Meta-NCU," must encapsulate the core concepts and relationships present in the provided NCUs.

Instructions:
1. Synthesize, Don't Just List: Do not simply list the topics. Create a coherent sentence that explains the overarching theme or the central point connecting all the NCUs.
2. Capture Core Concepts: Ensure the most critical concepts and their relationships from the input NCUs are integrated into the Meta-NCU.
3. Be Concise: The output must be a single, clear, and concise sentence.
4. Preserve Meaning: The Meta-NCU should accurately reflect the information given in the source NCUs without introducing new information or personal interpretations.

Input Format:
A list of NCU sentences.

Output Format:
produce only the final Meta-NCU text.

Figure 7: **Prompt template for representative summary generation (Section 2.3.2).** After single-fact sentences in $\mathcal{F}$ are grouped into cohesive clusters $\{C_1, \ldots, C_k\}$, each cluster C_j needs an explicit concept label to function as a graph node. The LLM synthesizes the facts in C_j into a single representative summary s_j that names the shared explanatory focus—for example, condensing facts about inflation and civil wars into "Internal Crisis" as in Section 2.3.2. The instructions require synthesis rather than enumeration and forbid information not present in the input, so that s_j faithfully labels C_j. Each (C_j, s_j) pair becomes a semantic node $v_j \in \mathcal{V}$, on top of which edges and perspectives are constructed in Sections 2.3.3 and 2.3.4.

You are a highly intelligent researcher specializing in identifying logical and narrative connections between concepts. You will be given two summaries of distinct topics, Topic A and Topic B. Your task is to generate all plausible hypotheses for the logical relationships between them.

TASK
Generate a list of all plausible logical hypotheses connecting Topic A and Topic B.
1. Search Exhaustively: You MUST consider relations in both directions (A -> B and B -> A).
2. Discover Multiple Perspectives: Output ALL valid hypotheses you can find, not just the single most obvious one. This is critical.
3. Stay Grounded: For each hypothesis, you MUST provide a concise reason that justifies the relation based ONLY on the provided Summaries or Key Evidence. Do not use any external knowledge.
4. Use Palette or Propose New:
- If a relation from the Relation Palette accurately fits, use its name string (e.g., "CAUSES").
- If NO relation from the palette is a good fit, but you find a valid logical link, you MUST propose a new relation. A new relation proposal MUST be a JSON object: {"new_relation": "NEW_RELATION_NAME", "new_description": "Your description of this new relation."}

OUTPUT FORMAT
Your response MUST be a single JSON object containing a key "hypotheses". This key must hold a list of hypothesis objects. Each hypothesis object MUST follow the structure below.
If no valid relations are found, return {"hypotheses": []}.

Figure 8: **Prompt template for edge generation (Section 2.3.3).** For each top-scoring node pair (v_i, v_j) that passes the heuristic relevance score $\text{Score}(v_i, v_j) = S_{\text{ent}} + S_{\text{prox}} + S_{\text{sem}}$, the LLM receives the two representative summaries with supporting facts and emits one directed labeled edge per explanatory relationship it identifies. The prompt requires (i) bidirectional search ($v_i \to v_j$ and $v_j \to v_i$), (ii) exhaustive enumeration rather than a single best label, and (iii) grounding each relation in the provided evidence, so a node pair may carry multiple edges with distinct labels in $\mathcal{E}$. This is what lets a single endpoint pair (v_s, v_e) later admit multiple valid reasoning paths $\Pi(v_s \to v_e)$—e.g., "External Invasion," "Political Collapse," and "Economic Collapse" from "Military Decay" to "Fall of Western Rome" in Figure 2.

You are a Benchmark Architect and Expert Educational Assessor.
Your task is to analyze a "path group" and generate a set of "controllable" (perspective-biased) benchmark questions, one for each major logical theme you discover.

Task
Your goal is to identify the main "logical themes" in the path group and generate a biased question for each.
1. Analyze and Identify Themes
Read all `valid_paths`. They share the same nodes but have different "relations" (edges).
Your first task is to identify the major logical themes (sub-groups). A "theme" is defined by a specific, recurring logical relation at a key hop.
Example Analysis (for 11->12->2->17): "I see two major themes for the `11->12` hop:
'IS_A_SPECIAL_CASE_OF' (Paths 0-5) and 'PRECEDES' (Paths 6-11). I will create one benchmark question for each theme."

2. Generate Biased Questions (One per Theme)
Identify the Theme: Your theme name (e.g., "Causal") MUST be derived from the relation names (e.g., CAUSES, EXPLAINS), not the reason fields.
Generate a Biased Hint: Your question must hint at this theme using general conceptual words (e.g., "focus on the physical mechanism," "focus on the conceptual sequence").
CRITICAL RULE: DO NOT READ THE REASONS
To prevent answer-leaking, you MUST NOT use the reason fields from the input paths to help you generate the question.
You MUST NOT use direct synonyms of the relation names (e.g., if the theme is CAUSES, do not use the word "cause" or "leads to" in the question).
Anchor the Question: The question must be anchored on all nodes (e.g., Topic 3, Topic 13).
CRITICAL METADATA RULE: The question must be a natural, human-readable sentence and MUST NOT contain any internal metadata or IDs (e.g., "(Topic 3)" or "path_0").
List the `valid_paths_for_this_theme` that exactly match this theme.

OUTPUT FORMAT
Your response MUST be a single JSON object. The output must be a list called `"controllable_benchmarks"`.
Each object in the list must contain the theme, the biased question, and the paths that validate it.

Figure 9: **Prompt template for perspective-targeted question generation (Section 2.3.4).** Given the valid paths $\Pi(v_s \to v_e)$ between an endpoint pair, the LLM groups paths that share the same explanatory focus into a perspective $\mathcal{P}_i \subseteq \Pi(v_s \to v_e)$ (e.g., causal, comparative, structural), generates one question q_i per perspective, and records the subset $\Pi_{\mathcal{P}}^{\text{target}}$ of paths that validate it. The prompt forbids using the `reason` fields of paths and direct synonyms of relation labels. It requires general conceptual hints to indicate the target perspective. Full step-by-step formalism is in Appendix D.

You are an expert educator. Your goal is to provide a focused, clear, and logically structured answer to the user's question.

TASK
Your answer must be a coherent narrative explanation that focuses exclusively on the logical perspective demanded by the user's question. Do not simply list facts.

INSTRUCTIONS
1. Read the [QUESTION] and the [CONTEXT] carefully.
2. If [CONTEXT] is empty (no documents provided): You must answer the [QUESTION] using only your internal expert knowledge.
3. If [CONTEXT] is provided: You MUST base your answer strictly and exclusively on the information within the [CONTEXT] section. Do not use any external knowledge or make assumptions not supported by the provided text. Your entire answer must be grounded in the given documents.

OUTPUT FORMAT
You MUST output only the generated answer text, and nothing else. Do not use JSON, Markdown, or any preamble like "Here is the answer:".

Figure 10: **Answer generation prompt for Base and Oracle settings (Appendix G).** The evaluated model receives a perspective-targeted question q_i produced by the Step 4 prompt (Figure 9) and is instructed to return a coherent narrative explanation focused exclusively on the requested perspective, not an enumerated list of facts. The template branches on the `[CONTEXT]` slot: when empty, the model answers from parametric knowledge alone (Base setting); when populated with the ground-truth single-fact sentences along target paths $\Pi_{\mathcal{P}}^{\text{target}}$, the model must ground its answer strictly in those reference documents (Oracle setting).

You are a meticulous logical evaluator.
Your task is to evaluate if a given "Answer" successfully reproduces any of the "Valid Paths" from a provided list.

TASK
Iterate and Verify Each Path. You must iterate through every path in the Valid_Paths_List. For each path:
1. Internal Check: Check if the "Answer to Evaluate" semantically reproduces ALL of the edges listed in that path's edges object.
2. All-or-Nothing Judgment:
- If ALL edges match, this path's alignment is "Yes".
- If ANY edge does not match, this path's alignment is "No".
3. Provide a reason for the judgment concisely.

Figure 11: **LLM-as-a-Judge prompt for the Perspective Alignment Score (Section 3.2).** Following the CheckEval-style checklist formulation (Lee et al., 2025), the judge does not score a response holistically. It iterates through every reasoning path $\pi \in \Pi(q)$ in the supplied `Valid_Paths_List` and, for each path, returns an all-or-nothing judgment of whether the response semantically reproduces *all* of its constituent edges. The resulting identified set $A = \{\pi \in \Pi(q) : \text{Judge}(\pi, \text{response}) = \text{present}\}$ yields perspective precision $P_{\text{Prec}} = |A \cap T|/|A|$, recall $P_{\text{Rec}} = |A \cap T|/|T|$, and their harmonic mean PAS. Treating each path as an atomic unit is a deliberate granularity choice; sub-path violations are discussed in the Limitations.

You are a meticulous fact-checker. Your task is to evaluate the Factual Accuracy of a given "Answer" based only on a provided list of "Ground Truth Sources".
You are an expert educator. Your goal is to evaluate the accuracy of the user's answer to the question.

TASK
Step 1: Deconstruct the Answer
- Break down the "Answer to Evaluate" into a list of its core factual claims (claims_made).

Step 2: Verify Each Claim (The "Checklist")
Iterate through each claim from Step 1. For each claim, check if it is directly supported by any text in the "Ground Truth Sources".
- If supported, set status to "Yes" and provide the supporting_evidence (the exact NCU from the sources).
- If not supported (i.e., it is a hallucination or not in the sources), set status to "No" and supporting_evidence to null.

Figure 12: **LLM-as-a-Judge prompt for Faithfulness (Section 3.2).** Following the standard claim-level verification protocol (Es et al., 2024), the prompt is organized into two stages. **Stage 1 (Deconstruct):** the judge breaks the response into a list of discrete factual claims (`claims_made`). **Stage 2 (Verify):** each claim is checked against the `Ground_Truth_Sources`—the ground-truth single-fact sentences drawn from target paths $\Pi_{\mathcal{P}}^{\text{target}}$—and marked Yes (with the exact supporting single-fact sentence) or No (unsupported). Faithfulness is the proportion of supported claims. In the Oracle setting these same documents are exposed to the model during generation; in the Base setting they are used only post-hoc to verify the claims in the response (Section 3.2).