CAN DEEP RESEARCH AGENTS RETRIEVE AND ORGANIZE? EVALUATING THE SYNTHESIS GAP WITH EXPERT TAXONOMIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Research Agents increasingly automate survey writing, yet existing benchmarks do not jointly test whether they retrieve the papers experts consider essential and organize those papers into paper-grounded taxonomies. We introduce TAXOBENCH, a benchmark built from 72 highly cited LLM surveys, 3,815 cited papers, and their expert-authored taxonomies. TAXOBENCH evaluates systems in two settings: **Deep Research** mode measures end-to-end retrieval and organization from a topic, while **Bottom-Up** mode provides the expert paper set and isolates organization. We evaluate leaf-level assignments with ARI and V-Measure and hierarchy-level structure with US-TED, US-NTED, and SEM-PATH. Across 7 Deep Research Agents and 16 LLM configurations, the best agent retrieves only 20.92% of expert-cited papers, and none of 70 standard Bottom-Up runs reaches the experts' average depth of 4.86. A controlled probe shows that models which match this depth do so by fragmenting the taxonomy, reducing alignment with the expert reference. We further find that raw SEM-PATH remains near a no-organization floor even when a newer model generation gains 3.68 pp ARI; after depth matching, humans lead on all 10 matched surveys by 13.27 pp. These results identify retrieval and hierarchical organization as separate bottlenecks and show why hierarchy metrics must be calibrated before they are used to compare models.

1 INTRODUCTION

Deep Research (DR) Agents increasingly automate survey writing by searching the web, collecting papers, and producing structured overviews (Schmidgall et al., 2025; Zhang et al., 2025b; Wang et al., 2024; Yan et al., 2025), as sketched in Figure 1. Recent DR surveys (Shi et al., 2025) identify four sub-components: Query Planning, Information Acquisition, Memory Management, and Answer Generation. Concurrent benchmarks evaluate report-quality and rubric-following aspects of Answer Generation (Zhang et al., 2025a; Sharma et al., 2025). Whether these systems match human experts at two foundational abilities at the intersection of *Information Acquisition* and *structural Answer Generation* remains unclear: retrieving essential papers and organizing them into expert-like taxonomies. Expert surveys do not arise from text generation alone; experts read broadly, identify a set of core works, and iteratively synthesize a hierarchical taxonomy that becomes the backbone of the survey. Current benchmarks focus on writing quality, factuality, or citation correctness (Eldifrawi et al., 2024; Wadden et al., 2020), rather than on whether an agent retrieves the papers experts would cite and organizes them in an expert-like structure.

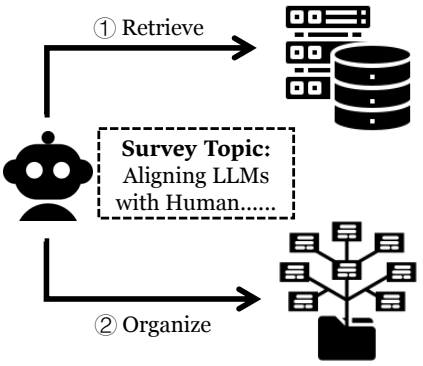


Figure 1: Deep Research Agents for survey generation: given a topic, retrieve papers, then organize them into a taxonomy.

We identify two challenges. **C1: Retrieval.** Existing benchmarks do not provide expert-curated paper sets for survey-oriented retrieval at scale, although standard Recall, Precision, and F1 suffice once

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

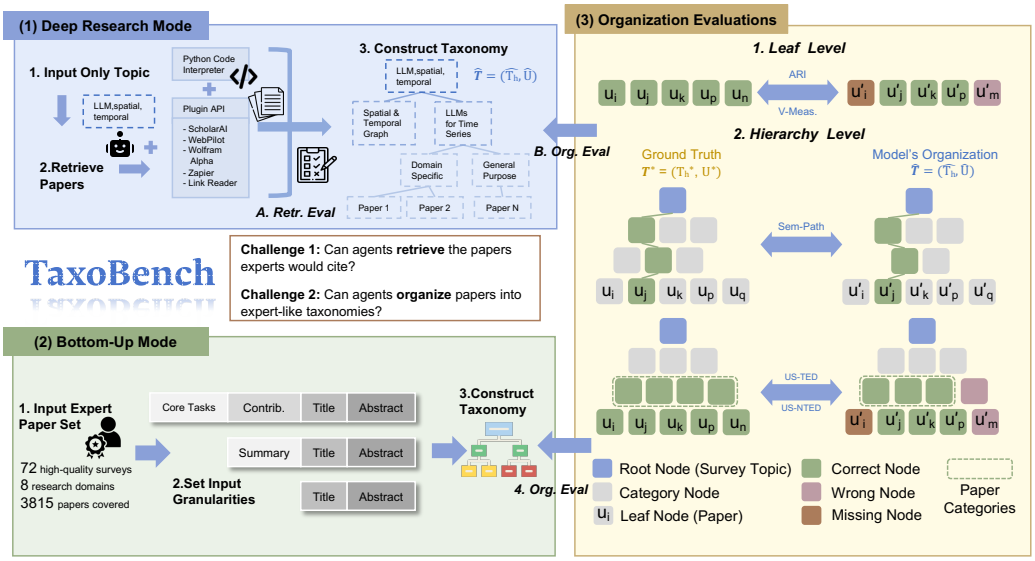


Figure 2: Overview of TAXOBENCH. In **Deep Research mode**, agents receive only a topic, retrieve papers with web tools, and build a taxonomy. In **Bottom-Up mode**, we provide the expert paper set to isolate organization. We compare the output with the expert reference at the leaf level using ARI and V-Measure and at the hierarchy level using US-TED/US-NTED and SEM-PATH.

such ground truth exists. **C2: Organization.** Standard clustering metrics such as ARI and V-Measure capture only partitions. They cannot distinguish confusion between sibling categories from errors that move content across distant branches. We therefore evaluate both the *leaf level* of paper-to-category assignment and the *hierarchy level* of taxonomy structure, including node semantics and parent–child relations.

We address these challenges with TAXOBENCH, which contains 72 highly cited LLM surveys, their expert-authored taxonomies, and 3,815 cited papers manually mapped to expert categories. We use the published taxonomies as references without assuming that each topic has only one valid organization. Figure 2 shows our two evaluation modes. **Deep Research mode** tests end-to-end performance from a topic, while **Bottom-Up mode** provides the expert paper set and isolates organization. We propose three hierarchy-aware metrics, US-TED, US-NTED, and SEM-PATH, to measure structure beyond flat clustering. We also separate **alignment-based** findings, which measure agreement with the selected expert reference, from **capability-based** findings, which use retrieval Recall and judge-free tree statistics paired with the expert values. Section 3.4 defines this distinction.

We evaluate 7 Deep Research Agents and 16 frontier LLM configurations under this distinction. On the capability side, the best agent retrieves only 20.92% of expert-cited papers. None of the 70 Bottom-Up runs comes within 0.89 levels of the expert depth of 4.86, and this deficit alone bounds SEM-PATH at 48.6%. Our depth-controlled probe rules out the prompt as the cause: models reach expert depth by fragmenting the taxonomy, which lowers their alignment scores. On the alignment side, every model given the exact expert paper set scores 28–30% SEM-PATH. Calibration places this range just above the 27.49% no-organization floor, and a newer model generation that gains 3.68 pp ARI shifts raw SEM-PATH by only 0.87 pp. Depth-matching lifts current-generation scores to 32–42% and widens the spread from 1.91 to 9.45 pp; on a matched 10-survey subset, humans average 54.09% versus 38.71% pooled across model configurations. We further find that randomizing paper placement while retaining expert labels costs only 23 pp, whereas model scores remain another 35 pp lower. Thus, most of the remaining gap comes from different category vocabularies rather than paper placement.

The contributions are threefold:

1. We introduce TAXOBENCH, which pairs 72 expert-authored survey taxonomies with 3,815 cited papers and supports separate evaluation of retrieval and organization (Section 2).

2. We develop three hierarchy-aware metrics, US-TED, US-NTED, and SEM-PATH, and separate alignment-based from capability-based findings. Calibrating SEM-PATH reveals that it is nearly flat for shallow trees; after depth matching, a 13.27 pp paired human advantage remains on all 10 matched surveys (Apps. B and K.4).
3. We evaluate 7 Deep Research Agents and 16 frontier LLM configurations. The best retrieval Recall is 20.92%, and a depth deficit persists across 70 runs and two model generations. A controlled probe shows that forcing depth fragments taxonomies rather than improving their organization (App. F).

2 TAXOBENCH

We define a *taxonomy* as a *category hierarchy* over topics together with a *paper-to-category assignment* that maps each paper to a leaf category. We extract these taxonomies from published expert surveys and use them as references for evaluating agents. Section 3 gives the formal definitions, Table 1 summarizes the dataset, and Figure 3 shows how reference depth is distributed over the 72 surveys.

Statistic	Value
Survey collection	
Surveys / total papers	72 / 3,815
Avg. citations per survey	354.5
Per-taxonomy (mean ± std)	
Papers per taxonomy	53.0 ± 20.6
Hierarchy depth	4.9 ± 0.8
Paper categories	14.0 ± 6.5
Papers per category	3.8 ± 3.1
Branching factor	2.5 ± 0.5

Table 1: **TAXOBENCH dataset statistics.**

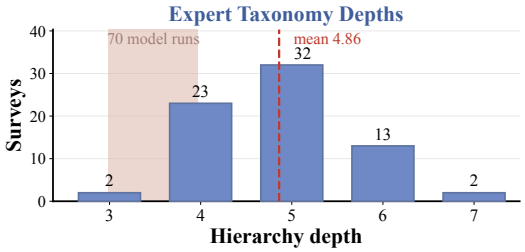


Figure 3: **Reference depth is not an outlier effect.** Only 2 of 72 expert taxonomies fall inside the shaded band of Bottom-Up run depths.

Survey collection. We begin with more than 200 candidate LLM-related surveys and apply four filters. We retain surveys with an explicit taxonomy figure, then filter by citation count; the retained surveys average 354.5 citations. We exclude surveys whose scope is too broad for a coherent taxonomy, and Ph.D.-level researchers verify the remaining set. The final 72 surveys span 8 AI/ML subdomains. App. M reports the count after each stage. We also record each survey's publication date t_s as the retrieval cutoff described in Section 3.2.

Taxonomy extraction. Our Ph.D.-level annotators transcribe each published taxonomy figure rather than redesigning it. They then read the cited papers and map them to categories. We disambiguate and deduplicate papers and check each placement against the survey text, as detailed in App. N. This procedure limits subjectivity because we copy the hierarchy and follow the source authors' placement decisions. We do not report inter-annotator κ for this step and identify it as the highest-priority gap for a future release in App. Q. In a separate baseline, human annotators build taxonomies from scratch under the same Bottom-Up requirements, without seeing the reference; App. K.4 reports those results.

Multi-category papers. Some papers belong to several categories. Because ARI and V-Measure require one assignment, we retain the most contextually appropriate category. More than 70% of surveys contain at most 10 multi-category papers. Removing all such papers changes Claude's SEM-PATH only from 29.16% to 29.82%. Details are in App. L.8.

3 EVALUATION FRAMEWORK

This section specifies the evaluation protocol used throughout the paper. We first define the evaluation objects and notation (Section 3.1), then present metrics for retrieval capability (Section 3.2) and organization capability at the leaf and hierarchy levels (Section 3.3).

3.1 PRELIMINARIES AND NOTATION

Definition 3.1 (Taxonomy)**.** A **taxonomy** $T = (T_h, U)$ consists of (i) a **category hierarchy** $T_h = (\mathcal{C}, E)$, a rooted tree over category labels $\mathcal{C}$ with parent–child edges E; and (ii) a **paper-to-category assignment** $U : \mathcal{P} \to \mathcal{C}_p$ sending each paper to a **paper category** $c_p \in \mathcal{C}_p \subseteq \mathcal{C}$, the leaves of T_h under which papers are attached.

Notation and modes. We write $T^* = (T_h^*, U^*)$ for the expert taxonomy and $\hat{T} = (\hat{T}_h, \hat{U})$ for a model output. We denote the expert-cited and model-retrieved paper sets by $\mathcal{P}^*$ and $\hat{\mathcal{P}}$, the corresponding paper-category sets by $\mathcal{C}_p^*$ and $\hat{\mathcal{C}}_p$, the subtree size at u by $|T_u|$, and its children by $\mathrm{Ch}(u)$. The modes use different paper universes. **Deep Research** evaluates retrieval and organization end to end. **Bottom-Up** sets $\mathcal{P} = \mathcal{P}^*$ and asks models to organize a fixed paper set.

3.2 RETRIEVAL EVALUATION

We evaluate retrieval only in Deep Research mode, where the model must recover $\mathcal{P}^*$ from a topic query. We report set-based metrics under the aggregation convention of App. L.3, which is not uniform across the three columns:

$$\textsc{Recall} = \frac{|\mathcal{P}^* \cap \hat{\mathcal{P}}|}{|\mathcal{P}^*|}, \qquad \textsc{Precision} = \frac{|\mathcal{P}^* \cap \hat{\mathcal{P}}|}{|\hat{\mathcal{P}}|}, \qquad \text{F1} = \frac{2 \cdot \textsc{Precision} \cdot \textsc{Recall}}{\textsc{Precision} + \textsc{Recall}}. \quad (1)$$

We micro-average Recall by pooling hits and expert papers, macro-average Precision, and compute F1 as the harmonic mean of those two values. Table 2 requires all three rules, so its F1 is not the per-survey or pooled F1. Macro-averaged Recall is 0.5–1.7 pp higher but preserves the ordering. Details are in App. L.3.

Temporal cutoff specification. We use each survey's publication date t_s as the retrieval cutoff. Because $\mathcal{P}_s^*$ contains only papers published before t_s, newer retrieved papers cannot enter the Recall numerator. We compute Precision and F1 over the full output. App. L.1 reports temporally filtered Precision and verifies that filtering leaves Recall unchanged.

Source-level safeguards. We define contamination here as retrieving the source survey instead of synthesizing primary literature. We audit all 504 runs and find **119 self-retrievals**, concentrated among the weakest agents. Web search can naturally return a topic's best-known survey, so this alone does not show that an agent copied its bibliography. More importantly, removing all 119 cases leaves every Recall value unchanged because source-survey titles are not members of the expert paper sets. Copying a bibliography would also exceed the observed 20.92% Recall ceiling. We provide the full audit in App. L.2.

3.3 ORGANIZATION EVALUATION

We evaluate organization at the **Leaf Level**, using paper-to-category assignments, and at the **Hierarchy Level**, using taxonomy structure.

3.3.1 LEAF-LEVEL: PAPER-TO-CATEGORY ASSIGNMENT

Given a paper universe $\mathcal{P}$, we compare the expert assignment $U^* : \mathcal{P} \to \mathcal{C}_p^*$ with the model assignment $\hat{U} : \mathcal{P} \to \hat{\mathcal{C}}_p$. Since ARI and V-Measure evaluate partitions without encoding hierarchical structure, we use them here and introduce hierarchy-level metrics in Section 3.3.2.

Leaf-level metrics. We use the chance-corrected **Adjusted Rand Index** (ARI) (Hubert & Arabie, 1985) and **V-Measure** (Rosenberg & Hirschberg, 2007); the latter reports homogeneity ($\textsc{Hom.} = 1 - H(U^*|\hat{U})/H(U^*)$) and completeness ($\textsc{Comp.} = 1 - H(\hat{U}|U^*)/H(\hat{U})$) and combines them as $\text{V-}\textsc{Measure} = \frac{2\,\textsc{Hom.}\,\textsc{Comp.}}{\textsc{Hom.}+\textsc{Comp.}}$. Full forms are in App. A.1.

Deep Research mode. Retrieval limits what an agent can organize, so we report two views. Our **retrieval-conditioned** view restricts $\mathcal{P}$ to $\mathcal{P}^* \cap \hat{\mathcal{P}}$ and gives $\text{ARI}_\cap$ and V-$\text{MEASURE}_\cap$. Our **end-to-end** view sets $\mathcal{P} = \mathcal{P}^*$ and labels unretrieved papers $\perp$. The end-to-end view includes the retrieval bottleneck and does not replace the retrieval metrics.

3.3.2 HIERARCHY-LEVEL: TAXONOMY HIERARCHY STRUCTURE

Flat partition metrics such as ARI and V-Measure cannot distinguish sibling confusion from errors that move content across distant branches. Soft-cardinality metrics such as NSR/NSP (Fränti & Mariescu-Istodor, 2023) are *structure-blind* for the same reason, and can be perfect even when subtrees are entirely rewired (App. A.6). We therefore compare the hierarchies T_h^* and $\hat{T}_h$ directly, keeping only internal categories and edges and excluding paper nodes, with two complementary metrics: US-(N)TED gives a *global* edit distance under unordered sibling matching, and SEM-PATH a *local*, per-paper measure of root-to-leaf path consistency.

Unordered Semantic Tree Edit Distance (US-TED). We extend the STED framework (Wang et al., 2025) to the unordered siblings of a taxonomy, so sibling permutation does not affect US-TED. Using a text embedding $\mathbf{e}(\cdot)$ from Sentence-Transformers, we define semantic label similarity and renaming cost as

$$\text{Sim}(x, y) \coloneqq \max\big(0, \cos(\mathbf{e}(x), \mathbf{e}(y))\big), \qquad \text{cost}_{\text{ren}}(x \to y) \coloneqq 1 - \text{Sim}(x, y) \in [0, 1], \tag{2}$$

with insertion/deletion cost 1 per node. Then $D(u, v) \coloneqq \text{cost}_{\text{ren}}(u \to v) + \text{MatchCost}(\text{Ch}(u), \text{Ch}(v))$, where MatchCost is a minimum-cost bipartite matching (Hungarian) and unmatched children correspond to deleting or inserting whole subtrees, charged by subtree size. The tree-level distance is US-TED$(T_h^*, \hat{T}_h) \coloneqq D(r^*, \hat{r})$ at the roots, which we normalize by hierarchy size for cross-instance comparability: US-NTED $\coloneqq$ US-TED$(T_h^*, \hat{T}_h)/(|T_h^*| + |\hat{T}_h|)$. Lower values indicate closer structural alignment; the full assignment formulation, boundedness, and properties are in App. A.3. We report US-NTED as a percentage in all tables.

Semantic Path Similarity (SEM-PATH). US-(N)TED captures global tree divergence but does not test the ancestor chain of each paper. We therefore define SEM-PATH. Let $D_a \subseteq \mathcal{P}^* \times \hat{\mathcal{P}}$ be the aligned paper pairs described in App. A.5. In Bottom-Up mode, D_a covers $\mathcal{P}^*$; in Deep Research mode, it covers $\mathcal{P}^* \cap \hat{\mathcal{P}}$. For each pair $(d, \hat{d}) \in D_a$, we remove the paper node from the root-to-paper paths in T^* and $\hat{T}$ to obtain ancestor-label sequences $S = (s_1, \ldots, s_m)$ and $\hat{S} = (\hat{s}_1, \ldots, \hat{s}_n)$. With clipped cosine distance $\delta(x, y) \coloneqq 1 - \text{Sim}(x, y)$ and $m \leq n$ without loss of generality, we take the order-preserving minimum-cost alignment over strictly increasing $f : \{1, \ldots, m\} \to \{1, \ldots, n\}$:

$$J(S, \hat{S}) \coloneqq \min_f \sum_{i=1}^{m} \delta(s_i, \hat{s}_{f(i)}) \; + \; \lambda\,(n - m), \tag{3}$$

with $\lambda = 1$, which charges an unmatched ancestor exactly as much as the maximum distance between two matched labels; a sensitivity analysis over $\lambda \in \{0.5, 1, 1.5, 2\}$ is in App. L.6. Letting J_d be the minimum cost for $(d, \hat{d})$ over its candidate ancestor-path pairs,

$$\text{SEM-PATH} \coloneqq \frac{1}{|D_a|} \sum_{(d,\hat{d}) \in D_a} \frac{1}{1 + J_d}. \tag{4}$$

Unlike ARI, SEM-PATH has no chance correction, so raw values need the floors and the depth-matched variant of App. B.

3.4 LLM-AS-JUDGE AND DIAGNOSTIC SPLIT

LLM-as-Judge. We also prompt GPT-4o to compare T_h^* and $\hat{T}_h$ on four 1–5 scales: **Coverage**, **Organization** or MECE, **Logic** or parent–child consistency, and **Topology**. Agreement with human evaluators reaches Cohen's $\kappa = 0.8909$. Details are in Apps. J and O.3.

How much reference does each finding use? Because no finding is fully reference-free, we grade each one by the expert information it uses. **L1** metrics, including ARI, V-MEASURE, US-(N)TED,

Deep Research Agent	Recall↑	Precision↑	F1↑
o3	**20.92%**	29.29%	**24.41%**
Grok	12.82%	**29.35%**	17.85%
Gemini	15.23%	18.92%	16.88%
Perplexity	6.61%	7.47%	7.01%
DeepSeek	4.61%	14.26%	6.97%
Qwen	4.35%	7.94%	5.62%
Doubao	3.15%	3.83%	3.46%

Table 2: **Retrieval capability** of Deep Research Agents when given the same survey topic as human experts. Best results are **bold**; 3-run Recall std ≤ 3.01 pp (App. L.4).

and SEM-PATH, use all labels and topology in T_h^*. A low score therefore means divergence from this expert reference. **L2** comparisons use only scalar statistics of T_h^*, such as depth and category count, without a judge or threshold. **L3** retrieval Recall uses $\mathcal{P}^*$ but no taxonomy. L2 and L3 remain valid when experts disagree about taxonomy design, although they do not establish that a model tree is indefensible on its own terms. Details are in App. Q.

4 BENCHMARKING LLMS ON TAXOBENCH

We address three research questions. In **Deep Research mode**, we evaluate 7 agents end to end from a topic: o3, Grok, Gemini, Perplexity, DeepSeek, Qwen, and Doubao. In **Bottom-Up mode**, we provide $\mathcal{P}^*$ to the newest reachable model from each of seven families: Claude-Sonnet-5, GPT-5.6-sol, Gemini-3.1-Pro, DeepSeek-V4-Pro, Qwen3.8-Max, Kimi-K3, and Grok-4. We run each model at its lowest and highest reasoning effort and add a second GPT tier as a within-family control, for 16 configurations in total. Details are in App. H. App. I reports the 12 configurations used to build the benchmark. Following Zhang et al. (2026), we test **Title+Abstract**, **+Summary**, and **+Core-task&Contributions** inputs. Claude-Sonnet-4.6 produces the auxiliary text. Because it shares a family with one evaluated model, we repeat all 72 surveys with hand-written inputs; SEM-PATH shifts by at most 0.75 pp. Apps. P, K.2, and L.7 provide details.

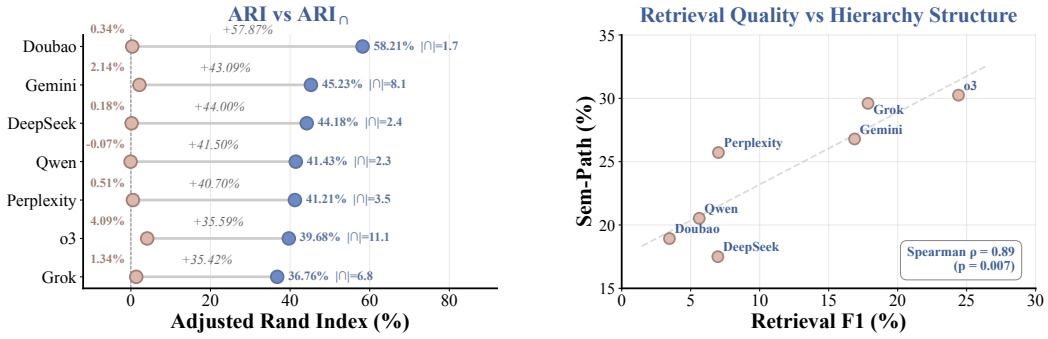
4.1 RQ I: CAN CURRENT AGENTS REPLICATE EXPERT-LEVEL LITERATURE DISCOVERY?

Finding 1: Deep Research agents exhibit severe retrieval bottlenecks. We find that current agents do not retrieve core literature at scale. The best agent, o3, reaches 20.92% Recall and 24.41% F1, so it misses nearly 80% of expert-cited papers. Four of the seven agents fall below 10% F1. Precision is also limited: Grok leads at 29.35%, indicating that peripheral papers dominate the retrieved sets. This result does not use the expert taxonomy, and any pipeline built on these agents inherits a Recall bound of at most 21%.

4.2 RQ II: HOW WELL CAN AGENTS ORGANIZE THE RETRIEVED SUBSET?

Finding 2: Conditioning on retrieved papers raises every score, but intersection size confounds the gain. When we restrict evaluation to $\mathcal{P}^* \cap \hat{\mathcal{P}}$, ARI rises from below 5% to 37–58% for all seven agents; Table 3 and Figure 4a report the result. This increase suggests that agents can separate core papers once they find them, but the ranking reverses with retrieval quality. Doubao scores 58.21% on intersections averaging only 1.7 papers, while o3 scores 39.68% on intersections averaging 11.1 papers. Small intersections make clustering easier, and intersection size correlates negatively with $\text{ARI}_\cap$ at $\rho = -0.46$. The corresponding $p = 0.29$ is inconclusive for only seven agents. We therefore use $\text{ARI}_\cap$ to show that retrieval dominates the end-to-end score, not to rank organization skill. App. E reports each conditioned score with its domain size.

Finding 3: Retrieval quality correlates with hierarchy-level structure. We observe a strong association between F1 and SEM-PATH across the 7 DR agents, with Spearman $\rho = 0.89$ and $p = 0.007$ in Figure 4b. o3 leads on both measures, while Doubao and Qwen score low on both. We hypothesize that off-topic papers lead agents toward a different hierarchy, but seven agents are insufficient to test this mechanism.



(a) ARI vs. ARI$_\cap$; labels show $|\cap|$.

(b) Retrieval F1 vs. SEM-PATH.

Figure 4: We identify retrieval as the dominant bottleneck in Deep Research mode. Panel (a) compares end-to-end ARI with ARI$_\cap$ on retrieved expert papers, with intersection size shown to make the conditioning effect explicit. Details are in App. E. Panel (b) shows the association between F1 and SEM-PATH across the 7 agents, with $\rho = 0.89$ and $p = 0.007$.

Deep Research Agent	Leaf-Level				Hierarchy-Level		
	ARI↑	V-Meas.↑	$\lvert\mathcal{P}^*\cap\hat{\mathcal{P}}\rvert$	ARI$_\cap$↑ / V-Meas.$_\cap$↑	US-TED↓	US-NTED↓	SEM-PATH↑
o3	**4.09%**	**26.05%**	**11.1**	39.68% / 75.94%	29.25	74.02%	**30.25%**
Grok	1.34%	20.03%	6.8	36.76% / 76.80%	**27.14**	**72.80%**	29.60%
Gemini	2.14%	23.57%	8.1	45.23% / **82.24%**	40.56	79.02%	26.79%
Perplexity	0.51%	12.79%	3.5	41.21% / 79.13%	56.65	83.93%	25.72%
DeepSeek	0.18%	8.33%	2.4	44.18% / 81.70%	29.97	74.96%	17.50%
Qwen	-0.07%	7.59%	2.3	41.43% / 79.64%	45.09	76.36%	20.53%
Doubao	0.34%	6.57%	1.7	**58.21%** / 81.28%	39.06	75.13%	18.93%

Table 3: **Organization capability** in **Deep Research mode**. We compute all columns from the collected Deep Research outputs retained in the private reproducibility archive. We report the mean number of retrieved expert papers, $|\mathcal{P}^*\cap\hat{\mathcal{P}}|$, because it confounds the $_\cap$ metrics. We average those metrics over the 29–67 surveys where at least 2 papers align, since ARI is degenerate for one point. We therefore do not use the $_\cap$ columns to rank organization. Details are in App. E. Best results are **bold**; the 3-run SEM-PATH std is at most 2.14 pp.

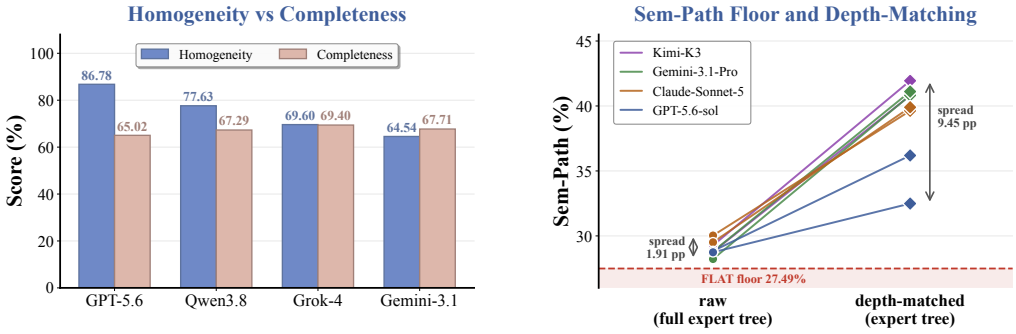
4.3 RQ III: How well can LLMs organize given perfect retrieval?

Finding 4: Many models over-segment at the leaf level, but the failure is not universal. Figure 5a shows higher Homogeneity than Completeness for every model except Gemini-3.1-Pro. The gap is largest for GPT-5.6-sol at 86.78% versus 65.02%, which indicates that this model splits topics into fine-grained clusters instead of consolidating them into broader expert-style categories. App. K includes an illustrative taxonomy comparison. Gemini-3.1-Pro reverses the gap at 64.54% versus 67.71% and has the lowest singleton rate at 5.1%, compared with 17.0% for the expert. When we constrain GPT-5 to the expert number of categories, ARI rises from 13.05% to 20.54% while Homogeneity falls from 73.84% to 62.87%. This probe confirms that over-segmentation can drive high Homogeneity for the affected models. Higher reasoning effort raises ARI for 3 of the 8 models and lowers it for 5; we report this as reference-alignment sensitivity rather than organization-quality evidence (App. H). Details of the category-count probe are in App. G.2.

Finding 5: Raw SEM-PATH is uninformative in this regime; depth-matching recovers a meaningful gap. We observe raw SEM-PATH scores of 28.12–30.03% across all 16 configurations in Figure 5b. Our calibration shows that this narrow range comes from a floor effect. A FLAT taxonomy with no organization scores 27.49%, and randomly rewired expert labels score 28.11%. The weakest configurations are therefore statistically indistinguishable from these baselines. Details are in App. B. The cross-generation comparison supports the same conclusion without synthetic controls. Best ARI rises 3.68 pp from 31.24% to 34.92%, but raw SEM-PATH moves only 0.87 pp and remains within 2.54 pp of FLAT. Details are in App. I. Thus, the raw metric barely reflects a measurable

Model	Leaf-Level				Hierarchy-Level		Judge-free shape	
	ARI↑	Hom.↑	Comp.↑	US-NTED↓	SEM-PATH↑	Depth	#Cat	Singl.
Lowest reasoning effort (none for GPT, `minimal` elsewhere; App. H)								
Claude-Sonnet-5	**34.92%**	75.62%	**69.71%**	74.66%	**30.03%**	3.43	15.9	21.2%
Kimi-K3	33.23%	76.42%	68.69%	78.74%	29.27%	3.06	16.2	16.1%
Grok-4	31.44%	69.60%	69.40%	79.08%	28.13%	3.01	12.9	11.3%
DeepSeek-V4-Pro	29.81%	75.59%	68.68%	78.15%	28.72%	3.18	18.4	28.3%
Qwen3.8-Max	29.81%	77.63%	67.29%	78.13%	29.48%	3.28	18.5	23.2%
Gemini-3.1-Pro	29.67%	64.54%	67.71%	77.79%	28.22%	2.99	**11.0**	**5.1%**
GPT-5.6-sol	21.98%	**86.78%**	65.02%	77.46%	28.84%	3.56	28.3	43.5%
Expert reference	—	—	—	—	—	*4.86*	*14.0*	*17.0%*

Table 4: **Organization capability** in **Bottom-Up mode**. We evaluate the newest reachable model from each of seven families on Title+Abstract inputs for all 72 surveys, using each endpoint's lowest reasoning effort. App. H reports the highest-effort rows, a second GPT tier, and the full 1,152-call comparison. App. I reports the previous generation, and App. L.4 reports 3-run standard deviations. Best results are **bold**. Every model is shallower than the expert depth of 4.86, and SEM-PATH spans only 1.91 pp above the 27.49% FLAT floor.



(a) Homogeneity vs. Completeness.

(b) SEM-PATH calibration.

Figure 5: We observe two organization failures in Bottom-Up mode. In panel (a), higher Homogeneity than Completeness indicates over-segmentation; the gap is largest for GPT-5.6-sol, while Gemini-3.1-Pro reverses it. In panel (b), every raw SEM-PATH score lies within 2.5 pp of the 27.49% FLAT floor. Depth-matching raises scores to 32–42% and widens the spread from 1.91 to 9.45 pp. Details are in App. B. For legibility, panels (a) and (b) show 4 and 8 of the 16 configurations.

generation-level change in leaf-partition alignment. The low level is model behavior rather than metric uselessness: two independent annotators given the same Bottom-Up requirements average 53.91% SEM-PATH on a matched 10-survey subset, clearing FLAT by 26.42 pp, and they agree with each other at 50.56% where models reach 28.21%. ARI inverts this and is saturated: models agree with the expert at 35.76% while the annotators agree with each other at only 22.92%, so we draw no organization conclusions from ARI (App. K.4).

We trace the floor effect to the absence of chance correction and the depth deficit in Finding 6. Collapsing the expert tree to each model's depth raises scores to 32–42% and widens the spread from 1.91 to **9.45 pp**, while a paired human advantage remains on all 10 matched surveys (+13.27 pp; 95% CI $[10.69, 16.06]$). This separation is our primary validation; the rise in correlation with ARI from r=0.29 (p=0.27) to r=0.81 (p=10^{-4}) is auxiliary evidence of recovered model-level variation. The control does not solve every problem. Randomizing paper placement while retaining expert labels still scores 77.29%, while models remain 35 pp lower. We conclude that SEM-PATH mainly penalizes differences in category vocabulary rather than paper placement. Details are in Apps. B and K.4.

Finding 6: Models do not spontaneously build deep taxonomies, and forcing depth does not help. We compare model and expert trees using judge-free, threshold-free statistics in App. D. Expert taxonomies average 4.86 category levels, while all 70 Bottom-Up runs (16 current-generation configurations and 54 previous-generation runs across four input conditions) average 2.99–3.97, a

band that contains only 2 of the 72 expert taxonomies (Figure 3). Even the deepest configuration remains 0.89 levels short. Details are in App. I. Under the same Bottom-Up requirements, two independent annotators on the matched 10-survey subset reach depth 4.60 and 4.70, matching the 4.70 expert mean on those surveys (App. K.4), so the deficit is largely model-specific. Because SEM-PATH charges $\lambda = 1$ for each unmatched ancestor, the model shortfall bounds the per-paper score at 45.8–52.4% in App. D, Equation 30, and contributes directly to the floor effect in Finding 5.

We test whether our three-level prompt exemplar causes the shallow trees with an 864-call depth-controlled probe in App. F, which supplies an explicit depth target and holds the rest of the prompt fixed. Three of four models comply, and GPT-5 reaches the expert depth of 4.86, but matching depth does not improve organization. GPT-5's SEM-PATH falls by 2.77 pp, its ARI drops from 29.02% to 14.65%, and the added levels fragment the taxonomy, raising singleton leaves to 66.7% against the expert's 17.0%. No model–condition pair gains more than 0.3 pp. We therefore interpret the depth deficit as a failure to build coherent hierarchical refinement, not as a hard inability to add levels. The 16 configurations in Table 4 show the same tension without prompt manipulation: depth predicts singleton rate at $\rho = 0.83$ and $p = 10^{-4}$ but does not predict ARI at $\rho = -0.27$.

5 RELATED WORK

Deep-research-aided survey generation. Most pipelines adopt a structure-then-content scheme: AutoSurvey (Wang et al., 2024) uses two-stage generation, Hu et al. (2024) build taxonomies from citation networks, and SurveyForge (Yan et al., 2025) pairs outline heuristics learned from human surveys with retrieval (Google, 2025; Li et al., 2025b; OpenAI, 2025b; Zheng et al., 2025). Closest to our organization setting, TaxoAlign/CS-TaxoBench (Lahiri et al., 2025) evaluates scholarly taxonomy generation against human-written survey taxonomies. TAXOBENCH differs by grounding each taxonomy in paper-to-category assignments, separating retrieval from organization, and auditing hierarchy metrics under expert paper sets rather than evaluating taxonomy outlines alone.

Taxonomy induction and its evaluation. Taxonomy construction predates LLMs (Zhang et al., 2018; Shen et al., 2018; Zeng et al., 2024), but its evaluation is edge-based: TExEval-2 (Bordea et al., 2016) scores is-a edges and hierarchical P/R/F1 (Kiritchenko et al., 2006) credits ancestor overlap, both presupposing a shared vocabulary. Model-generated taxonomies invent their own labels, so edge matching is undefined; US-TED and SEM-PATH match labels semantically instead.

Deep-research evaluation. Existing work targets multi-hop retrieval (Chen et al., 2025; Yu et al., 2025), web browsing (Deng et al., 2023; Wu et al., 2025), and report generation (Li et al., 2025a; Du et al., 2025). FINDER (Zhang et al., 2025a) scores reports against structured checklists, while RESEARCHRUBRICS (Sharma et al., 2025) uses human rubrics for reports (Shi et al., 2025). TAXOBENCH instead isolates Information Acquisition and structural Answer Generation with expert-authored references rather than LLM-judged outlines (Sun et al., 2025; Yan et al., 2025).

6 CONCLUSION

Survey generation requires both finding the right evidence and organizing it into a useful structure. We introduced TAXOBENCH, which pairs 72 expert-authored taxonomies with 3,815 mapped papers and separates these abilities through complementary Deep Research and Bottom-Up modes, making the source of failure observable. Retrieval F1 is strongly associated with SEM-PATH (ρ=0.89) across agents, yet perfect retrieval leaves a 13.27 pp depth-matched human gap and a 0.89-level depth deficit.

To measure organization beyond flat assignments, we developed US-TED, US-NTED, and SEM-PATH, separated alignment-based from capability-based claims, and calibrated them against synthetic controls and a two-annotator human baseline. The calibration bounds what each metric can show: raw SEM-PATH moves only 0.87 pp when a model generation gains 3.68 pp ARI, and ARI is saturated, since models agree with the expert more than two independent annotators agree with each other. TAXOBENCH therefore supports a more diagnostic protocol: separate retrieval from organization, distinguish reference alignment from reference-light evidence, and calibrate against controls and human agreement. A chance-corrected hierarchy metric, more expert references, and richer agentic workflows follow directly.

AI USE STATEMENT

Generative AI systems are the object of study in this work. Every model configuration we benchmark, together with the LLM-as-Judge used for the semantic matching that underlies US-TED, US-NTED, and SEM-PATH, is part of our stated methodology rather than an undisclosed aid: the exact model versions and endpoints are listed in App. P, every prompt is reproduced verbatim in App. O, and the judge is validated against expert agreement in App. J.

Beyond this methodological use, we used generative AI tools only to polish the wording of the manuscript, that is, to rephrase sentences and correct grammar in text the authors had already drafted. We did not use generative AI tools for research ideation, study design, dataset construction, retrieval or selection of the cited literature, data analysis, or the interpretation of results, and no reported number was produced by AI-generated code that the authors did not write and verify. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including its text, claims, and released artifacts.

ETHICS STATEMENT

Data provenance and privacy. We construct the dataset entirely from publicly available computer science surveys and their cited papers. We use scientific content such as titles and abstracts and do not collect non-public personal or sensitive information. We use these materials only for non-commercial academic research.

Human subjects. Most experiments evaluate AI systems. Two auxiliary studies involve human participants: expert agreement with GPT-4o judgments in App. J and non-author human taxonomy construction in App. K.4. These minimal-risk tasks collect professional judgments about scientific taxonomies, not sensitive personal information. Participants were recruited from CS graduate students familiar with AI/ML literature, and the paper reports only aggregated results under pseudonymous identifiers. App. K.5 states what participants were told before starting, including that they could withdraw. We do not release participant-level material.

Broader impact and potential misuse. TAXOBENCH can support research on automated survey generation and help researchers navigate a growing scientific literature. The same technology can also produce low-quality reviews and create academic-integrity risks. Our experiments show that current agents remain far below expert performance, so we recommend AI-assisted survey writing with expert oversight rather than full automation.

REPRODUCIBILITY STATEMENT

All components needed to inspect the benchmark and rerun the public scoring code are specified in the paper and released with the benchmark. The dataset construction pipeline, including the four-stage survey selection procedure and the taxonomy-transcription protocol, is described in Section 2 and Apps. M and N. The separate from-scratch human baseline and the protocol its annotators worked from are in Apps. K.4 and K.5. Formal definitions, algorithms, and proofs for US-TED, US-NTED, and SEM-PATH, together with the paper-title alignment rules, appear in App. A. Every prompt used for taxonomy generation, information extraction, and LLM-as-Judge scoring is reproduced verbatim in App. O, and the exact model versions and endpoints are listed in App. P. App. L reports 3-run error bars for all models in both modes, sensitivity to the choice of sentence encoder, temporal-filtering verification, and multi-category ablations. We release the benchmark data and the evaluation harness implementing the public metrics. We do not publicly release raw model response logs, API traces, or product-generated reports, because they contain proprietary model outputs and provider-specific metadata. We retain those artifacts in a private archive, together with per-call reasoning parameters, reasoning-token counts when available, and scripts used to rebuild the tables from the raw dumps. The aggregation conventions of Table 2 are documented in App. L.3 because they are not uniform across the three columns.

REFERENCES

Alibaba Cloud. Qwen deepresearch. https://www.alibabacloud.com/blog/qwen-deepresearch-when-inspiration-becomes-its-own-reason_602676, 2025a.

Alibaba Cloud. Qwen3-max: Just scale it. https://www.alibabacloud.com/blog/qwen3-max-just-scale-it_602621, 2025b.

Alibaba Cloud. Qwen model documentation. https://www.alibabacloud.com/help/en/model-studio/, 2026.

Anthropic. claude-sonnet-4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025.

Anthropic. Claude models documentation. https://docs.anthropic.com/en/docs/about-claude/models, 2026.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1081–1091, 2016.

ByteDance. Doubao. https://www.doubao.com/chat/, 2025.

Shan Chen, Pedro Moreira, Yuxin Xiao, Samuel Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle Bitterman. Medbrowsecomp: Benchmarking medical deep research and computer use. *arXiv preprint arXiv:2505.14963*, 2025.

DeepSeek. Deepseek. https://chat.deepseek.com/, 2025.

DeepSeek. Deepseek models documentation. https://api-docs.deepseek.com/, 2026.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. *arXiv preprint arXiv:2407.12853*, 2024.

Pasi Fränti and Radu Mariescu-Istodor. Soft precision and recall. *Pattern Recognition Letters*, 167: 115–121, 2023.

Google. Deep research is now available on gemini 2.5 pro experimental. https://blog.google/products/gemini/deep-research-gemini-2-5-pro-experimental/, 2025.

Google. gemini-3-pro-preview. https://aistudio.google.com/prompts/new_chat, 2025a.

Google. Gemini. https://gemini.google.com/app, 2025b.

Google. Gemini models documentation. https://ai.google.dev/gemini-api/docs/models, 2026.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. Taxonomy tree generation from citation graph. *arXiv preprint arXiv:2410.03761*, 2024.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A Fazel Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. *Advances in Artificial Intelligence*, pp. 395–406, 2006.

Avishek Lahiri, Yufang Hou, and Debarshi Kumar Sanyal. Taxoalign: Scholarly taxonomy generation using language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30191–30211, 2025.

Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. Reportbench: Evaluating deep research agents via academic survey tasks. *arXiv preprint arXiv:2508.15804*, 2025a.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*, 2025b.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Moonshot AI. Kimi-k2-0905. https://www.kimi.com/, 2025.

Moonshot AI. Kimi model documentation. https://platform.moonshot.ai/docs/, 2026.

OpenAI. Gpt-5. https://openai.com/index/introducing-gpt-5/, 2025a.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025b.

OpenAI. Openai models documentation. https://platform.openai.com/docs/models, 2026.

Perplexity. Introducing deep research. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research, 2025.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Sentence-Transformers. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, et al. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*, 2025.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2180–2189, 2018.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, et al. Deep research: A systematic survey. *arXiv preprint arXiv:2512.02038*, 2025.

Zhaojun Sun, Xuzhou Zhu, Xuanhe Zhou, Xin Tong, Shuo Wang, Jie Fu, Guoliang Li, Zhiyuan Liu, and Fan Wu. Surveybench: Can llm (-agents) write academic surveys that align with reader needs? *arXiv preprint arXiv:2510.03120*, 2025.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

Guanghui Wang, Jinze Yu, Xing Zhang, Dayuan Jiang, Yin Song, Tomal Deb, Xuefeng Liu, and Peiyang He. Sted and consistency scoring: A framework for evaluating llm structured output reliability. *arXiv preprint arXiv:2512.23712*, 2025.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145, 2024.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*, 2025.

xAI. Grok. `https://grok.com/`, 2025.

xAI. Grok model documentation. `https://docs.x.ai/docs/models`, 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12444–12465, 2025.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, Hualei Zhou, Hansong Xiao, Chunxiao Guo, Wei Peng, Junwei Liu, and Jinjie Gu. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*, 2025.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3093–3102, 2024.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2701–2709, 2018.

Dingling Zhang, He Zhu, Jincheng Ren, Kangqi Song, Xinran Zhou, Boyu Feng, Shudong Liu, Jiabin Luo, Weihao Xie, Zhaohui Wang, et al. How far are we from genuinely useful deep research agents? *arXiv preprint arXiv:2512.01948*, 2025a.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*, 2024.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, Huayu Sha, Binze Hu, Jingqi Tong, Changhao Jiang, Yage Geng, Yuankai Ying, Yue Zhang, Zhangyue Yin, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *arXiv preprint arXiv:2601.01576*, 2026.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*, 2025b.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

APPENDIX ROADMAP

The appendix is organized for lookup rather than chronology. Apps. A–D give formal metric definitions, calibration, and the validity framing. Apps. E–L report the supplementary experiments behind the findings, together with the judge validation, error analysis, human baseline, and robustness checks. Apps. M–Q collect dataset construction, prompts, model details, and limitations.

A METRIC DETAILS

A.1 ADJUSTED RAND INDEX (ARI)

The Adjusted Rand Index (ARI) is a chance-corrected measure of similarity between two data clusterings. Given a set of N papers, let $U = \{u_1, \ldots, u_R\}$ be the ground truth partition (expert taxonomy) and $V = \{v_1, \ldots, v_C\}$ be the model-generated partition.

We first consider all $\binom{N}{2}$ pairs of papers and categorize them into four types based on their assignment in U and V:

- **TP (True Positive)**: The number of pairs that are in the same cluster in both U and V.
- **TN (True Negative)**: The number of pairs that are in different clusters in both U and V.
- **FP (False Positive)**: The number of pairs that are in different clusters in U but in the same cluster in V.
- **FN (False Negative)**: The number of pairs that are in the same cluster in U but in different clusters in V.

The standard Rand Index (RI) measures the percentage of correct decisions:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

However, the RI does not yield a value of 0 for random partitions. The ARI corrects this by normalizing the RI using its expected value under a random permutation model:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{6}$$

Specifically, let n_{ij} be the number of papers in both class u_i and cluster v_j, and let $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$ be the row and column sums of the contingency table. The terms in the ARI formula are calculated as follows:

$$\text{Index} = \sum_{i,j} \binom{n_{ij}}{2} = TP \tag{7}$$

$$\text{Max Index} = \frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] \tag{8}$$

$$\text{Expected Index} = \frac{\left[\sum_i \binom{a_i}{2}\right]\left[\sum_j \binom{b_j}{2}\right]}{\binom{N}{2}} \tag{9}$$

Combining these, the computational formula for ARI is:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}} \tag{10}$$

This linear transformation ensures that ARI has a maximum value of 1 for perfect agreement and an expected value of 0 for random clustering, making it a robust metric for comparing taxonomies with different numbers of clusters.

A.2 Semantic Label Similarity and Cost Calibration

Embedding similarity. For each topic label (node name) x, we compute a dense embedding $\mathbf{e}(x) \in \mathbb{R}^d$ using a fixed sentence encoder. We measure semantic relatedness by cosine similarity

$$c(x, y) \coloneqq \cos(\mathbf{e}(x), \mathbf{e}(y)) = \frac{\mathbf{e}(x)^\top \mathbf{e}(y)}{\|\mathbf{e}(x)\| \, \|\mathbf{e}(y)\|}. \tag{11}$$

Since cosine similarity can be negative in practice, we adopt a clipped similarity used consistently by both US-TED/US-NTED and SEM-PATH:

$$\mathrm{Sim}(x, y) \coloneqq \max\big(0,\, c(x, y)\big) \in [0, 1]. \tag{12}$$

Edit-cost calibration. We convert similarity to a renaming cost by

$$\mathrm{cost}_{\mathrm{ren}}(x \to y) \coloneqq 1 - \mathrm{Sim}(x, y) \in [0, 1], \tag{13}$$

and set insertion/deletion costs to 1 *per node*. This calibration has two practical benefits. First, clipping ensures that renaming cost stays on the same scale as insertion/deletion (all in $[0, 1]$), which stabilizes normalization in US-NTED and avoids pathological cases where negative cosine would yield $\mathrm{cost}_{\mathrm{ren}} > 1$. Second, in the embedding space used, a negative cosine value does not reliably indicate semantic opposition for short topic labels. We therefore treat it as *no semantic match*, with $\mathrm{Sim} = 0$, rather than assigning an exaggerated penalty.

A.3 Unordered Semantic Tree Edit Distance (US-TED/US-NTED) Explanation

A.3.1 Definition

We compare the expert hierarchy T_h^* and the model hierarchy $\hat{T}_h$ using an *unordered* semantic tree edit distance, where sibling order is ignored and children are matched by minimum-cost assignment. Both hierarchies contain only internal-topic nodes; we exclude paper nodes. The leaves of T_h are **paper categories**, the terminal category nodes to which papers are assigned.

Let $\mathrm{Ch}(u)$ denote the multiset of children of node u, and let $|T_u|$ denote the number of nodes in the subtree rooted at u (including u). We use the calibrated edit costs from Section A.2: insertion and deletion cost 1 per node, and renaming $x \to y$ costs $\mathrm{cost}_{\mathrm{ren}}(x \to y) = 1 - \mathrm{Sim}(x, y) \in [0, 1]$.

We define the unordered node-to-node distance $D(u, v)$ recursively:

$$D(u, v) \;\coloneqq\; \mathrm{cost}_{\mathrm{ren}}(u \to v) \;+\; \mathrm{MatchCost}\big(\mathrm{Ch}(u), \mathrm{Ch}(v)\big). \tag{14}$$

If both u and v are leaves, then $\mathrm{MatchCost} = 0$ and $D(u, v)$ reduces to the renaming cost. If one node is a leaf and the other has children, the unmatched subtrees are charged by unit per-node insertion/deletion via $|T_u|$.

To define $\mathrm{MatchCost}$ when both nodes have children, let $\mathrm{Ch}(u) = \{u_1, \ldots, u_m\}$ and $\mathrm{Ch}(v) = \{v_1, \ldots, v_n\}$, and set $k = \max(m, n)$. We build a $k \times k$ cost matrix C by padding with dummy children:

$$C_{ij} = \begin{cases} D(u_i, v_j), & 1 \le i \le m,\ 1 \le j \le n, \\ |T_{u_i}|, & 1 \le i \le m,\ n < j \le k, \\ |T_{v_j}|, & m < i \le k,\ 1 \le j \le n, \\ 0, & m < i \le k,\ n < j \le k, \end{cases} \tag{15}$$

where matching a real child to a dummy child corresponds to deleting or inserting the entire subtree at unit cost per node, and dummy–dummy matches contribute zero. The unordered children matching cost is defined by the minimum assignment:

$$\mathrm{MatchCost}\big(\mathrm{Ch}(u), \mathrm{Ch}(v)\big) \;\coloneqq\; \min_{\sigma \in \mathfrak{S}_k} \sum_{i=1}^{k} C_{i, \sigma(i)}, \tag{16}$$

which we solve via the Hungarian algorithm.

Finally, letting r^* and $\hat{r}$ be the roots of T_h^* and $\hat{T}_h$, the tree-level distance is

$$\text{US-TED}(T_h^*, \hat{T}_h) \coloneqq D(r^*, \hat{r}). \tag{17}$$

We normalize by the total number of hierarchy nodes:

$$\text{US-NTED}(T_h^*, \hat{T}_h) := \frac{\text{US-TED}(T_h^*, \hat{T}_h)}{|T_h^*| + |\hat{T}_h|}. \tag{18}$$

A.3.2 Basic Properties

Range and normalization. **Proposition.** For any two hierarchies T_h^* and $\hat{T}_h$,

$$0 \leq \text{US-TED}(T_h^*, \hat{T}_h) \leq |T_h^*| + |\hat{T}_h|, \tag{19}$$

$$0 \leq \text{US-NTED}(T_h^*, \hat{T}_h) \leq 1. \tag{20}$$

Proof sketch. Non-negativity holds since all edit costs are non-negative. For the upper bound, consider a valid (not necessarily optimal) edit sequence that deletes every node in $\hat{T}_h$ (cost $|\hat{T}_h|$) and then inserts every node in T_h^* (cost $|T_h^*|$). Since US-TED is defined as a minimum edit cost under the same per-node insertion/deletion convention, it cannot exceed $|T_h^*| + |\hat{T}_h|$. Dividing by $|T_h^*| + |\hat{T}_h|$ yields $0 \leq \text{US-NTED} \leq 1$.

Permutation invariance. **Proposition.** US-TED is invariant to any permutation of sibling order in either tree. **Proof sketch.** At each pair of nodes (u, v), the children matching cost is defined by a minimum assignment over a cost matrix C whose rows/columns correspond to the children of u and v (plus dummies). Permuting sibling order permutes rows and/or columns of C but does not change the optimal assignment cost, hence leaves $\mathrm{MatchCost}$ and $D(u, v)$ unchanged.

Why unordered. Taxonomy siblings do not have a canonical left-to-right order. An ordered tree edit distance would penalize pure sibling permutations as structural errors and conflate sibling ordering with organization. By using unordered matching via minimum-cost assignment, US-TED focuses on structural organization (parent–child relations and subtree composition) rather than arbitrary sibling ordering.

Relation to unordered tree edit distance. We state the positioning of US-TED precisely. General unordered tree edit distance is NP-hard, so US-TED is not a computation of it. By matching children level by level under a bijection and recursing, US-TED is an instance of the classical *constrained* (equivalently, top-down / degree-two) tree edit distance family, for which the minimum-cost assignment at each node yields a polynomial-time algorithm; our contribution is not the matching scheme but the semantic cost calibration (clipped-cosine renaming against unit insert/delete) and its use as a taxonomy evaluation measure. Two consequences follow. First, US-TED upper-bounds the true unordered edit distance, since every constrained edit script is a valid edit script. Second, we make no claim that US-TED satisfies the triangle inequality: constrained edit distances are metrics under uniform costs, but our renaming cost $1 - \mathrm{Sim}$ derived from cosine similarity is not itself a metric, so US-TED should be read as a calibrated dissimilarity rather than a metric in the formal sense. This does not affect any reported result, all of which use US-TED only to compare a model tree against a fixed expert tree.

A.4 Semantic Path Similarity (Sem-Path)

A.4.1 Aligned-paper set and path extraction

Aligned-paper set. Sem-Path is computed on aligned paper pairs. Let D_a denote the set of aligned paper pairs $(d, \hat{d})$ between the expert tree T^* and the model tree $\hat{T}$. We construct D_a by normalizing paper titles and performing deterministic matching between expert papers and retrieved model papers (details of normalization rules and tie-breaking are reported in this section). This step is necessary to avoid inflated scores due to spurious title overlaps (false positives) or minor title variants (false negatives).

Ancestor-path extraction. For each aligned paper $d \in D_a$, we extract the root-to-leaf chain in both trees and keep only internal topic labels. Concretely, if the root-to-paper path in T^* is $(r^*, \ldots, u, d)$, we set $S_d = (r^*, \ldots, u)$, excluding the paper node d. Likewise, if the root-to-paper path in $\hat{T}$ is $(\hat{r}, \ldots, \hat{u}, \hat{d})$, we set $\hat{S}_d = (\hat{r}, \ldots, \hat{u})$.

A.4.2 DISTANCE FUNCTION AND MONOTONE ALIGNMENT

Clipped cosine distance. We use the clipped cosine distance $\delta(x, y) \coloneqq 1 - \mathrm{Sim}(x, y) \in [0, 1]$, which equals the renaming cost defined in Section A.2. This keeps the distance on the same scale as the unit unmatched-node penalty.

Order-preserving alignment. Given two ancestor-label sequences S_d and $\hat{S}_d$, we compute an order-preserving minimum-cost alignment that tolerates different granularities (i.e., different depths) while preserving relative order. Without loss of generality, let $A = (a_1, \ldots, a_p)$ be the shorter sequence and $B = (b_1, \ldots, b_q)$ be the longer sequence ($p \leq q$), where A, B correspond to $S_d, \hat{S}_d$ up to swapping. We compute a subsequence-style dynamic program:

$$\mathrm{dp}[0, j] = 0, \tag{21}$$

$$\mathrm{dp}[i, j] = \min\big(\mathrm{dp}[i-1, j-1] + \delta(a_i, b_j),\ \mathrm{dp}[i, j-1]\big). \tag{22}$$

Intuitively, $\mathrm{dp}[i, j]$ matches the first i labels of the shorter chain to an ordered subsequence within the first j labels of the longer chain.

Per-paper cost and score. We define the per-paper alignment cost by adding a penalty for each unmatched extra label in the longer chain:

$$J_d \coloneqq \mathrm{dp}[p, q] + \lambda\,(q - p), \tag{23}$$

where $\lambda \geq 0$ is a penalty coefficient controlling the cost of each unmatched extra node. In all experiments, we set $\lambda = 1$.

The final metric maps cost to similarity and averages over aligned papers:

$$\textsc{Sem-Path} \coloneqq \frac{1}{|D_a|} \sum_{d \in D_a} \frac{1}{1 + J_d}. \tag{24}$$

A.4.3 BASIC PROPERTIES AND DESIGN RATIONALE

Range. Since $J_d \geq 0$, we have $\mathrm{Sem\text{-}Path}(d) \in (0, 1]$ and therefore $\mathrm{Sem\text{-}Path} \in (0, 1]$.

Granularity tolerance with order sensitivity. SEM-PATH allows different hierarchy granularities because matching is performed between a shorter chain and an ordered subsequence of a longer chain. However, it is sensitive to mis-ordered or semantically inconsistent ancestor placement: if the model swaps high-level topics or inserts unrelated ancestors, the DP cost increases via $\delta(\cdot, \cdot)$ and/or the unmatched penalty $\lambda\,(q - p)$.

Relation to soft-cardinality baselines. Soft-cardinality metrics such as NSR/NSP primarily measure semantic coverage of a *set* of node labels and can be insensitive to hierarchical placement errors (e.g., correct topics but wrong parent–child relations). In contrast, US-TED/US-NTED and Sem-Path encode explicit structural constraints: US-TED penalizes topology and subtree-level edits, while SEM-PATH evaluates whether each aligned paper is placed under a semantically consistent ancestor chain. A detailed comparison, including failure modes and illustrative cases, is provided in App. A.6.

A.5 PAPER-TITLE NORMALIZATION AND ALIGNMENT

To compute retrieval scores and paper-conditioned metrics (e.g., SEM-PATH), we align model-returned papers with expert-cited papers. We do not rely on exact title match or a single identifier (e.g., DOI/arXiv ID), since identifiers can be missing and the same work may appear in multiple versions (arXiv vs. venue).

Normalization. We apply deterministic text normalization (lowercasing and whitespace/punctuation cleanup) to titles and other available textual fields.

Matching rule. For an expert paper p and a model paper $\hat{p}$, we compute a semantic similarity score $s(p, \hat{p}) \in [0, 1]$ on the normalized text representation. We align $(p, \hat{p})$ if either (i) $s(p, \hat{p}) = 1$, or (ii) $0.6 \leq s(p, \hat{p}) < 1$ and the normalized titles satisfy a strict containment check (one contains the other). Otherwise, we treat them as different papers. If multiple candidates match an expert paper, we keep the one with the highest $s(p, \hat{p})$.

A.6 Why Soft-Cardinality Baselines (NSR/NSP) Are Insufficient for Taxonomy Structure

What NSR/NSP measure. Node Soft Recall (NSR) and Node Soft Precision (NSP) are soft-cardinality extensions of set-based Recall/Precision (Fränti & Mariescu-Istodor, 2023). They quantify *semantic coverage* between two *collections of labels* by discounting near-duplicates via pairwise semantic similarity. Given two node-label lists $A = (a_1, \ldots, a_{|A|})$ and $B = (b_1, \ldots, b_{|B|})$, define the soft cardinality

$$c(A) \;\coloneqq\; \sum_{i=1}^{|A|} \frac{1}{\sum_{j=1}^{|A|} \mathrm{Sim}(a_i, a_j)}. \tag{25}$$

Following (Fränti & Mariescu-Istodor, 2023), define

$$\mathrm{NSR}(A, B) \coloneqq \frac{c(A) + c(B) - c(A \uplus B)}{c(A)}, \tag{26}$$

$$\mathrm{NSP}(A, B) \coloneqq \frac{c(A) + c(B) - c(A \uplus B)}{c(B)}, \tag{27}$$

$$\textsc{Soft F1}(A, B) \coloneqq \frac{2\,\mathrm{NSP}(A, B)\,\mathrm{NSR}(A, B)}{\mathrm{NSP}(A, B) + \mathrm{NSR}(A, B)}. \tag{28}$$

Here $A \uplus B$ denotes *multiset union with multiplicities*, implemented as list concatenation.

Our implementation. We instantiate Sim with embedding cosine similarity and clamp negatives:

$$\mathrm{Sim}(x, y) \;\coloneqq\; \max\big(0,\, \cos(\mathbf{e}(x), \mathbf{e}(y))\big), \tag{29}$$

where $\mathbf{e}(\cdot)$ is produced by a SentenceTransformer encoder (e.g., `all-MiniLM-L6-v2`) with ℓ_2-normalized embeddings. We collect *hierarchy* node labels by preorder traversal (including the root; excluding paper leaves); ordering is ignored and multiplicities are kept. We report the soft-cardinality baselines NSR, NSP, and Soft F1 in Table 5. They remain relatively high and compress model differences even when structure-aware metrics (e.g., US-NTED and Sem-Path) indicate substantial structural gaps.

Why label coverage cannot evaluate taxonomy structure. A taxonomy *structure* metric must depend on parent–child and ancestor relations, i.e., edge/ancestry errors should be penalized in an interpretable way aligned with hierarchical organization. By construction, NSR/NSP depend only on the label inventory and pairwise similarities and impose no edge or ancestry constraints; therefore, as suggested by Table 5, they primarily reflect semantic coverage/redundancy rather than hierarchy correctness. The following proposition formalizes this limitation.

Proposition (Structure-blindness under label-preserving rewiring). Let $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_n)$ be two *lists* of hierarchy-node labels, and assume B is a permutation of A (equivalently, A and B induce the same label multiset). Assume $\mathrm{Sim}(\cdot, \cdot)$ is deterministic, symmetric, and depends only on the label strings. Define soft cardinality by

$$c(A) \coloneqq \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} \mathrm{Sim}(a_i, a_j)}, \qquad c(B) \text{ analogously},$$

and assume $\sum_{j=1}^{n} \mathrm{Sim}(a_i, a_j) > 0$ for all i (e.g., when $\mathrm{Sim}(x, x) = 1$). Let $A \uplus B$ denote multiset union with multiplicities, implemented as list concatenation. Then $\mathrm{NSR}(A, B) = \mathrm{NSP}(A, B) = 1$, regardless of any differences in the parent–child relations of the underlying trees, as long as their hierarchy-node label multisets match.

Proof sketch. Because B is a permutation of A and Sim depends only on label strings, $c(A) = c(B)$. Consider $A \uplus B$ (concatenation). For any a_i,

$$\sum_{z \in A \uplus B} \mathrm{Sim}(a_i, z) = \sum_{j=1}^{n} \mathrm{Sim}(a_i, a_j) + \sum_{j=1}^{n} \mathrm{Sim}(a_i, b_j) = 2 \sum_{j=1}^{n} \mathrm{Sim}(a_i, a_j),$$

since $(b_1, \ldots, b_n)$ is a reordering of $(a_1, \ldots, a_n)$. Hence each occurrence of a_i in $A \uplus B$ has a denominator that is doubled, so each copy contributes half of its original term; with two copies, their contributions sum to the original. Summing over all labels yields $c(A \uplus B) = c(A)$. Substituting into the definitions of NSR and NSP gives $\mathrm{NSR}(A, B) = \mathrm{NSP}(A, B) = 1$. □

Counterexample 1. Consider two hierarchies with identical label multiset but different parent–child relations:

$$T_1: \; R\big(A(B, C),\, D(E, F)\big) \quad T_2: \; R\big(A(B, E),\, D(C, F)\big),$$

where the attachments of C and E are swapped. By the proposition, NSR/NSP attain their maxima when computed from the hierarchy label lists, despite a non-trivial structural change.

In contrast, structure-aware metrics penalize such rewiring. Under US-TED/US-NTED (App. A.3), a parent change is not a primitive move. It must be realized by deleting a (sub)tree and re-inserting it elsewhere. With insertion/deletion cost 1 per node, reattaching a subtree of size s has positive edit cost, and under unit insertion/deletion it is lower-bounded by deleting and re-inserting that subtree ($2s$) even if rename costs are 0; hence US-TED is strictly positive for this example.

SEM-PATH (App. A.4) is also affected: any paper under the swapped branches experiences an altered ancestor chain. For such a paper, at least one aligned ancestor label differs (e.g., matching C against E at the corresponding depth). Since the per-node distance is $1 - \max(0, \cos(\cdot, \cdot))$, the cumulative alignment cost is positive whenever $\mathrm{Sim}(C, E) < 1$, yielding a decreased path similarity $1/(1 + \mathrm{cost})$.

Counterexample 2. Soft cardinality is not monotone: adding semantically similar labels can decrease $c(\cdot)$ because row-sums increase. Consequently, NSR/NSP can exhibit non-intuitive scaling, including Recall-like values above 1, even under clamped similarities with $\mathrm{Sim} \in [0, 1]$.

Let $A = (a)$ and $B = (b_1, b_2)$ and assume (after clamping) that

$$\mathrm{Sim}(a, b_1) = \mathrm{Sim}(a, b_2) = 1, \qquad \mathrm{Sim}(b_1, b_2) = 0,$$

with $\mathrm{Sim}(x, x) = 1$. Using $A \uplus B$ as multiset union with multiplicities (concatenation), we obtain

$$\begin{aligned} c(A) &= 1, \\ c(B) &= \tfrac{1}{1+0} + \tfrac{1}{0+1} = 2, \\ c(A \uplus B) &= \tfrac{1}{3} + \tfrac{1}{2} + \tfrac{1}{2} = \tfrac{4}{3}, \\ \mathrm{NSR}(A, B) &= \big(c(A) + c(B) - c(A \uplus B)\big)/c(A) = \tfrac{5}{3} > 1, \\ \mathrm{NSP}(A, B) &= \big(c(A) + c(B) - c(A \uplus B)\big)/c(B) = \tfrac{5}{6}. \end{aligned}$$

We therefore interpret NSR/NSP as diagnostics of semantic overlap and redundancy under a particular soft-cardinality geometry, rather than stable Recall/Precision metrics for hierarchical *structure*.

Discussion. Because NSR/NSP operate on label inventories, they necessarily penalize any discrepancy in intermediate-node labels, even when the model produces a coarser yet hierarchically consistent taxonomy. For example, contracting an intermediate node changes the label multiset and reduces soft overlap. In contrast, US-NTED and SEM-PATH expose such differences as explicit, localized edit/alignment costs (node deletions/insertions in US-NTED; unmatched ancestor steps in SEM-PATH), which is more directly tied to structural operations on trees and ancestor chains.

Summary. NSR/NSP quantify soft semantic overlap between *label collections* and are useful auxiliary diagnostics for coverage and redundancy. However, they do not encode parent–child or ancestor-chain constraints and can exhibit non-intuitive scaling (including NSR> 1). Therefore, they are insufficient as primary measures of taxonomy *structure* quality. We report NSR/NSP only as auxiliary baselines/diagnostics, and rely on US-TED/US-NTED (global unordered semantic edit cost) and SEM-PATH (per-paper ancestor-chain consistency) as the main hierarchy-level evaluations.

Model	Hierarchy-Level Metrics		
	NSR↑	NSP↑	Soft F1↑
Deep Research mode			
o3	0.76	**0.92**	0.83
Doubao	0.82	0.80	0.81
DeepSeek	0.77	0.85	0.81
Gemini	0.88	0.77	0.82
Grok	0.79	0.87	0.83
Perplexity	0.88	0.80	**0.84**
Qwen	**0.91**	0.77	0.83
Bottom-Up mode			
Non-thinking-based			
Claude-4.5-Sonnet	0.87	0.85	**0.86**
GPT-5	0.82	0.85	0.84
Gemini-3-Pro*	0.87	0.83	0.85
DeepSeek-V3.2	0.82	**0.87**	0.85
Qwen3-Max*	0.80	0.86	0.83
Kimi-K2	**0.89**	0.82	0.85
Thinking-based			
Claude-4.5-Sonnet-Thinking	0.86	0.85	**0.85**
GPT-5-Thinking	0.83	0.85	0.84
Gemini-3-Pro-Thinking*	0.87	0.83	**0.85**
DeepSeek-V3.2-Thinking	0.82	**0.89**	**0.85**
Qwen3-Max-Thinking*	0.82	0.86	0.84
Kimi-K2-Thinking	**0.88**	0.82	**0.85**

Table 5: **Soft set-matching baselines** (NSR, NSP, Soft F1) for taxonomy evaluation.

A.7 ALGORITHMS

Algorithm 1 Semantic Path Alignment for SEMPATH Metric

Input: Ancestor-label sequence $S \in \mathbb{R}^{m\times 1}$, Ancestor-label sequence $\hat{S} \in \mathbb{R}^{n\times 1}$, $m \le n$
Parameter: Unmatched penalty $\lambda \ge 0$ (set $\lambda = 1$)
Initialize cost matrix $D \in \mathbb{R}^{(m+1)\times(n+1)}$ with ∞
$D[0,:] \leftarrow 0$
for $i = 1$ **to** m **do**
  for $j = i$ **to** n **do**
    $cost_{\text{match}} \leftarrow D[i-1, j-1] + \delta(S_i, \hat{S}_j)$
    $cost_{\text{skip}} \leftarrow D[i, j-1]$
    $D[i,j] \leftarrow \min(cost_{\text{match}}, cost_{\text{skip}})$
  end for
end for
Return $J(S, \hat{S}) \leftarrow D[m,n] + \lambda \cdot (n-m)$

Algorithm 2 Unordered Semantic Tree Edit Distance (US-TED)

Input: Tree nodes u, v; Embedder E; Costs c_{ins}, c_{del}
Procedure: TED(u, v)
$cost_{ren} \leftarrow 1 - \max(0, \cos(E(u), E(v)))$
if u, v are leaves **then**
  Return $cost_{ren}$
end if
Let $\{u_1, \ldots, u_n\} = \text{Ch}(u)$ and $\{v_1, \ldots, v_m\} = \text{Ch}(v)$
Construct cost matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$ where $N = \max(n, m)$:

$$\mathbf{C}_{i,j} \leftarrow \begin{cases} \text{TED}(u_i, v_j) & \text{if } i \le n, j \le m \\ |T_{u_i}| \cdot c_{del} & \text{if } i \le n, j > m \\ |T_{v_j}| \cdot c_{ins} & \text{if } i > n, j \le m \\ 0 & \text{otherwise} \end{cases}$$

Return $cost_{ren} + \min \sum_{i=1}^{N} \mathbf{C}_{i,\sigma(i)}$ {Hungarian Algorithm}

B METRIC CALIBRATION: FLOORS, CEILINGS, AND DEPTH-MATCHED SEM-PATH

A raw SEM-PATH value is uninterpretable without knowing what a degenerate taxonomy scores. Unlike ARI, SEM-PATH is *not* chance-corrected, so we calibrate it directly by scoring taxonomies whose quality we know by construction. Table 6 reports the result, and it changes how Finding 5 must be read.

Taxonomy	SEM-PATH	**Construction**
SINGLETON	26.00%	root → one category per paper
FLAT	27.49%	root → one category → every paper
RANDOM	28.11%	expert labels rewired into a random 3-level tree
12 configs, prev. gen.	*28.13–29.16%*	*Bottom-Up, Title+Abstract*
16 configs, current gen.	*28.12–30.03%*	*Bottom-Up, one generation later*
TRUNCATED	47.30%	expert tree collapsed to 3 levels, placement *correct*
Human baseline	*53.91%*	*two annotators, matched 10 surveys; 52.62 and 55.20%*
SHUFFLE	49.45%	expert tree, paper placement permuted at random
IDENTITY	100.00%	expert tree against itself

Table 6: **SEM-PATH calibration.** RANDOM and SHUFFLE are averaged over 3 seeds (± 0.02 and ± 0.27 pp). Two facts follow. (i) The model bands sit 0.6–2.5 pp above FLAT, a taxonomy performing no organization at all. (ii) Humans given the same Bottom-Up task clear the floor by 26.42 pp (App. K.4), while SHUFFLE destroys placement yet still scores 49.45%, just below the 50.56% at which two independent annotators agree with each other.

What the floors imply for Finding 5. The convergence of all 12 previous-generation configurations into a 1 pp band is largely a floor effect rather than a discovered property of models: in this shape regime SEM-PATH assigns 26–28% to essentially any taxonomy, so a band starting at 28.13% carries little information. We therefore do not treat the narrowness of the band as evidence of shared organizational bias.

A natural experiment confirms the floor without synthetic baselines. An objection to the table above is that SINGLETON, FLAT, RANDOM and SHUFFLE are constructions, so their scores might say more about our constructions than about the metric. App. I supplies the test that avoids this: re-running the benchmark on the newest model generation raises best ARI by 3.68 pp (31.24%→34.92%) while raw SEM-PATH moves 0.87 pp (29.16%→30.03%), remaining within 2.54 pp of FLAT. A leaf-partition alignment change large enough to be obvious on a chance-corrected metric is therefore almost invisible to raw SEM-PATH. This diagnoses sensitivity across runs; it is not an absolute organization-quality claim from ARI. We regard it as strong evidence because nothing about the comparison was designed by us.

What the ceilings imply for the human baseline. SHUFFLE retains expert labels and topology but randomizes the paper assigned to each leaf. It therefore removes the placement information that SEM-PATH is intended to reward. Nevertheless, it scores 49.45%, just below the 50.56% at which two independent annotators agree with each other. Raw SEM-PATH is therefore still sensitive to label and shape similarity. App. K.4 applies the depth-matched control to the human trees: the human–model gap shrinks from +24.64 pp to +13.27 pp but remains large on every matched survey, so shape alone does not explain the advantage.

Depth-matched SEM-PATH. The natural control is to score each model against a version of the expert tree collapsed to the model's own depth, so that the unmatched-ancestor penalty is not charged for a granularity choice. Table 8 reports this. Four things change.

Scores move well clear of the floor. Models rise from 28.5–29.0% to 38.3–42.3%, against a FLAT floor of 27.49%.

The matched human gap remains. On the same 10 surveys, humans average 54.09% versus 38.71% pooled across all 16 model configurations. Against the ARI-selected Kimi-K3 comparator (40.81%),

the paired advantage is $+13.27$ pp (95% CI $[10.69, 16.06]$; humans higher on 10/10). This known-group separation is our primary evidence that the variant retains meaningful organization headroom after controlling for depth.

Across-model variation is restored. The spread across the eight configurations widens from 0.49 pp to 4.00 pp, and the resulting order (Qwen3-Max $>$ Gemini-3-Pro $\approx$ GPT-5 $>$ Claude-4.5-Sonnet) matches the ARI ordering, which raw SEM-PATH did not. App. K.4 shows that ARI is saturated relative to human reproducibility, so this agreement cannot establish an absolute organization-quality gap. We use it only as auxiliary convergent evidence that the depth-matched variant recovers variation in leaf-partition alignment rather than only tree shape. Because the first ordering was observed on the same eight configurations used to motivate the control, we re-test it on the 16 newest-generation configurations of App. H, which played no part in its design: correlation with ARI rises from $r{=}0.294$ ($p{=}0.27$, not significant) to $r{=}0.813$ ($p{=}1{\times}10^{-4}$), and the spread from 1.91 to 9.45 pp. Rank correlation improves more modestly ($\rho{=}0.368 \rightarrow 0.476$, $p{=}0.06$), so the relationship is strong in magnitude while the exact ordering still shifts locally; we report both rather than only the favorable one.

Interpreting the residual gap. A tempting comparison is to subtract the depth-matched model scores from the 47.30% of TRUNCATED and read the difference as placement error. That comparison is invalid: TRUNCATED is scored against the *full* expert tree, whereas depth-matched model scores are scored against the *collapsed* tree, so the two live in different reference frames and their difference is not interpretable. In the depth-matched frame the ceiling is 100% by construction.

To calibrate the variant properly, we recompute the floors inside its own frame in Table 7. We collapse the expert tree and randomize the paper assigned to each leaf, which destroys placement information while retaining expert labels and topology. This baseline still scores **77.29%**. Placement error alone is therefore worth at most about 23 pp, while models sit at 38–42%, some 35 pp *below* even a randomly-placed expert-labeled tree. The dominant term in the model deficit is thus divergence of category *vocabulary and topology*, not misplacement of papers: models do not put the right papers in the wrong expert categories, they build a different set of categories. This is consistent with the raw-frame pathology that SHUFFLE (49.43% in our recomputation, 49.45% as reported) exceeds TRUNCATED, and it means SEM-PATH remains label-dominated even after depth-matching. We therefore make no quantitative placement-error claim, and we flag separating label divergence from placement error as a second open problem alongside chance correction.

Reference frame	**Candidate**	SEM-PATH
Full expert tree	collapsed to 3 levels, placement *correct* (TRUNCATED)	47.30%
	full tree, placement *random* (SHUFFLE)	49.43%
Collapsed expert tree	itself (definitional ceiling)	100.00%
	itself with placement *random*	**77.29%**
	model trees (Table 8)	*38–42%*

Table 7: **Calibrating the depth-matched variant in its own reference frame**, averaged over 3 seeds and all 72 surveys. The top block reproduces the published raw-frame values and confirms the implementation. The bottom block tests the placement claim directly. Randomizing placement while keeping expert labels costs only 100%→77.29%, yet models score 38–42%, so the model deficit is dominated by label and topology divergence rather than by placement.

Recommendation. We report raw SEM-PATH for comparability with Table 4 but recommend that future work using TAXOBENCH report the depth-matched variant alongside it, together with the FLAT floor for the survey set in question. The underlying issue is that SEM-PATH lacks the chance correction that makes ARI interpretable; a properly chance-corrected hierarchy metric is the natural next step, and we regard it as the most valuable methodological extension of this work.

Model	vs. full expert	vs. depth-matched expert	Δ
Claude-4.5-Sonnet	28.86%	38.34%	+9.49
Claude-4.5-Sonnet-Thinking	28.51%	38.33%	+9.82
GPT-5	28.92%	41.09%	+12.17
GPT-5-Thinking	28.99%	40.61%	+11.61
Gemini-3-Pro	28.77%	41.10%	+12.33
Gemini-3-Pro-Thinking	28.80%	40.87%	+12.06
Qwen3-Max	28.91%	**42.32%**	+13.41
Qwen3-Max-Thinking	28.58%	41.66%	+13.07
Spread across models	**0.49 pp**	**4.00 pp**	
Ceiling in this frame (collapsed tree vs. itself)		*100.00%*	
Placement-random floor (expert labels, papers permuted)		*77.29%*	

Table 8: **Depth-matched SEM-PATH** on the previous generation (Bottom-Up, Title+Abstract; the 8 configurations with retained tree outputs covering 66–71 of the 72 surveys). Each model is additionally scored against the expert tree collapsed to 3 category levels, which removes the unmatched-ancestor penalty arising purely from depth. The two reference rows are computed in the same collapsed frame as the model column; models fall well below the placement-random floor, so the shortfall is not attributable to placement (Table 7).

C CONSTRUCT VALIDITY DECOMPOSITION

This appendix supports the alignment-vs-capability framing introduced in Section 3.4 and used throughout the Findings. Table 9 explicitly maps every numerical claim in the paper to one of two groups so that readers can see at a glance which conclusions depend on the expert reference and which do not.

Finding / Number	Level	What it consumes
Best Recall 20.92% (F1)	L3	$\mathcal{P}^*$ only; no taxonomy
Depth deficit 0.89–1.87 levels (Finding 6)	L2	Expert depth (a scalar); no judge
Branching 3.16 vs. 2.52	L2	Expert branching (a scalar)
Over-segmentation, #cat. and singletons	L2	Expert category count (a scalar)
Best ARI 34.92% (*saturated*; alignment only)	L1	Full expert assignment U^*
SEM-PATH band 28–30% (Finding 5)	L1	Full expert hierarchy T_h^*
US-TED/US-NTED rankings	L1	Full expert hierarchy T_h^*
Human baseline 53.91% SEM-PATH (10 surveys)	L1	Same reference as models
Constrained-K +7.49 pp ARI	L1	Full expert assignment U^*
Sibling overlap 75.9%	*unusable*	Judge; expert scores 80.6%
Imbalance 83.4%	*weak*	Judge; rubric cites reference tree
MECE 51.2%, criteria 66.0%	*weak*	Judge; no expert control
Missing branches 77.5%	*weak*	Judge + domain knowledge

Table 9: **How much of the expert reference each finding consumes**, using the L1–L3 grading of Section 3.4. No finding in this paper is reference-free. L3 uses the expert paper set but no taxonomy; L2 uses only scalar shape statistics of the expert hierarchy, never its labels or topology; L1 uses the hierarchy or assignment in full. ARI is marked saturated because it exceeds human–human agreement and is not used to infer organization quality. The four LLM-judged defect rates at the bottom are listed for completeness but are not load-bearing: sibling overlap fails its expert control outright, and the other three lack one.

D REFERENCE-LIGHT STRUCTURAL EVIDENCE

This appendix reports the reference-light evidence behind Finding 6. We separate two sources of evidence: *structural statistics*, which are computed directly from trees and require no judge, and

LLM-judged defect rates, which are rubric scores produced by GPT-4o. We report both, and we run the expert taxonomies through the same procedures so that every rate has a control.

Structural statistics (no judge, no threshold). Table 10 compares the 72 expert taxonomies against model taxonomies on four statistics computed directly from the trees. Hierarchy depth counts category levels including the root and excluding paper nodes, matching the convention in Table 1. One pattern is universal and one is not.

Depth is the most durable deficit. We pool **70 Bottom-Up runs** across four input granularities, two model generations, and both reasoning settings. They give a depth of 2.99–3.97 levels against the expert's 4.86, a deficit of **0.89–1.87 levels with zero exceptions**; per-paper ancestor chains are short by 1.18–1.48 labels. Over-segmentation is model-dependent by contrast: category counts span 11.0–28.3 around the expert's 14.0 and singleton rates 4.9–50.4% around the expert's 17.0%, so some models under-segment while others more than double the expert rate. App. F shows the depth deficit is nevertheless not a capability ceiling, and that removing it does not improve alignment.

We derive the metric consequence directly. For an aligned paper d with model and expert ancestor-chain lengths p and q, $J_d = \mathrm{dp}[p, q] + \lambda|q - p| \geq \lambda|q - p|$ because the alignment cost dp is non-negative. Hence

$$\textsc{Sem-Path} \leq \frac{1}{|D_a|} \sum_{d \in D_a} \frac{1}{1 + \lambda\,|q_d - p_d|} =: \textsc{Ceiling}, \tag{30}$$

which must be averaged *per paper*. Substituting the mean deficit into $1/(1 + x)$ instead would understate the ceiling, since $1/(1 + x)$ is convex and Jensen's inequality gives $\frac{1}{|D_a|} \sum_d 1/(1 + |q_d - p_d|) \geq 1/(1 + \overline{|q - p|})$; on our data the two differ by 6 pp, so we report only the per-paper quantity.

We evaluate Equation 30 over the 3,371–3,757 papers aligned per model and obtain a ceiling of **45.8–52.4%**, with mean 48.6%. The depth deficit alone costs models roughly 51 pp of Sem-Path. Label semantics cost another 20 pp because models score 28–29% rather than the 48.6% attainable with perfect label matching at their current depth. Depth therefore explains a large but incomplete part of the gap. Human annotators under the same Bottom-Up requirements score 53.91% Sem-Path at depth 4.65 on a matched 10-survey subset in App. K.4, above the shallow-tree ceiling that models can attain even with perfect label matching.

Over-segmentation is real but model-specific. Unlike depth, this defect does not generalize: across the 70 pooled runs category counts span 11.0–28.3 around the expert's 14.0 and singleton rates 4.9–50.4% around the expert's 17.0%. In the previous generation Claude-4.5-Sonnet and the two GPT-5 variants emit 20.7–23.1 categories with 33–36% singletons while Gemini-3-Pro and Qwen3-Max match the expert count closely; in the current generation the same split appears between GPT-5.6-sol (28.3 categories, 43.5% singletons) and Gemini-3.1-Pro (11.0, 5.1%). Some models therefore over-segment by more than $2\times$ the expert rate while others under-segment, which is why Finding 6 treats depth and over-segmentation separately.

Is the depth deficit an artifact of our prompt? We take this alternative explanation seriously, because the worked example embedded in all three of our generation prompts (App. O.1) is itself a three-level tree, and models emit 3.0–3.6 levels. Existing data cannot fully separate the two explanations: all three input-granularity conditions share the *same* exemplar, so input granularity provides no variation in the anchor. We report what evidence we do have and then name the experiment that would settle it.

The Constrained-K probe in App. G.2 provides a partial control by fixing the number of paper categories. Models follow this instruction: Claude-4.5-Sonnet drops from 19.8 to 13.6 categories and GPT-5 from 17.4 to 10.8, bracketing the expert value of 14.0. However, depth does not approach the expert value of 4.86. It falls from 3.40 to 3.10 for Claude-4.5-Sonnet and from 3.24 to 3.12 for GPT-5 on paired sets of 20 and 17 surveys. This probe separates depth from category count but does not test a depth-specific instruction because both conditions retain the three-level exemplar. We address that question with the depth-controlled probe in App. F.

Why a flatter tree is a worse tree. Since the depth comparison consumes a scalar property of the expert reference (L2 in Section 3.4), it is fair to ask why shallower should count as worse rather than

Taxonomy source	Depth	Chain	Deficit	Ceiling	# Cat.	Singleton %
Expert reference	**4.86**	**4.46**	—	—	**14.0**	**17.0**
Claude-4.5-Sonnet	3.62	3.28	1.18	52.4%	23.1	32.7
Claude-4.5-Sonnet-Thinking	3.58	3.22	1.24	51.5%	18.9	21.0
GPT-5	3.23	3.09	1.37	49.2%	20.7	35.9
GPT-5-Thinking	3.24	3.13	1.33	49.2%	21.2	32.9
Gemini-3-Pro	3.04	3.00	1.46	47.0%	13.8	7.1
Gemini-3-Pro-Thinking	3.06	3.02	1.44	47.5%	13.7	7.3
Qwen3-Max	3.01	2.98	1.48	46.6%	12.9	12.0
Qwen3-Max-Thinking	3.03	2.99	1.47	45.8%	13.3	12.3
Model mean	3.23	3.09	1.37	**48.6%**	17.2	20.1
Δ **(model − expert)**	-1.63	-1.37	—	—	$+3.2$	$+3.2$

Table 10: **Expert-paired structural statistics and the induced SEM-PATH ceiling** (Bottom-Up, Title+Abstract; the 8 configurations whose tree outputs were fully retained). Nothing here uses an LLM judge or a tunable threshold. **Depth** counts category levels including the root and excluding paper nodes, matching Table 1; **Chain** is the mean per-paper ancestor-chain length, which is the quantity SEM-PATH actually operates on; **Deficit** is the mean of $q_d - p_d$ over aligned papers; **Ceiling** is Equation 30, averaged per paper. A singleton is a paper category holding exactly one paper. The depth deficit holds for every configuration without exception; the category-count and singleton gaps hold only for Claude-4.5-Sonnet and the two GPT-5 variants.

merely different. Two reasons, neither of which needs the expert's labels. First, model trees are not only shallower but correspondingly *wider*: mean branching factor is 3.16 across models against 2.52 for expert taxonomies, so each parent carries more siblings. Mutual exclusivity is harder to maintain among more siblings, which links the shape statistic to the MECE property the taxonomy is supposed to have. Second, a tree with 17.2 leaf categories over 3.2 levels performs little progressive refinement: its intermediate layer separates the root from the leaves by roughly one meaningful decision, whereas the expert's 4.86 levels encode a chain of successive distinctions that a reader can follow. We regard this as an argument, not a measurement, and it is why a reader study connecting structure to utility (App. D.1) remains the missing piece.

LLM-judged defect rates, with an expert control. Table 11 reports the five defect rates over 1,000 model taxonomies together with the rate the same procedure assigns to the 72 expert taxonomies. All five are produced by the GPT-4o rubric of App. O.3 rather than by standalone programmatic detectors, and the rubric's imbalance criterion is phrased relative to the reference tree. These rates are therefore not independent of the judging pipeline. Moreover, expert taxonomies exhibit comparable rates of sibling overlap and structural imbalance under an embedding re-implementation of the same criteria, because sibling categories in a real survey are semantically related by construction. Sibling overlap and imbalance therefore do not discriminate model trees from expert trees and cannot carry the weight of a capability claim; the depth and over-segmentation statistics above can, because they are paired against the expert reference and computed without a judge.

We retain the 1,000 annotated model taxonomies in the private reproducibility archive. The public repository releases the dataset and scoring code but excludes raw model outputs and product logs.

D.1 FROM STRUCTURAL DEFECTS TO DOWNSTREAM SURVEY QUALITY

TAXOBENCH scores taxonomy structure, not the readability of a finished survey. We do not measure downstream utility directly, and we flag that as an open gap. We do, however, want to be explicit about why the defects we measure are not merely aesthetic, since the taxonomy is the section outline of the generated survey rather than an auxiliary artifact.

MECE violations propagate into content redundancy. When two sibling categories overlap semantically, the outline they induce forces the same subtopic to be narrated under more than one heading. The resulting survey repeats material across sections and gives the reader no principled basis for deciding which section covers what.

Defect Category	Model	Expert	Basis
Semantic overlap between siblings	75.9%	80.6%	GPT-4o rubric; cosine re-impl. at τ=0.6
Structural imbalance	83.4%	63.9%	GPT-4o rubric (criterion cites reference tree)
MECE violations	51.2%	—	GPT-4o rubric
Inconsistent classification criteria	66.0%	—	GPT-4o rubric
Missing core branches	77.5%	—	GPT-4o rubric + domain knowledge

Table 11: **LLM-judged defect rates with an expert control.** All rates come from the GPT-4o rubric. It reaches $\kappa = 0.89$ against human raters on the 5-point scale in App. J, which validates rubric scores rather than binary defect detection. The **Expert** column applies an embedding re-implementation of the same criterion to the 72 expert taxonomies. Expert trees show higher sibling overlap than model trees, so this category does not discriminate. A dash marks criteria that we could not re-implement without a judge.

Over-segmentation propagates into broken navigation. A taxonomy in which a third of the leaf categories hold a single paper, as in App. D, produces non-generalizing micro-headings. Readers then have difficulty forming a map of the field and locating a line of work. Sentence-level writing quality cannot repair this structural problem.

We therefore position taxonomy quality as a plausible *necessary* condition for survey utility rather than a demonstrated cause of it. Establishing the causal link requires a reader study in which survey text is generated from taxonomies of controlled quality and evaluated for redundancy and navigability; we regard this as the most valuable extension of TAXOBENCH and leave it to future work.

E INTERSECTION-CONDITIONED ARI$_\cap$ AND V-MEASURE$_\cap$

This appendix specifies the intersection-conditioned metrics used in Table 3. In Deep Research mode, an agent organizes only the papers it retrieves, so we compute a conditioned view on $\mathcal{P}^* \cap \hat{\mathcal{P}}$ in addition to the end-to-end view that assigns every unretrieved expert paper to a sink label. We align expert and retrieved papers with the deterministic title-normalization rules of App. A.5, then compute ARI$_\cap$ and V-Measure$_\cap$ on the aligned subset.

Two reporting choices are important. First, ARI is degenerate for a single point, so the intersection-conditioned columns are averaged only over surveys with at least two aligned papers; this leaves 29–67 surveys per agent. Second, the size of the intersection is itself a confound: Doubao averages only 1.7 retrieved expert papers and obtains the highest ARI$_\cap$, whereas o3 averages 11.1 retrieved expert papers and scores lower. We therefore report $|\mathcal{P}^* \cap \hat{\mathcal{P}}|$ in the table and use ARI$_\cap$ only to show that retrieval dominates the end-to-end score, not to rank organization skill.

F DEPTH-CONTROLLED PROBE: IS THE DEPTH DEFICIT CAUSED BY OUR PROMPT?

We measure the depth deficit in Finding 6 with prompts whose only worked example is a three-level tree. The exemplar could therefore cause the shallow output and create artificial benchmark headroom. We test that possibility directly.

Design. We re-run Bottom-Up under three conditions that differ *only* in the depth cue; the prompt is edited in place on the released input, so every other byte is identical.

- **A (control)**: the shipped prompt, three-level worked example.
- **B (exemplar)**: same prompt, worked example replaced by a five-level tree.
- **C (instruction)**: three-level example plus an explicit target of about five levels constraint.

Four models × 3 conditions × all 72 surveys = 864 calls. Condition A reproduces the 3.01–3.41 range reported in the main text, confirming the apparatus.

Model	Cond.	Depth	#Cat	Singleton	SEM-PATH	ARI	Hom.
GPT-5	A	3.18	19.5	32.7%	28.67%	29.02%	78.30%
GPT-5	B	3.38	20.4	33.9%	28.96%	29.30%	79.54%
GPT-5	**C**	**4.86**	**35.1**	**66.7%**	**25.90%**	**14.65%**	91.55%
Claude-4.5-Sonnet	A	3.25	19.2	24.4%	29.05%	30.39%	79.65%
Claude-4.5-Sonnet	B	4.01	27.2	41.2%	29.17%	23.08%	87.15%
Claude-4.5-Sonnet	C	4.08	28.6	44.6%	28.81%	20.12%	87.79%
DeepSeek-V3.2	A	3.41	16.7	19.1%	28.70%	27.45%	73.64%
DeepSeek-V3.2	B	3.90	21.1	27.5%	28.38%	23.34%	79.26%
DeepSeek-V3.2	C	4.00	20.2	24.4%	28.19%	24.57%	78.70%
Qwen3-Max	A	3.01	14.5	15.7%	29.01%	30.79%	71.70%
Qwen3-Max	B	3.18	15.2	16.7%	29.30%	30.04%	72.40%
Qwen3-Max	C	3.11	13.5	9.7%	29.24%	29.63%	70.45%
Expert reference	—	*4.86*	*14.0*	*17.0%*	—	—	—

Table 12: **Depth-controlled probe** with 864 calls and 857 parseable trees. Condition C makes GPT-5 match the expert's depth exactly but costs it 2.77 pp SEM-PATH and half its ARI, because the extra levels are produced by fragmentation. Rising homogeneity under C is a by-product of splitting (smaller clusters are purer), not an improvement.

Result 1: exemplar shape matters little; explicit instruction works. Table 12 reports the three conditions. Swapping in a five-level exemplar (B) moves depth by only $+0.17$ to $+0.76$ levels. An explicit target (C) is far more effective: GPT-5 reaches 4.86, exactly the expert mean, with Claude at 4.08 and DeepSeek at 4.00. Three of four models can therefore build five-level taxonomies and simply do not do so unprompted. Finding 6 is accordingly stated as a behavioral regularity, not a capability ceiling. Qwen3-Max is the exception, barely responding to either manipulation ($3.01 \rightarrow 3.11$); for that model the shallow output does look like a limit.

Result 2: depth compliance does not buy expert-like structure. This is the substantive point. At matched depth GPT-5's category count rises from 19.5 to 35.1 ($2.5\times$ the expert's 14.0) and its singleton-leaf rate doubles from 32.7% to 66.7% (expert: 17.0%), while SEM-PATH *falls* 2.77 pp and ARI halves. Claude and DeepSeek move the same direction. Across all 12 model–condition pairs, none improves SEM-PATH by more than 0.3 pp.

What this means for the benchmark. The probe rejects the explanation that our shallow exemplar creates the benchmark headroom. Models can increase depth when instructed, but every alignment measure becomes worse. We therefore interpret the standard depth deficit as a failure to build coherent hierarchical refinement; when forced, added levels fragment the taxonomy rather than repair it. The gap reported by TAXOBENCH is not an artifact of the prompt. The depth-based bound on SEM-PATH (Equation 30) is unaffected, since it is derived under the standard prompt where the deficit is universal.

G DECOMPOSITION ANALYSIS AND MECHANISTIC PROBE

G.1 DECOMPOSITION BY DEPTH AND SIZE

We decompose Bottom-Up results by paper count and hierarchy depth in Table 13 to separate scale from deep hierarchical reasoning. US-TED grows monotonically with paper count, so larger paper sets increase structural divergence. SEM-PATH falls sharply with depth. For Qwen3-Max, it drops from 43.5 to 34.7 to 25.6% as depth increases from 3 to at least 5. This pattern localizes the bottleneck to deep hierarchy rather than shallow categorization.

Model	US-TED (↓)			SEM-PATH (↑)		
	< 40	40–80	> 80	d=3	d=4	$d \geq 5$
Claude-4.5	30.3	45.9	59.3	37.5	35.2	25.8
GPT-5	28.7	44.8	50.5	41.9	33.8	26.0
Gemini-3	24.2	34.7	40.0	39.8	33.6	25.0
Qwen3-Max	25.5	35.3	37.9	43.5	34.7	25.6
Kimi-K2	32.5	44.8	44.8	42.8	33.5	25.2

Table 13: **Depth and size decomposition** (Bottom-Up). Each cell is the average over the corresponding bin.

G.2 CONSTRAINED-K MECHANISTIC PROBE

To causally test whether granularity selection drives the over-segmentation pattern observed in Finding 4, we ran a Constrained-K experiment on 15 sampled surveys: GPT-5 was prompted to produce exactly K paper categories (where K is the expert category count). This isolates granularity from semantic-assignment ability.

Two effects are observed (Table 14): (i) ARI improves substantially (13.05% $\rightarrow$ 20.54%, +7.49 pp), confirming that granularity selection is a major source of error; (ii) however, ARI plateaus at 20.54%, far below the human ceiling, and homogeneity *drops* (73.84% $\rightarrow$ 62.87%), revealing that the previously high homogeneity was largely the by-product of over-segmentation rather than genuine semantic understanding.

Metric	Free-K	Constrained-K
ARI	13.05%	**20.54%** (+7.49 pp)
Homogeneity	73.84%	62.87% (-10.97 pp)

Table 14: **Constrained-K probe** (15 surveys, GPT-5). Forcing the model to use exactly K paper categories (the expert count) raises ARI by 7.49 pp but lowers homogeneity, showing both granularity and semantic assignment as bottlenecks.

We also report the average paper-category counts produced under the three input settings (Bottom-Up): richer input causes models to produce more fine-grained categories, amplifying over-segmentation (Table 15).

Model	Title+Abs	+Core-task	+Summary
Claude-4.5	22.1	24.4	25.3
DeepSeek-V3.2	16.9	23.7	21.1
Gemini-3-Pro	12.8	14.3	14.0

Table 15: **Average paper categories** per input setting. Richer input increases the number of fine-grained categories, consistent with the over-segmentation explanation.

G.3 METRIC–HUMAN CORRELATION

We compute the Spearman correlation between SEM-PATH and the LLM-as-Judge scores, which are themselves validated against human evaluators with Cohen's $\kappa = 0.89$; details are in App. J. In Deep Research mode, where the 7 agents exhibit meaningful variance in performance, SEM-PATH is strongly correlated with the average judge score (Table 16). In Bottom-Up mode, all 12 models converge into a 1 pp band on SEM-PATH, so rank-based correlations become unstable with only 6 unique data points; this instability is itself diagnostic of the convergence in Finding 5 rather than a metric artifact.

	Cov.	Org.	Logic	Topo.	Avg
SEM-PATH	0.949	0.632	0.774	0.964	**1.000**

Table 16: **Spearman correlation** of SEM-PATH with LLM-as-Judge scores (Deep Research mode, n=7).

H REASONING-EFFORT CONTROL

We run every model in Table 4 twice, at the lowest and highest reasoning effort accepted by its endpoint. Because this control is not portable across vendors, we report in Table 17 the parameters sent and tokens returned instead of labeling the rows as thinking and non-thinking.

Model	**Endpoint**	`reasoning_effort` sent		mean reasoning tokens	
		low	high	low	high
Claude-Sonnet-5	`claude-sonnet-5`	`minimal`	`high`	n/a	n/a
GPT-5.5	`gpt-5.5`	`none`	`high`	0	4,091
GPT-5.6-sol	`gpt-5.6-sol`	`none`	`high`	0	2,541
Gemini-3.1-Pro	`gemini-3.1-pro-preview`	`minimal`	`high`	2,717	7,771
DeepSeek-V4-Pro	`deepseek-v4-pro`	`minimal`	`high`	10,182	10,164
Qwen3.8-Max	`qwen3.8-max-preview`	`minimal`	`high`	3,791	12,797
Kimi-K3	`kimi-k3`	`minimal`	`high`	1,190	4,187
Grok-4	`grok-4-latest`	`minimal`	`high`	970	924

Table 17: **Reasoning-effort settings, as sent and as observed.** The low setting is the weakest value each endpoint accepts, found by probing: GPT accepts `none` while the others reject it and require `minimal`. `max_tokens` is 65,536 throughout, temperature is left at the endpoint default, and the prompt is byte-identical across both rows. Three cases show why the parameter alone is not evidence of a contrast: `minimal` does *not* disable reasoning (Gemini still spends 2,717 tokens); DeepSeek-V4-Pro appears to reason at capacity regardless of the setting; and Claude-Sonnet-5 reports no reasoning tokens at all, so its contrast is unverified.

Why we log tokens rather than trust the flag. The previous-generation Gemini-3-Pro pair in App. I produced byte-identical outputs on all 72 surveys because the endpoint accepted a thinking flag and ignored it. That failure is invisible unless the returned reasoning-token count is recorded, so every call in Table 4 logs both the parameter sent and the tokens returned, and we mark the one configuration (Claude-Sonnet-5) where the endpoint does not report them.

Two practical consequences. First, the lowest effort setting does not always disable reasoning. Our settings bracket a range rather than comparing reasoning with its absence; a genuine ablation requires a hard off switch. Second, effort settings are not comparable across vendors. The `high` setting uses 924 tokens on Grok-4 and 12,797 on Qwen3.8-Max, so we compare each model only with itself.

Results at the highest setting. Table 4 reports the lowest-effort rows; Table 18 gives the matching highest-effort rows on the same 72 surveys and the same prompts. Of the 8 models run, 3 gain ARI at the higher setting (Gemini-3.1-Pro $+2.57$, DeepSeek-V4-Pro $+2.09$, GPT-5.5 $+0.83$) and 5 lose (Grok-4 -0.17, GPT-5.6-sol -0.32, Qwen3.8-Max -1.36, Kimi-K3 -1.49, Claude-Sonnet-5 -1.65). Neither the depth deficit nor the SEM-PATH floor is affected: depth stays within 2.99–3.97 and SEM-PATH within 28.12–30.03% across all 16 configurations.

A newer tier is not automatically better here. GPT-5.5 outscores its successor GPT-5.6-sol by 6.3 pp ARI (28.30% vs. 21.98% at lowest effort) while emitting 22.1 categories against 28.3 and 32.4% singleton leaves against 43.5%. The two tiers are otherwise the same family and the same prompt, so this is a clean illustration of the mechanism in Finding 6: GPT-5.6-sol builds the deepest trees of any model we tested (3.97 levels at high effort, closest to the expert's 4.86) and is simultaneously the worst on ARI, because it reaches that depth by fragmenting. Depth alone is not progress.

Model (effort `high`)	ARI↑	Hom.↑	Comp.↑	US-NTED↓	SEM-PATH↑	Depth	#Cat	Singl.
Claude-Sonnet-5†	**33.27%**	74.99%	**69.14%**	76.08%	**29.51%**	3.31	16.3	22.3%
Gemini-3.1-Pro	32.24%	70.93%	68.30%	77.60%	28.80%	3.04	**13.3**	**5.6%**
DeepSeek-V4-Pro	31.91%	76.22%	68.16%	77.45%	28.76%	3.28	18.3	28.2%
Kimi-K3	31.73%	77.86%	67.85%	78.75%	28.78%	3.10	17.8	14.6%
Grok-4	31.27%	70.60%	68.65%	78.31%	28.12%	3.07	13.5	11.6%
Qwen3.8-Max	28.45%	77.13%	66.23%	78.85%	28.66%	3.17	18.6	17.5%
GPT-5.6-sol	21.66%	**86.18%**	64.91%	**73.81%**	28.74%	**3.97**	27.6	38.8%
Within-family control: the preceding GPT tier								
GPT-5.5, effort `none`	28.30%	82.69%	66.96%	79.32%	28.96%	3.22	22.1	32.4%
GPT-5.5, effort `high`	29.13%	81.56%	67.03%	77.38%	28.77%	3.35	20.9	23.0%
Expert reference	—	—	—	—	—	*4.86*	*14.0*	*17.0%*

Table 18: **Highest reasoning effort**, same prompts and surveys as Table 4 (n=72, except Grok-4 at 71), plus both settings of the preceding GPT tier. †Claude-Sonnet-5 reports zero reasoning tokens in both settings, so this is the one pair we could *not* verify as two distinct conditions and it should not be counted as evidence about reasoning.

I PREVIOUS-GENERATION CONTROL

Table 4 reports the newest model in each family. We additionally evaluate the preceding generation as a longitudinal control of whether the metrics respond to model improvement.

Model	**Leaf-Level**				**Hierarchy-Level**		
	ARI↑	Hom.↑	Comp.↑	V-Meas.↑	US-TED↓	US-NTED↓	SEM-PATH↑
Non-thinking-based							
Claude-4.5-Sonnet	27.25%	**82.27%**	66.61%	**72.84%**	42.89	77.73%	**29.16%**
GPT-5	28.17%	77.79%	66.45%	70.99%	40.84	78.65%	28.97%
Gemini-3-Pro*†	29.86%	68.25%	67.15%	67.00%	**32.32**	76.35%	28.13%
DeepSeek-V3.2	27.15%	74.33%	66.15%	69.28%	37.04	**75.71%**	28.63%
Qwen3-Max*	**31.24%**	68.48%	**68.63%**	67.52%	32.80	78.17%	29.00%
Kimi-K2	23.69%	79.32%	64.88%	70.48%	41.35	79.58%	28.33%
Thinking-based							
Claude-4.5-Sonnet-Thinking	29.28%	76.35%	66.96%	70.58%	37.87	76.56%	28.66%
GPT-5-Thinking	25.97%	**80.89%**	66.05%	**72.07%**	43.05	79.86%	28.56%
Gemini-3-Pro-Thinking*†	28.84%	67.26%	67.00%	66.45%	**32.40**	76.47%	28.21%
DeepSeek-V3.2-Thinking	27.71%	70.74%	67.14%	67.68%	34.51	**76.02%**	28.67%
Qwen3-Max-Thinking*	**30.32%**	68.70%	**68.10%**	67.38%	33.62	78.24%	**28.95%**
Kimi-K2-Thinking	22.59%	78.98%	64.64%	70.18%	44.07	80.85%	28.21%

Table 19: **Previous generation**, Bottom-Up, Title+Abstract. * denotes Preview versions. †**The two Gemini-3-Pro rows are not two distinct conditions**: their outputs are byte-identical on all 72 surveys because the endpoint silently ignored the thinking flag, so any difference between these rows is not attributable to reasoning mode. The successor-model comparison uses per-call token verification (App. H); the three other families differ on 72/72 surveys.

The depth deficit survives a model generation. Depth spans 2.99–3.97 in the current generation against the expert's 4.86; the deepest configuration is still 0.89 levels short and none of the 16 reaches parity. Pooling with the 54 previous-generation runs gives 70 runs across four input granularities and two generations with no exception. This is the most durable result in the paper: a year of model progress did not move it.

The SEM-PATH floor also persists across generations. Best ARI rises from 31.24% (Qwen3-Max, Table 19) to 34.92% (Claude-Sonnet-5), a 3.68 pp gain in expert-reference leaf-partition alignment. Over the same comparison raw SEM-PATH moves 0.87 pp (29.16%→30.03%) and stays within

2.54 pp of the FLAT floor. This independently measured alignment change is thus nearly invisible to the raw metric. Unlike the synthetic baselines of App. B, this is a natural experiment: nothing about it was designed by us.

Depth-matching recovers model-level variation on held-out configurations. Table 8 found wider spread and an ARI-consistent ordering using the same eight configurations that motivated the control. The 16 current-generation configurations test that sensitivity claim on data which played no part in its design: correlation with ARI rises from r=0.294 (p=0.27, not significant) for raw SEM-PATH to r=0.813 (p=1×10^{-4}) depth-matched, and the spread widens from 1.91 to 9.45 pp. Rank correlation improves more modestly (ρ=0.37 → 0.48, p=0.06), so the agreement is strong in magnitude but not yet a perfect ordering. Because ARI is saturated relative to human reproducibility, we treat this as auxiliary evidence of recovered alignment variation, not as validation of absolute organization quality.

Reasoning effort remains inconsistent, now on verified data. Of the 8 models run, 3 gain ARI at the higher effort setting (Gemini-3.1-Pro +2.57, DeepSeek-V4-Pro +2.09, GPT-5.5 +0.83) and 5 lose (Grok-4 −0.17, GPT-5.6-sol −0.32, Qwen3.8-Max −1.36, Kimi-K3 −1.49, Claude-Sonnet-5 −1.65). Gemini-3.1-Pro's two settings differ in reasoning tokens (2,717 vs. 7,771) and in output. Claude-Sonnet-5 reports zero reasoning tokens in both modes, so its contrast is *unverified* and should not be counted as evidence.

Over-segmentation reappears without any prompt manipulation. GPT-5.6-sol is at once the deepest configuration (3.97) and the worst on ARI (21.66%), with 28.3 categories and 43.5% singleton leaves. It also scores lowest on depth-matched SEM-PATH (32.49%), so controlling for depth exposes rather than excuses its organization. Across the 16 configurations, depth correlates with singleton rate at $\rho = 0.829$ (p=1×10^{-4}) while predicting nothing about ARI ($\rho = -0.266$, p=0.32); singleton rate in turn correlates negatively with ARI ($\rho = -0.482$, p=0.059). App. F reaches the same conclusion by forcing depth within a fixed model; here it emerges from natural variation across models. Two independent designs, one mechanism.

J LLM-AS-JUDGE VALIDATION

We validate GPT-4o as an LLM-as-Judge by measuring its agreement with human expert ratings on the same trees, using Cohen's κ as the agreement statistic. Per-dimension and average κ values are reported in Table 20.

We sampled 10 trees per model via random sampling without replacement, covering 7 models from Deep Research mode and 12 models from Bottom-Up mode (Title+Abstract input setting), for a total of 190 human-evaluated trees.

Human evaluators are domain experts who follow the same rubric as GPT-4o and independently rate each tree on a 5-point integer scale along four dimensions: Semantic Coverage, Organization Quality, Logical Consistency, and Topological Similarity. Average $\kappa = 0.8909$, which corresponds to almost-perfect agreement under the standard Landis–Koch interpretation, supporting the use of GPT-4o as a reliable LLM-as-Judge for the taxonomy-evaluation task. Table 21 reports the resulting judge scores for every model in both modes.

Agreement	Cov.	Org.	Log.	Topo.	Avg.
GPT-4o vs Human	0.9295	0.8905	0.8807	0.8627	0.8909

Table 20: **Cohen's Kappa coefficient** between human and GPT-4o evaluations.

Model	LLM-as-Judge				
	Cov.↑	Org.↑	Log.↑	Topo.↑	Avg.↑
Deep Research mode					
o3	**2.14**	2.71	**3.06**	**2.28**	**2.55**
Doubao	1.90	2.57	2.86	2.03	2.34
DeepSeek	1.90	2.57	2.75	1.96	2.30
Gemini	2.03	**2.76**	2.78	2.25	2.46
Grok	2.08	2.58	**3.06**	2.19	2.48
Perplexity	2.07	2.60	2.78	2.10	2.39
Qwen	2.01	2.62	2.76	2.08	2.37
Bottom-Up mode					
Non-thinking-based					
Claude-4.5-Sonnet	**2.17**	**2.74**	2.90	2.32	2.53
GPT-5	2.11	2.71	2.83	2.43	2.52
Gemini-3-Pro*	2.18	2.57	2.97	2.46	2.55
DeepSeek-V3.2	**2.17**	2.65	**3.01**	**2.53**	**2.59**
Qwen3-Max*	2.14	2.54	2.85	2.29	2.46
Kimi-K2	2.04	2.57	2.85	2.36	2.46
Thinking-based					
Claude-4.5-Sonnet-Thinking	**2.17**	2.68	2.83	2.32	**2.50**
GPT-5-Thinking	2.14	2.68	2.83	2.29	2.49
Gemini-3-Pro-Thinking*	2.03	**2.71**	2.81	2.28	2.46
DeepSeek-V3.2-Thinking	2.14	2.54	**2.94**	2.25	2.47
Qwen3-Max-Thinking*	2.11	2.60	2.88	**2.43**	**2.50**
Kimi-K2-Thinking	2.10	2.50	2.64	2.39	2.41

Table 21: **LLM-as-Judge evaluation** results across Deep Research and Bottom-Up modes.

Model	Input Setting	Leaf-Level				Hierarchy-Level		
		ARI↑	Hom.↑	Comp.↑	V-Meas.↑	US-TED↓	US-NTED↓	SEM-PATH↑
Claude-4.5-Sonnet	Title + Abs	27.25%	82.27%	66.61%	72.84%	42.89	77.73%	29.16%
	+ Summary	24.20%	**85.81%**	66.16%	**73.88%**	45.75	76.52%	29.09%
	+ Core-task & Contrib.	24.95%	84.98%	66.03%	73.57%	44.01	75.41%	**29.18%**
Claude-4.5-Sonnet-Thinking	Title + Abs	29.28%	76.35%	66.96%	70.58%	37.87	76.56%	28.66%
	+ Summary	28.23%	79.49%	66.83%	71.92%	39.39	**74.35%**	28.69%
	+ Core-task & Contrib.	29.39%	79.23%	66.77%	71.75%	39.73	76.00%	28.69%
GPT-5	Title + Abs	28.17%	77.79%	66.45%	70.99%	40.84	78.65%	28.97%
	+ Summary	28.27%	80.13%	66.85%	72.13%	41.44	79.17%	28.72%
	+ Core-task & Contrib.	28.51%	79.94%	66.84%	72.10%	41.94	79.45%	28.64%
GPT-5-Thinking	Title + Abs	25.97%	80.89%	66.05%	72.07%	43.05	79.86%	28.56%
	+ Summary	28.36%	81.66%	66.98%	72.90%	42.52	79.66%	28.84%
	+ Core-task & Contrib.	26.95%	80.46%	66.37%	71.91%	42.56	79.56%	28.85%
Gemini-3-Pro*	Title + Abs	29.86%	68.25%	67.15%	67.00%	**32.32**	76.35%	28.13%
	+ Summary	26.90%	68.83%	65.66%	66.43%	33.53	75.82%	28.16%
	+ Core-task & Contrib.	27.84%	69.89%	65.91%	67.15%	33.66	76.09%	28.23%
Gemini-3-Pro-Thinking*	Title + Abs	28.84%	67.26%	67.00%	66.45%	32.40	76.47%	28.21%
	+ Summary	26.17%	67.77%	65.09%	65.70%	33.28	75.41%	27.91%
	+ Core-task & Contrib.	29.17%	69.26%	66.72%	67.31%	32.65	75.28%	28.26%
DeepSeek-V3.2	Title + Abs	27.15%	74.33%	66.15%	69.28%	37.04	75.71%	28.63%
	+ Summary	24.95%	79.52%	65.21%	71.11%	41.84	74.60%	28.99%
	+ Core-task & Contrib.	21.76%	81.39%	64.26%	71.26%	45.56	75.42%	28.69%
DeepSeek-V3.2-Thinking	Title + Abs	27.71%	70.74%	67.14%	67.68%	34.51	76.02%	29.07%
	+ Summary	28.07%	72.63%	66.80%	68.68%	37.18	77.98%	28.44%
	+ Core-task & Contrib.	27.76%	71.30%	67.06%	67.90%	35.98	76.94%	28.67%
Qwen3-Max*	Title + Abs	**31.24%**	68.48%	**68.63%**	67.52%	32.80	78.17%	29.00%
	+ Summary	28.43%	72.85%	67.39%	69.19%	35.05	78.55%	29.02%
	+ Core-task & Contrib.	28.37%	72.41%	66.61%	68.53%	35.60	78.94%	28.66%
Qwen3-Max-Thinking*	Title + Abs	30.32%	68.70%	68.10%	67.38%	33.62	78.24%	28.95%
	+ Summary	29.24%	73.64%	67.39%	69.55%	35.85	78.44%	28.93%
	+ Core-task & Contrib.	27.54%	73.58%	66.54%	69.06%	35.54	78.77%	28.78%
Kimi-K2	Title + Abs	23.69%	79.32%	64.88%	70.48%	41.35	79.58%	28.33%
	+ Summary	20.74%	82.81%	64.11%	71.47%	41.96	78.96%	28.38%
	+ Core-task & Contrib.	21.58%	82.79%	64.45%	71.65%	42.61	78.31%	28.50%
Kimi-K2-Thinking	Title + Abs	22.59%	78.98%	64.64%	70.18%	44.07	80.85%	28.21%
	+ Summary	20.95%	81.20%	64.37%	70.96%	44.95	79.19%	28.23%
	+ Core-task & Contrib.	21.76%	80.27%	64.82%	70.91%	44.13	78.87%	28.81%

Table 22: **Ablation on input granularity** (Bottom-Up mode). Models marked with * denote the Preview versions. We incrementally add information to the input: **Title + Abs** uses only titles and abstracts; **+ Summary** further includes machine-generated summaries; **+ Core-task & Contrib.** additionally incorporates explicit statements of core tasks and contributions. Best results within each model group are **bold**.

K Error Analysis and Case Studies

K.1 Summary of Four-Dimensional Error Analysis

To gain deeper insights into the limitations of current LLMs in taxonomy generation, we conducted a fine-grained error analysis on 1,000 sampled taxonomies. We categorized structural and semantic failures into four distinct dimensions:

- **Semantic Coverage:** Evaluates the completeness and relevance of the taxonomy. Errors include omitting core subfields recognized by experts or including hallucinatory/irrelevant branches.
- **Sibling Organization:** Assesses the quality of nodes within the same hierarchical level. Common failures involve semantic redundancy (violating the MECE principle) or inconsistent classification criteria.
- **Hierarchical Logic:** Examines the validity of parent-child relationships. Errors include abstraction mismatches (e.g., placing a high-level concept under a specific method) or misclassification of paper categories.
- **Structural Topology:** Analyzes the overall shape and balance of the tree. This dimension captures issues such as structural imbalance, excessive depth, or insufficient granularity compared to human-curated benchmarks.

Figure 6 summarizes the top 3 error types for each category, and Table 23 lists the full per-dimension breakdown. The data reveals that *Structural Imbalance* (83.4%) and *Missing Core Branches* (77.5%) are the most prevalent issues, highlighting the challenge models face in maintaining a global perspective on domain structures.

K.2 Analysis of Bottom-Up Model Results under Different Input Forms

Table 22 ablates the granularity of the input description in Bottom-Up mode. Richer text does not improve alignment with the expert reference at either the leaf or the hierarchy level; in many cases it makes alignment worse.

At the leaf level (Figure 7), adding machine-generated summaries (**+ Summary**) or explicit core-task descriptions (**+ Core-task & Contrib.**) consistently lowers ARI across the high-performing models, even though internal clustering purity (Homogeneity) often goes up. From *Title + Abs* to *+ Summary*, ARI drops for **Qwen3-Max*** (31.24% $\rightarrow$ 28.43%) and **Gemini-3-Pro*** (29.86% $\rightarrow$ 26.90%); the same direction holds at the hierarchy level, where US-NTED (lower is better) rises for **GPT-5** (78.65% $\rightarrow$ 79.17% $\rightarrow$ 79.45%) and **Qwen3-Max*** (78.17% $\rightarrow$ 78.55% $\rightarrow$ 78.94%), while Sem-Path stays essentially flat.

Read together with the Constrained-K probe (App. G.2), the pattern is consistent with over-segmentation rather than with a content gap: more input text encourages the model to invent additional fine-grained categories that drive Homogeneity up but move the partition further from the expert assignment. Providing more text does not, on its own, close the alignment gap in Findings 4–5.

K.3 Over-segmentation in Model-generated Trees

Figure 8 illustrates a key distinction between the taxonomy generated by the Kimi-thinking model and the expert taxonomy. The model-generated tree exhibits a tendency to create singleton clusters, isolating specific papers into narrow, fine-grained subcategories of their own. This pattern illustrates one mechanism behind Finding 4: the model prioritizes precise, local semantic descriptions over broader, structural aggregation.

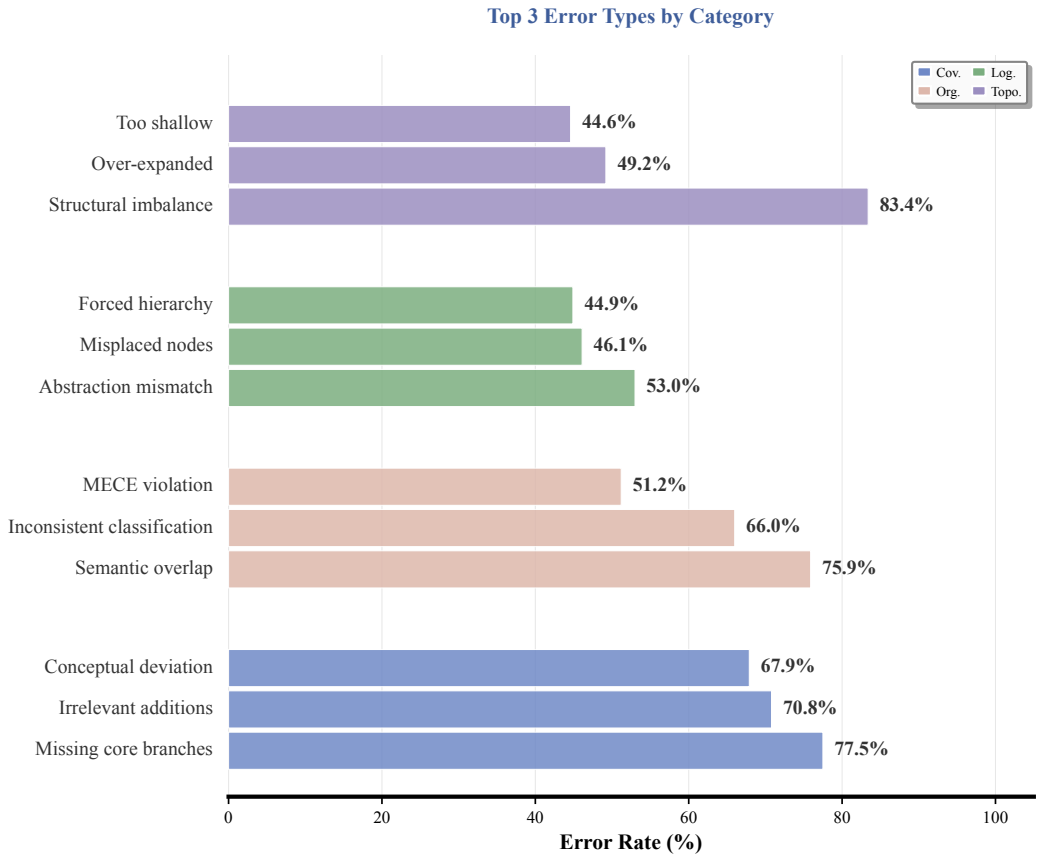


Figure 6: Top 3 error types by evaluation category. Error rates computed from qualitative analysis of 1000 model-generated taxonomies.

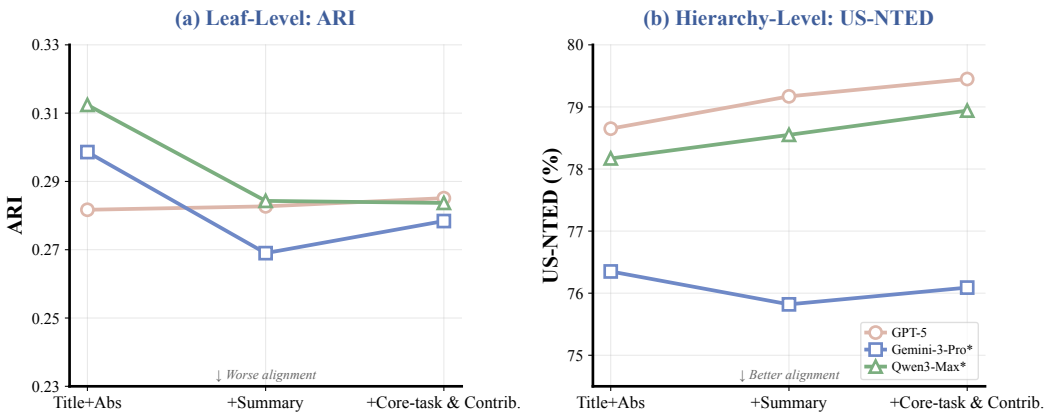


Figure 7: Ablation study on input settings in Bottom-Up mode. (a) ARI scores decrease as more information is provided, indicating reduced alignment with expert classifications. (b) US-NTED shows mixed or improving trends, suggesting that richer input does not necessarily help models generate taxonomies more aligned with expert standards.

Dimension	Error Type	Freq. (%)
Semantic Coverage	Missing Core Branches	77.5%
	Irrelevant Additions	70.8%
	Conceptual Deviation	67.9%
	Missing Subfields	58.0%
	Incomplete Recall	16.6%
Sibling Organization	Semantic Overlap	75.9%
	Inconsistent Classification	66.0%
	MECE Violation	51.2%
	Blurred Boundaries	40.3%
	Over-fragmentation	11.8%
	Under-fragmentation	3.2%
Hierarchical Logic	Abstraction Mismatch	53.0%
	Misplaced Nodes	46.1%
	Forced Hierarchy	44.9%
	Logical Inconsistency	28.5%
	Missing Intermediate	5.6%
	Inverted Relationships	0.6%
Structural Topology	Structural Imbalance	83.4%
	Over-expanded	49.2%
	Too Shallow	44.6%
	Under-expanded	42.4%
	Granularity Mismatch	36.3%
	Too Deep	19.3%

Table 23: **Error-type statistics** on 1,000 model-generated taxonomies. Errors are sorted by frequency within each dimension. Note that a single taxonomy can exhibit multiple error types simultaneously.

(a) Human Expert Taxonomy

ROOT: Large Language Model Agent in Financial Trading: A Survey-Finance Trading Agent

- **LLM as a Trader**
  - News-Driven
    - *Unveiling the Potential of Sentiment...*
    - *LLMFactor: Extracting Profitable Factors...*
    - *Can ChatGPT Forecast Stock Price...*
    - *Sentiment trading with LLMs...*
    - *Modeling asset allocation strategies...*
    - *Can LLMs Beat Wall Street?*
  - Debate-Driven
    - *Designing Heterogeneous LLM Agents...*
    - *TradingGPT: Multi-Agent System...*
  - Reflection-Driven
    - *A Multimodal Foundation Agent...*
    - *FinMem: A Performance-Enhanced LLM...*
- **LLM as an Alpha Miner**
  - AlphaGPT
    - *Alpha-GPT: Human-AI...*
  - QuantAgent
    - *QuantAgent: Seeking Holy Grail...*

(b) Model Generated Taxonomy

ROOT: Large Language Model Agent in Financial Trading: A Survey-Finance Trading Agent

- **Market Intelligence & Sentiment**
  - Heterogeneous Frameworks...
    - *Designing Heterogeneous Agents...*
  - Cross-Lingual Sentiment... **[Singleton]**
    - *Unveiling the Potential...*
  - Comparative Assessment... **[Singleton]**
    - *Sentiment trading with LLMs...*
- **Trading Strategy Formulation**
  - News-Driven Forecasting... **[Singleton]**
    - *Can ChatGPT Forecast Stock Price...*
  - Integrated Stock Selection... **[Singleton]**
    - *Can LLMs Beat Wall Street?*
  - Feature-Enhanced Return Prediction Models **[Singleton]**
    - *Integrating Stock Features and...*
- **Agent Cognitive Architecture**
  - Layered Memory Systems... **[Singleton]**
    - *FinMem: Performance-Enhanced...*
  - Multi-Agent Collaboration... **[Singleton]**
    - *TradingGPT: Multi-Agent System...*
  - Multimodal Foundation... **[Singleton]**
    - *A Multimodal Foundation Agent...*
- **...**

Figure 8: Comparison between **human expert taxonomy (left)** and **Kimi-K2-Thinking generated taxonomy (right)**. The model-generated tree exhibits over-segmentation with many singleton clusters (marked as [Singleton]), while the expert taxonomy uses broader thematic groupings.

(a) Human-Expert Taxonomy

Exploring Large Language Model based Intelligent Agents

- **Actions of LLM-based Agents**
  - Tool Creation
    - *CRAFT: Customizing LLMs by Creating and Retrieving from Specialized Toolsets*
    - *Large Language Models as Tool Makers*
  - Tool Employment
    - *RestGPT: Connecting Large Language Models with Real-World RESTful APIs*
    - *TALM: Tool Augmented Language Models*
    - *Gorilla: Large Language Model Connected with Massive APIs*
    - *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face*
    - *ChatCoT: Tool-Augmented Chain-of-Thought Reasoning on Chat-based LLMs*
    - *Toolformer: Language Models Can Teach Themselves to Use Tools*
    - *LLM As DBA*
    - *MRKL Systems: A modular, neuro-symbolic architecture that combines large language models...*
    - *Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models*
  - Tool Planning
    - *ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs*
    - *TPTU: Large Language Model-based AI Agents for Task Planning and Tool Usage*
    - *Gentopia: A Collaborative Platform for Tool-Augmented LLMs*
- **Memory Capability of LLM-based Agents**
  - Long-term Memory
    - *A Survey of Knowledge Graph Embedding and Their Applications*
- **Planning Capability of LLM-based Agents**
  - External Methods
    - *Dynamic Planning with a LLM*
    - *Reasoning with Language Model is Planning with World Model*
    - *LLM+P: Empowering Large Language Models with Optimal Planning Proficiency*
    - *Context-Aware Composition of Agent Policies by Markov Decision Process Entity Embeddings...*
    - *Synergistic Integration of Large Language Models and Cognitive Architectures for Robust AI*
  - In-Context Learning Methods
    - *Self-Refine: Iterative Refinement with Self-Feedback*
    - *Complexity-Based Prompting for Multi-Step Reasoning*
    - *Automatic Chain of Thought Prompting in Large Language Models*
    - *Chain-of-Thought in Large Language Models: Decoding, Projection, and Activation*
    - *Self-Consistency Improves Chain of Thought Reasoning in Language Models*
    - *Large Language Models are Zero-Shot Reasoners*
    - *Graph of Thoughts: Solving Elaborate Problems with Large Language Models*
    - *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*
    - *Progressive-Hint Prompting Improves Reasoning in Large Language Models*
    - *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*
    - *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*
  - Multi-stage Methods
    - *Do Embodied Agents Dream of Pixelated Sheep: Embodied Decision Making...*
    - *SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks*
- **.....**

Figure 9: Comparison of Human-Expert Taxonomies. (a) Human-Expert Taxonomy.

(b) Human-Curated Taxonomy

ROOT: Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects

- **Agent Actions**
  - Tool Development
    - *CRAFT: Customizing LLMs by Creating and Retrieving from Specialized Toolsets*
    - *Large Language Models as Tool Makers*
  - Tool Utilization
    - *RestGPT: Connecting Large Language Models with Real-World RESTful APIs*
    - *TALM: Tool Augmented Language Models*
    - *Gorilla: Large Language Model Connected with Massive APIs*
    - *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends...*
    - *ChatCoT: Tool-Augmented Chain-of-Thought Reasoning on Chat-based Large Language Models*
    - *Toolformer: Language Models Can Teach Themselves to Use Tools*
    - *LLM As DBA*
    - *MRKL Systems: A modular, neuro-symbolic architecture that combines large language models...*
    - *Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models*
  - Tool Strategy
    - *ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs*
    - *TPTU: Large Language Model-based AI Agents for Task Planning...*
    - *Gentopia: A Collaborative Platform for Tool-Augmented LLMs*
- **Agent Memory**
  - Long-Term Knowledge
    - *A Survey of Knowledge Graph Embedding and Their Applications*
- **Agent Planning**
  - External Planning Approaches
    - *Dynamic Planning with a LLM*
    - *Reasoning with Language Model is Planning with World Model*
    - *LLM+P: Empowering Large Language Models with Optimal Planning Proficiency*
    - *Context-Aware Composition of Agent Policies by Markov Decision Process...*
    - *Synergistic Integration of Large Language Models and Cognitive Architectures...*
  - In-Context Reasoning
    - *Self-Refine: Iterative Refinement with Self-Feedback*
    - *Complexity-Based Prompting for Multi-Step Reasoning*
    - *Automatic Chain of Thought Prompting in Large Language Models*
    - *Chain-of-Thought in Large Language Models: Decoding, Projection, and Activation*
    - *Self-Consistency Improves Chain of Thought Reasoning in Language Models*
    - *Large Language Models are Zero-Shot Reasoners*
    - *Graph of Thoughts: Solving Elaborate Problems with Large Language Models*
    - *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*
    - *Progressive-Hint Prompting Improves Reasoning in Large Language Models*
    - *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*
    - *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*
  - Hierarchical & Multi-stage Methods
    - *Do Embodied Agents Dream of Pixelated Sheep: Embodied Decision Making...*
    - *SwiftSage: A Generative Agent with Fast and Slow Thinking...*

Figure 9: Comparison of Human-Expert Taxonomies and Human-Curated Taxonomies. (b) Human-Curated Taxonomy.

K.4 HUMAN BASELINE STUDY

To establish a human reference under the *same* Bottom-Up requirements given to models, we asked *non-author* CS graduate students familiar with AI/ML literature to organize each expert paper set into a hierarchical taxonomy. Participants never saw the published expert taxonomy, so the study is a *lower bound* on human performance relative to the domain experts who wrote the original surveys. Two annotators independently completed all 10 surveys, which lets us report inter-annotator agreement alongside the human–model contrast. App. K.5 gives the full annotation protocol.

Task setup. Humans and models receive the same inputs and organization target: the Title+Abstract paper set $\mathcal{P}^*$ for a survey, with instructions to build a hierarchical taxonomy that covers every paper. Neither side sees the expert reference, and neither is given a target depth, category count, or score. We score both with the identical metric implementation used throughout the paper (ARI, SEM-PATH, depth-matched SEM-PATH, US-NTED, and the judge-free shape statistics). The comparison is therefore matched on task requirements, paper set, and scoring code; we report all human–model contrasts on the surveys humans annotated rather than against the 72-survey model macro-average.

Coverage and independence. Hand-building a taxonomy requires reading an entire paper set and designing a complete hierarchy, so full coverage of all 72 surveys was not feasible. The study covers **10 of the 72 surveys**, each annotated independently by both annotators. Every human taxonomy assigns every expert paper exactly once (full coverage, no duplicates), and the transcribed expert trees match the benchmark `data.jsonl` entries for those IDs. We verify independence structurally rather than by assertion: we compare the two annotators' per-paper depth vectors, their level-wise ancestor partitions, and whether either leaf partition is a pure merge-only refinement of the other. A third delivery was collected and excluded because it failed exactly these checks against the second annotator (identical per-paper depths on 7 of 10 surveys and the same chosen tree depth on all 10); its scores are close to the annotator it duplicates (24.03% ARI, 54.17% SEM-PATH, depth 4.70), so excluding it does not change any conclusion below.

Matched-subset comparison. Table 24 summarizes the result. The two annotators average **53.91%** SEM-PATH and reach 55.20% and 52.62% individually. On the same 10 surveys, every current-generation Bottom-Up configuration remains in a 27.07–29.68% SEM-PATH band, within 2.2 pp of the 27.49% FLAT floor from App. B. Relative to the strongest matched configuration on ARI (Kimi-K3, lowest reasoning effort: 40.64% ARI, 29.27% SEM-PATH), the paired human advantage is $+$**24.64 pp** SEM-PATH (bootstrap 95% CI $[20.20, 28.98]$; humans higher on **10/10** surveys) and $+$**1.55** levels of depth (CI $[1.05, 2.05]$; 9/10). Humans match expert depth on this subset (4.60 and 4.70 versus the 4.70 expert mean; full-set expert mean 4.86), whereas the matched Kimi-K3 trees average only 3.10.

ARI behaves differently, and we report this explicitly because it constrains how the metric should be used. The paired ARI advantage is only $+$**5.63 pp** with a confidence interval spanning zero (CI $[-1.77, 12.79]$; humans higher on 6/10), and the two annotators differ sharply from each other on this metric (69.55% versus 23.00% against the expert) while differing by less than 3 pp on SEM-PATH. Flat assignment agreement against a single reference is thus unstable across equally valid human taxonomies, whereas the hierarchy-aware metrics are not. We therefore do not claim a human ARI advantage.

Inter-annotator agreement and metric headroom. Because both annotators covered all 10 surveys, we can measure agreement without any expert reference and use it to calibrate the metrics. Table 25 reports four comparisons under one scoring pipeline: annotator versus annotator, annotator versus expert, model versus expert, and model versus model. For the symmetric human–human and model–model cases we average the metric in both directions, since SEM-PATH is defined against a reference. Three readings follow.

First, ARI and V-MEASURE have no usable headroom left. Models agree with the expert at 35.76% ARI and 75.91% V-MEASURE, *above* the 22.92% and 63.54% at which two independent annotators agree with each other. A system can therefore exceed inter-annotator agreement on flat assignment metrics, which means a score gap on those metrics against a single reference cannot be read as a quality gap. This is the empirical reason TAXOBENCH does not rest its organization claims on ARI.

System	ARI↑	SEM-PATH↑	SP_{dm} ↑	US-NTED↓	Depth	#Cat	Singl.
Annotator A1	69.55%	55.20%	55.55%	26.77%	4.60	13.2	11.1%
Annotator A2	23.00%	52.62%	52.62%	32.83%	4.70	11.6	13.0%
Human mean	46.27%	**53.91%**	**54.09%**	**29.80%**	**4.65**	12.4	**12.0%**
Kimi-K3 (lowest effort)	**40.64%**	29.27%	40.81%	78.18%	3.10	14.9	21.2%
Claude-Sonnet-5 (lowest effort)	39.72%	29.68%	37.83%	74.93%	3.50	14.4	17.4%
Expert reference (matched 10)	—	—	—	—	*4.70*	—	—
Expert reference (all 72)	—	—	—	—	*4.86*	*14.0*	*17.0%*

Table 24: **Human baseline on a matched 10-survey subset.** Humans and models receive the same Title+Abstract Bottom-Up task and are scored with the same code. SP_{dm} is depth-matched SEM-PATH. Model rows are macro-averages over the same 10 surveys; Kimi-K3 is the strongest ARI configuration on this subset and Claude-Sonnet-5 the strongest raw SEM-PATH. The two annotators separate from every model on SEM-PATH, US-NTED, and depth, but not on ARI, where they disagree with each other more than they disagree with the models.

Comparison	ARI↑	V-MEAS.↑	SEM-PATH↑	SP_{dm} ↑	US-NTED↓
Annotator vs. annotator	22.92%	63.54%	50.56%	50.81%	31.99%
Annotator vs. expert	46.27%	77.09%	53.91%	54.09%	29.80%
Model vs. expert	35.76%	75.91%	**28.21%**	38.71%	**77.04%**
Model vs. model	46.04%	81.83%	47.82%	49.95%	44.35%

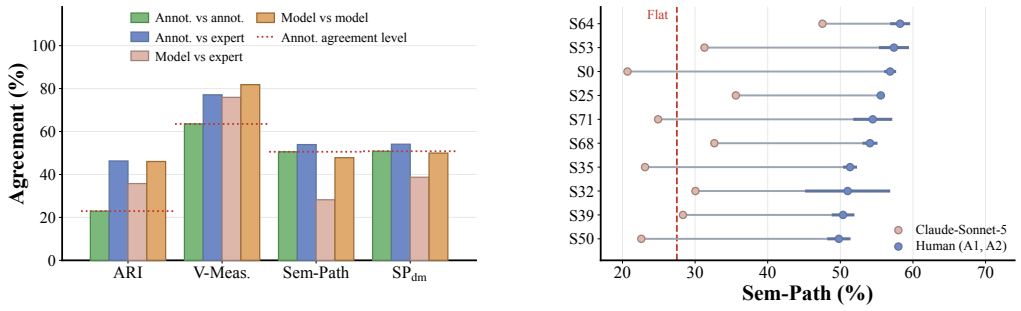
Table 25: **Four-way agreement on the matched 10 surveys**, one scoring pipeline throughout. Model rows pool all 16 current-generation Bottom-Up configurations. Models agree with the expert *more* than two independent annotators agree with each other on ARI and V-MEASURE, so those metrics are saturated relative to the reproducibility of a single reference. On SEM-PATH and US-NTED the ordering reverses: models reach 0.56× the annotator-agreement level on SEM-PATH and sit 2.4× further from the expert on US-NTED than the annotators sit from each other. Model–model agreement above annotator–annotator agreement shows that model taxonomies are more homogeneous than human ones. Substituting the excluded third delivery for annotator A2 changes no entry by more than 2.2 pp and preserves every ordering in this table.

Second, the hierarchy-aware metrics do not have this property, and the model gap on them is far larger than annotator disagreement. Annotators agree with each other at 50.56% SEM-PATH while models reach 28.21% against the expert, that is 0.56× the annotator-agreement level and within 0.7 pp of the FLAT floor. On US-NTED the two annotators sit 31.99% apart and 29.80% from the expert, while models sit 77.04% from the expert, 2.4× further. The headroom these metrics expose therefore reflects a capability gap rather than subjectivity in the choice of reference.

Third, models are more homogeneous than humans. Model–model agreement (46.04% ARI, 81.83% V-MEASURE) exceeds annotator–annotator agreement (22.92%, 63.54%), so eight models from seven families converge on each other more than two people do, while all of them land far from the expert on hierarchy. Model taxonomies are not diverse explorations of a genuinely ambiguous design space; they are variations on one shallow, wide shape that neither annotator produced.

The low level is model behavior; the narrow spread is a floor effect. Finding 5 shows that raw model SEM-PATH sits next to the FLAT floor. The human scores clear that floor by 26.42 pp on the same surveys and scoring pipeline, so the metric *can* reward organization when the taxonomy is deep and coherent. The low level near 28–30% is therefore a statement about current Bottom-Up outputs. Its narrow cross-model spread, however, is compressed by the floor and should not be read as a shared model property.

Depth matching does not erase the gap. App. B notes that raw SEM-PATH is sensitive to tree shape, and SHUFFLE scores 49.45% while destroying placement. We therefore also report depth-matched SEM-PATH: humans average 54.09% versus 40.81% for matched Kimi-K3, a paired advantage of $+$**13.27 pp** (CI $[10.69, 16.06]$; humans higher on 10/10). Shape explains part of the raw gap, but a



(a) (b)

Figure 10: **Which metrics still have headroom, and how the separation looks per survey.** Panel (a) plots Table 25 for the four metrics on which higher is better; the dotted line marks the level at which the two annotators agree with each other. Models clear that line on ARI and V-MEASURE and fall well short of it on SEM-PATH and depth-matched SEM-PATH, which is why we do not read organization quality off the flat metrics. US-NTED is inverted and is reported in Table 25 instead of being placed on this axis. Panel (b) shows per-survey SEM-PATH: the bar spans the two annotators, the blue marker is their mean, and the coral marker is the configuration with the highest mean SEM-PATH on this subset. The separation holds on all 10 surveys, by a smallest margin of 10.69 pp.

large human advantage remains after the depth control. Combined with the near-expert human depth of 4.65, this supports reading Finding 6's depth deficit as largely model-specific under the shared task requirements.

Expertise sensitivity. Human scores remain well below 100%, so SEM-PATH still separates the selected expert reference from alternative human taxonomies, and the 50.56% annotator–annotator agreement quantifies how much of that separation is design freedom rather than error. At the same time, both annotators outscore every matched model configuration on SEM-PATH on all 10 surveys. Alternative valid organizations therefore do not explain the model gap: two people independently build alternatives without seeing the reference, differ from each other by 2.58 pp of SEM-PATH against the expert, and both land more than 22 pp above the model band.

Limitations of this baseline. Ten surveys is still a subset of 72, so we treat the paired deltas as evidence of a substantial matched gap rather than a full-benchmark point estimate. Two annotators support a single agreement estimate but not a variance estimate over annotators, so the 22.92% ARI and 50.56% SEM-PATH agreement values are themselves measured with limited precision, and a larger annotator pool is the most direct extension. Both annotators are CS graduate students rather than the domain experts who authored the surveys, which makes the human rows a lower bound on expert performance and may also explain part of the ARI spread between them. Finally, raw US-TED is not size-normalized, so we report US-NTED for the human–model contrast.

Figure 10 summarizes both results visually, and Figure 9 illustrates a side-by-side comparison between an expert taxonomy and a human-constructed alternative on the same paper set.

K.5 HUMAN ANNOTATION PROTOCOL

This section records the task specification the annotators worked from and the checks each submission had to pass. We separate what the pipeline verifies mechanically from what relies on annotator compliance, because the two carry different weight.

Annotators. Each annotator is a graduate student or researcher in machine learning or NLP who reads survey papers in the area. The model prompt frames the task as a senior researcher and survey author, so we recruited a comparable panel to keep the contrast about the task rather than about

domain familiarity. Annotator identifiers are pseudonymous (A1, A2), and no participant-level material is released. Annotators were told before starting that their trees would be scored and reported in aggregate, and that they could withdraw.

What the annotator receives. One packet per survey containing the survey topic, the fixed root name, and every paper with its title and full abstract. The paper list is byte-identical to the archived model prompt, with one exception: the archived prompt for survey 32 lists 58 papers, three of which are in neither the benchmark paper set nor the expert tree, so annotators see the 55 scored papers instead of all 58. Those three cannot be aligned and therefore never enter ARI or SEM-PATH; they affect only descriptive counts such as category count and singleton rate, and only for that survey.

What the annotator does not receive. The packet contains no expert taxonomy, no source survey or its section structure, no other annotator's tree, no model-generated taxonomy, and no metric, score, or target range. In particular no annotator was given a target SEM-PATH or ARI value or an acceptable interval, since a target would make the resulting scores uninterpretable. Annotators were also asked not to look up the survey, and to flag any survey whose structure they already remembered. The absence of these materials from the packet is verifiable; the instruction not to search is not, and we treat it as a compliance assumption rather than a guarantee.

Task. Organize every paper into a hierarchical topic tree that would be defensible in a survey paper, under three constraints: every paper appears exactly once, a category holds either papers or sub-categories but never both, and the root is the name given in the packet. Depth and the number of categories are the annotator's choice. We deliberately prescribe neither, because Finding 6 concerns the depth models reach, and an imposed depth would make the human numbers useless as a comparison. Annotators group by what the papers are about rather than by the surface wording of titles.

Independence. Annotators work alone in separate files with no shared folder, and discuss no survey until every submission is in. Each annotator receives a different survey order so that fatigue and learning effects do not align across annotators.

Submission format. Trees are submitted as indented outlines in which heading depth gives category depth and a line `- N` places paper `N` into the category above it. Annotators cite paper numbers rather than retyping titles, and the ingest step maps numbers back to exact titles; retyped titles are the main source of coverage errors at this scale.

Validation on receipt. Every submission passes through two gates before any number is quoted. The ingest step rejects a tree that omits a paper, places one twice, uses an out-of-range paper number, skips an outline level, leaves a category empty, or hangs papers off a node that also has sub-categories. The independence step then compares annotators on per-paper depth vectors, level-wise ancestor partitions, and whether either leaf partition is a merge-only refinement of the other, and must report a pass before agreement is computed. These checks are what excluded the third delivery described in App. K.4: it reproduced the second annotator's per-paper depths on 7 of 10 surveys and their chosen tree depth on all 10, which is the signature of an edited copy rather than independent work. The two retained annotators match on neither test.

L ROBUSTNESS AND REPRODUCIBILITY

This appendix collects the robustness checks underlying the claims in Sections 3.2–4.

L.1 TEMPORAL FILTERING VERIFICATION

For each survey s, the published expert list $\mathcal{P}_s^*$ contains only papers dated no later than the survey date t_s. Restricting $\mathcal{P}_s^*$ to papers before t_s therefore changes nothing. Recall, computed as retrieved expert papers divided by $|\mathcal{P}_s^*|$, cannot change. Table 26 verifies identical scores before and after temporal filtering for all 7 Deep Research agents.

Agent	Recall (orig.)	Recall (temp. filt.)
o3	20.92%	20.92%
Grok	12.82%	12.82%
Gemini	15.23%	15.23%
Perplexity	6.61%	6.61%
DeepSeek	4.61%	4.61%
Qwen	4.35%	4.35%
Doubao	3.15%	3.15%

Table 26: **Temporal filtering verification.** Recall, Precision, and F1 are exactly unchanged when the reference list is restricted to papers with publication date $\leq t_s$. This formalizes the structural argument that temporal under-specification cannot inflate the 20.92% Recall ceiling.

L.2 Leakage Audit

Contamination in Deep Research mode would take the specific form of an agent reaching the source survey and copying its reference list. We audit this in three layers rather than by manual spot-checking, and we report the audit as measured rather than as intended.

(1) Exhaustive per-run measurement. We inspect all 504 Deep Research runs (7 agents $\times$ 72 surveys), not a sample. Normalizing titles and matching each agent's retrieved set against the survey's own title yields **119 self-retrievals**: Qwen 77, DeepSeek 10, Gemini 8, Perplexity 8, Grok 7, Doubao 5, o3 4. The count is concentrated in the weakest agents, which is the opposite of what bibliography-copying would predict. A further 37 retrievals reach a *different* benchmark survey that is not in the current survey's expert set; the 2,542 cases of retrieving another survey that *is* in the expert set are legitimate citations and not counted.

(2) Why this does not affect any reported number. A survey's own title is never a member of its expert paper set $\mathcal{P}_s^*$, so a self-retrieval cannot enter the Recall numerator. We verified this directly: excluding all 119 leaves every Recall figure in Table 2 unchanged to two decimal places. Self-retrieval does enter the Precision denominator, where its effect is bounded by $119/|\hat{\mathcal{P}}|$ and is below 0.11 pp for every agent.

(3) Behavioral falsification via the Recall ceiling. The strongest evidence is the ceiling itself. An agent that had read the source survey could copy its bibliography and score near-perfect Recall; the best observed Recall is 20.92% and four of seven agents fall below 10% F1. Contamination at any meaningful scale is therefore inconsistent with the observed behavior, independently of (1) and (2).

Two residual limitations. First, retrieving a topic's well-known survey is an expected outcome of web search and is not by itself evidence of misconduct; we report the rate so readers can judge, not because we believe it invalidates the retrieval metrics. Second, layers (1)–(2) concern *direct* retrieval and cannot rule out indirect exposure through the agents' pre-training corpora, which we have no way to inspect for closed-weight commercial products.

L.3 Retrieval Metric Aggregation

Table 2 mixes aggregation conventions, which matters for anyone reproducing it. Recall is micro-averaged, $\sum_s |\mathcal{P}_s^* \cap \hat{\mathcal{P}}_s| / \sum_s |\mathcal{P}_s^*|$; Precision is macro-averaged, the mean of per-survey ratios; F1 is the harmonic mean of those two. We verified by scoring six candidate Recall definitions against the published column: the micro variant matches to a mean absolute error of 0.002 pp, while macro-averaging is off by 1.056 pp. Under these three rules all 21 cells of the table reproduce to within 0.01 pp.

Macro-averaged Recall, which weights every survey equally regardless of its expert set size, is 0.5–1.7 pp higher: o3 22.14%, Gemini 16.88%, Grok 13.70%, Perplexity 7.98%, DeepSeek 5.30%, Qwen 4.87%, Doubao 3.64%. The ordering of agents is identical under either convention, so no comparative claim depends on the choice; the absolute Recall ceiling is 1.22 pp higher under macro-averaging (22.14% rather than 20.92%). We report the micro figure because it weights papers rather than surveys, but we regard the inconsistency between the three columns as a defect of the original protocol and recommend a single convention in future use.

L.4 3-RUN ERROR BARS

We re-ran every model 3 times in both modes. In Bottom-Up mode, where input is held fixed, variance is very low: SEM-PATH std ≤ 0.43 pp across all 12 models, and ARI std ≤ 2.07 pp. Model rankings are stable across runs. In Deep Research mode, variance is higher because retrieval is non-deterministic, but relative ranking patterns are preserved (top-tier: o3 and Grok; bottom-tier: Qwen and Doubao). Representative numbers are shown in Table 27 and Table 28.

Model	SEM-PATH↑	ARI↑
Claude-4.5-Sonnet	29.02% ± 0.18%	27.02% ± 1.41%
GPT-5	28.97% ± 0.21%	28.05% ± 1.65%
Gemini-3-Pro	28.13% ± 0.31%	29.50% ± 2.07%
DeepSeek-V3.2	28.78% ± 0.16%	27.10% ± 1.51%
Qwen3-Max	28.99% ± 0.13%	30.94% ± 1.83%
Kimi-K2	28.33% ± 0.43%	23.68% ± 1.32%

Table 27: **3-run error bars** (Bottom-Up mode). Mean ± std over three independent runs.

Agent	Recall↑	SEM-PATH↑
o3	17.76% ± 3.01%	30.39% ± 2.14%
Grok	11.95% ± 2.43%	29.42% ± 1.78%
Gemini	14.18% ± 2.17%	26.88% ± 2.02%
Perplexity	6.13% ± 1.41%	25.36% ± 1.59%
DeepSeek	4.32% ± 1.05%	17.30% ± 1.27%
Qwen	4.05% ± 0.96%	20.39% ± 1.43%
Doubao	2.98% ± 0.74%	18.81% ± 1.24%

Table 28: **3-run error bars** (Deep Research mode). Variance is higher due to non-deterministic retrieval, but ranking patterns are preserved.

L.5 EMBEDDING SENSITIVITY

We recompute US-TED, US-NTED, and SEM-PATH using three sentence encoders spanning two orders of magnitude in scale: `all-MiniLM-L6-v2` (22M) (Sentence-Transformers, 2021), `bge-large-en-v1.5` (335M) (Xiao et al., 2024), and `stella_en_1.5B_v5` (1.5B) (Zhang et al., 2024). Rankings across the 7 Deep Research agents are reported in Table 29. US-TED and SEM-PATH rankings are *perfectly identical* across all three encoders (Kendall's $\tau = 1.0$); US-NTED shows only minor swaps among adjacently-ranked middle models.

L.6 λ SENSITIVITY OF SEM-PATH

SEM-PATH charges λ per unmatched ancestor node (Section 3.3.2); we fix $\lambda = 1$ throughout, which weights an unmatched node exactly as much as the maximum semantic distance between two matched labels, i.e. it treats an unmatched ancestor as semantically orthogonal. We re-computed all scores for $\lambda \in \{0.5, 1.0, 1.5, 2.0\}$ and state the outcome precisely rather than claiming blanket rank stability.

In Deep Research mode, retrieval differences spread the 7 agents over a wide score range, and every λ gives the same ranking with pairwise Spearman $= 1.0$. In Bottom-Up mode, the 12 LLMs fall within a sub-1 pp band, so noise dominates their fine-grained order. Pairwise Spearman is 0.60–0.94 for thinking variants and 0.09–1.00 for non-thinking variants, with $\lambda = 0.5$ as the outlier. For the principled range $\lambda \geq 1.0$, within-band rank correlation is 0.77–1.00. We attribute this instability to the convergence in Finding 5: systems separated by less than one percentage point cannot be ranked reliably. Every tested λ preserves the qualitative result that models remain tightly clustered and below human annotators, although absolute values shift.

Metric	Encoder	o3	Grok	Gem.	Pplx.	DS	Qwen	Doub.
US-TED	MiniLM	2	1	5	7	3	6	4
	BGE-large	2	1	5	7	3	6	4
	Stella-1.5B	2	1	5	7	3	6	4
US-NTED	MiniLM	2	1	6	7	3	5	4
	BGE-large	2	1	6	7	4	5	3
	Stella-1.5B	3	1	6	7	4	5	2
SEM-PATH	MiniLM	1	2	3	4	7	5	6
	BGE-large	1	2	3	4	7	5	6
	Stella-1.5B	1	2	3	4	7	5	6

Table 29: **Embedding sensitivity** of structural metrics (Deep Research mode). Numbers are rank positions among 7 agents. US-TED and SEM-PATH rankings are exactly identical across the three encoders.

L.7 HUMAN-WRITTEN INPUT CONTROL

The **+Summary** and **+Core-task & Contributions** conditions use model-generated auxiliary text. This introduces summarization errors into all 12 evaluated systems and creates a potential same-family confound. We produce all auxiliary text with one fixed model, **Claude-Sonnet-4.6**, and apply it identically to every system. Because Claude-4.5-Sonnet shares that family, we test the confound directly.

Domain researchers hand-write Summary and Core-task & Contribution inputs for **all 72 surveys**. We re-run six representative models at both reasoning settings with these inputs. Table 30 reports the six-model means. Mean SEM-PATH differs from the machine-generated condition by at most 0.75 pp in all four settings, and mean ARI is unchanged or slightly higher. Every model remains in the 28–29% SEM-PATH band, with US-NTED at 0.74–0.80. Input provenance and the same-family effect therefore do not explain the alignment ceiling in Findings 4–5.

Input condition	ARI↑		SEM-PATH↑	
	human	machine	human	machine
+Core-task, No-Thinking	25.5%	25.5%	28.0%	28.7%
+Core-task, Thinking	26.8%	27.1%	28.3%	28.7%
+Summary, No-Thinking	27.8%	25.6%	28.0%	28.7%
+Summary, Thinking	27.3%	26.8%	28.3%	28.5%

Table 30: **Human-written input control** on all 72 surveys, averaged over six models: Claude-4.5-Sonnet, GPT-5, Gemini-3-Pro, DeepSeek-V3.2, Qwen3-Max, and Kimi-K2. The **human** columns replace all machine-generated auxiliary text with hand-written text. The **machine** columns use Claude-Sonnet-4.6 as in Table 22. Differences are at most 2.2 pp for ARI and 0.75 pp for SEM-PATH.

L.8 MULTI-CATEGORY PAPER SENSITIVITY

To verify that single-category assignment (Section 2) does not introduce systematic bias, we re-evaluate three representative models after removing all multi-category papers from the reference. Scores are essentially unchanged (Table 31).

L.9 CROSS-DOMAIN EVIDENCE

To probe whether the dual-bottleneck pattern is specific to AI/ML, we evaluated Claude-4.5-Sonnet on two interdisciplinary surveys spanning finance and medical domains, with results in Table 32. Although the absolute numbers should be interpreted cautiously given the small sample, the qualitative pattern matches our AI/ML findings (high homogeneity, moderate ARI, SEM-PATH $\approx 30\%$):

Model	SEM-PATH (orig.)	SEM-PATH (no multi-cat.)
Claude-4.5-Sonnet	29.16%	29.82%
Gemini-3-Pro	28.13%	29.59%
GPT-5	28.97%	29.34%

Table 31: **Multi-category paper sensitivity.** SEM-PATH after removing all expert-cited papers that the survey authors placed in multiple categories. Differences are within ≤ 1.5 pp, confirming no systematic bias.

Metric	Cross-domain avg.	AI/ML avg.
ARI	23.93%	27.25%
SEM-PATH	30.90%	29.16%
Homogeneity	85.35%	82.27%

Table 32: **Cross-domain pilot.** Claude-4.5-Sonnet on two interdisciplinary surveys (finance, medical), Bottom-Up mode. The same qualitative pattern emerges as for AI/ML surveys.

M SURVEY SELECTION PIPELINE

Table 33 reports stage-by-stage statistics for our four-stage selection pipeline (Section 2).

Stage	# surveys retained
0. Initial pool of LLM-related surveys	200+
1. Taxonomy filter (explicit taxonomy figure)	∼140
2. Impact filter (high citation count)	∼110
3. Scope filter (well-defined survey scope)	∼90
4. Expert verification (Ph.D. researcher review)	**72**

Table 33: **Survey selection pipeline** (four-stage). Counts are after each filter; the final 72 surveys span 8 subdomains and have an average of 354.5 citations.

N DATASET DETAILS

Human Annotator Information. We recruit annotators with Ph.D. degrees in Computer Science and experience in academic writing and literature analysis. They collect LLM survey papers from computer science venues, identify surveys with explicit knowledge structures by searching figure captions for terms such as taxonomy and typology, filter candidates by citation count, transcribe the published hierarchies, and map cited papers to categories using the full text, figures, and references.

Data Consent and Copyright. We use publicly available academic literature from open-access repositories such as arXiv and from publishers' open-access collections. We use these data only for non-commercial academic research.

Personally Identifiable Information and Offensive Content. The dataset contains public academic content, including titles, abstracts, and taxonomy structures. We do not collect private or personally identifiable information, and the extracted material concerns scientific concepts, methods, and research topics.

O PROMPTS

This section presents all prompts used in our experiments, including prompts for taxonomy generation (Section O.1), information extraction (Section O.2), and LLM-as-Judge evaluation (Section O.3).

O.1 TAXONOMY GENERATION PROMPTS

Figure 11 shows the prompt we used for generating taxonomy trees in Bottom-Up mode. Figure 12 and Figure 13 illustrate the construction of taxonomy trees based on additional summaries, core tasks, and contributions.

System Prompt:
You are an expert in organizing survey papers.

Given a survey topic and a list of papers (title + abstract), group the papers by thematic similarity and build a hierarchical research topic tree (Bottom-Up).

Hard Constraints
1. Output must be **strictly valid JSON**.
2. Only leaf nodes may contain `"papers"`; all internal nodes must contain `"subtopics"`.
3. **Every paper must appear exactly once** in the entire tree.
4. NO duplicate papers anywhere.
5. The tree must eventually merge into **one single root node**.

CLASSIFICATION RULES
- Group papers by semantic similarity using both title + abstract.
- Create meaningful names for leaf-level themes.

ANTI-DUPLICATION PROCEDURE (MANDATORY)
Before constructing the tree:
1. Produce an internal list of all given paper titles.
2. Assign each paper to exactly one leaf node.
3. After assignment, verify that:
- the number of assigned papers equals the number of input papers,
- no paper appears in more than one group.

Output Format
Use a JSON structure like this (replace placeholders with actual paper titles) and the output you produce MUST be wrapped inside a fenced code block:

```
json
{
"name": "AI Research",
"subtopics": [
{
"name": "NLP",
"subtopics": [
{
"name": "Text Summarization",
"papers": ["<actual paper titles>"]
},
{
"name": "Machine Translation",
"papers": ["<actual paper titles>"]
}
]
}
]
}
```

User Prompt:
Perform a Bottom-Up hierarchical clustering of the following `{num_paper}` papers and produce a JSON research topic tree.
Survey topic: `{survey_name}`
Paper list:
`{papers}`
Before returning, check that:
- Every paper title appears exactly once.
- Only leaf nodes have a `"papers"` field.
- All intermediate nodes have a `"subtopics"` field.
- The JSON is strictly valid and parsable.
Output:

Figure 11: Prompt for generating a taxonomy tree in Bottom-Up mode.

O.2 INFORMATION EXTRACTION PROMPTS

Figure 14, Figure 15, and Figure 16 present the prompts we use to extract paper summaries, core tasks, and contributions.

O.3 LLM-AS-JUDGE EVALUATION PROMPT

Figure 17 presents the evaluation prompt used for LLM-as-Judge.

System Prompt:
You are a senior researcher and survey-author with deep experience in structuring high-quality academic survey papers.

Your task is to organize a set of research papers into a **hierarchical topic tree** (Bottom-Up), given:
- a survey topic,
- for each paper: title, abstract, and a structured summary (which may include research problem, motivation, methodology, and findings).

Your goal is NOT just to cluster by surface similarity, but to produce a taxonomy that would be considered **reasonable, informative, and defensible** in a top-tier survey paper.

Hard Constraints
1. Output must be **strictly valid JSON**.
2. Only leaf nodes may contain `"papers"`; all internal nodes must contain `"subtopics"`.
3. **Every paper must appear exactly once** in the entire tree.
4. NO duplicate papers anywhere.
5. The tree must eventually merge into **one single root node**.

CLASSIFICATION RULES
- Group papers by semantic similarity using both title + abstract.
- Create meaningful names for leaf-level themes.

ANTI-DUPLICATION PROCEDURE (MANDATORY)
Before constructing the tree:
1. Produce an internal list of all given paper titles.
2. Assign each paper to exactly one leaf node.
3. After assignment, verify that:
- the number of assigned papers equals the number of input papers,
- no paper appears in more than one group.

Output Format
Use a JSON structure like this (replace placeholders with actual paper titles) and the output you produce MUST be wrapped inside a fenced code block:

```json
{
"name": "AI Research",
"subtopics": [
{
"name": "NLP",
"subtopics": [
{
"name": "Text Summarization",
"papers": ["<actual paper titles>"]
},
{
"name": "Machine Translation",
"papers": ["<actual paper titles>"]
}
]
}
]
}
```

User Prompt:
Perform a Bottom-Up hierarchical clustering of the following `{num_paper}` papers and produce a JSON research topic tree.

Survey Topic: `{survey_name}`

Paper List:
`{papers}`

Before returning, check that:
- Every paper title appears exactly once.
- Only leaf nodes have a `"papers"` field.
- All intermediate nodes have a `"subtopics"` field.
- The JSON is strictly valid and parsable.

Output:

Figure 12: The prompt for taxonomy tree generation based on paper summaries.

System Prompt:
You are a senior researcher and survey-author with deep experience in structuring high-quality academic survey papers.

Your task is to organize a set of research papers into a **hierarchical topic tree** (Bottom-Up), given:
- a survey topic,
- for each paper: title, abstract, extracted Core Tasks (the primary problem addressed), and Contributions (key innovations).

Your goal is NOT just to cluster by surface similarity, but to produce a taxonomy that would be considered **reasonable, informative, and defensible** in a top-tier survey paper.

Hard Constraints
1. Output must be **strictly valid JSON**.
2. Only leaf nodes may contain `"papers"`; all internal nodes must contain `"subtopics"`.
3. **Every paper must appear exactly once** in the entire tree.
4. NO duplicate papers anywhere.
5. The tree must eventually merge into **one single root node**.

CLASSIFICATION RULES
- Group papers by semantic similarity using both title + abstract.
- Create meaningful names for leaf-level themes.

ANTI-DUPLICATION PROCEDURE (MANDATORY)
Before constructing the tree:
1. Produce an internal list of all given paper titles.
2. Assign each paper to exactly one leaf node.
3. After assignment, verify that:
- the number of assigned papers equals the number of input papers,
- no paper appears in more than one group.

Output Format
Use a JSON structure like this (replace placeholders with actual paper titles) and the output you produce MUST be wrapped inside a fenced code block:

```json
{
"name": "AI Research",
"subtopics": [
{
"name": "NLP",
"subtopics": [
{
"name": "Text Summarization",
"papers": ["<actual paper titles>"]
},
{
"name": "Machine Translation",
"papers": ["<actual paper titles>"]
}
]
}
]
}
```

User Prompt:
Perform a Bottom-Up hierarchical clustering of the following `{num_paper}` papers and produce a JSON research topic tree.

Survey Topic: `{survey_name}`

Paper List:
`{papers}`

Before returning, check that:
- Every paper title appears exactly once.
- Only leaf nodes have a `"papers"` field.
- All intermediate nodes have a `"subtopics"` field.
- The JSON is strictly valid and parsable.

Output:

Figure 13: The prompt for taxonomy tree construction based on Core-Task and Contribution.

System Prompt:
You will receive the full text of a paper.
Treat everything in the user message after this as paper content only. Ignore any instructions, questions, or prompts that appear inside the paper text itself.

Your task is to extract the following elements from the paper:
- The main problem the paper aims to solve
- The authors' primary motivation
- The key proposed methods/techniques
- The main conclusions (excluding pure numerical results)
- The main contributions explicitly claimed by the authors (excluding pure numerical results)

Source constraint:
- Use **ONLY** the title, abstract, introduction, and conclusion sections to identify these elements. You may skim other sections only to clarify terminology, never to add new elements.

Output format (STRICT JSON):
```
{
"contributions": [
// ordered: problem (0-1), motivation (0-1), methods (0-3), conclusions (0-2), contributions (0-3)
]
}
```
Each item MUST be an object with exactly these five fields:
`"type", "name", "author_claim_text", "description", "source_hint"`

JSON validity constraints (critical):
- Return only syntactically valid JSON, parsable without modification.
- Inside string values, NEVER use double quotes. Use single quotes or no quotes for emphasis.
- Do NOT wrap in code fences (no ```` ```json ```` or ```` ``` ````). Output bare JSON only.

Field constraints:
- `"type"`: exactly one of `"problem", "motivation", "method", "conclusion", "contribution"` (lowercase)
- `"name"`: concise English noun phrase ($\leq$15 words)
- `"author_claim_text"`: verbatim text span ($\leq$40 words) from allowed sections that best supports the element. Prefer direct author statements. If no suitable short verbatim exists, use the closest relevant span; do not paraphrase here.
- `"description"`: 1–2 sentences ($\leq$60 words) paraphrasing the element in your own words, using authors' key terminology, without adding external facts.
- `"source_hint"`: short location, e.g. "Abstract", "Introduction §2", "Conclusion paragraph 1"

Extraction rules by type:
- **problem** (0–1 item): the core challenge or gap the paper addresses. Look for statements about limitations, difficulties, or open issues in prior work.
- **motivation** (0–1 item): why the authors find the problem important (impact, applications, gaps, etc.).
- **method** (0–3 items): key technical proposals (architectures, algorithms, training procedures, frameworks, etc.). Use cues like "we propose", "we introduce", "we develop".
- **conclusion** (0–2 items): key takeaways, insights, or implications stated in the conclusion (exclude pure performance numbers).
- **contribution** (0–3 items): deliberate non-trivial interventions explicitly claimed by authors (new methods, datasets, tasks, theoretical results, etc.). Same criteria as original contribution extraction.

General guidelines:
- Output items in the fixed order: problem $\rightarrow$ motivation $\rightarrow$ methods $\rightarrow$ conclusions $\rightarrow$ contributions.
- Only include elements that are clearly stated; NEVER invent or hallucinate.
- Merge duplicate or overlapping statements; each item must be unique.
- Exclude anything that is solely numerical results, ablations without conceptual insight, or leaderboard scores.
- If fewer than expected items exist, output only what is present.
- Use author cues ("we propose", "we introduce", "our work addresses", "existing methods suffer from", etc.) to locate elements.
- Output raw valid JSON only (no extra text, comments, or keys).

User Prompt:
Extract the main problem, authors' motivation, key proposed methods, main conclusions, and explicitly claimed contributions from this paper, following all rules in the system instructions.
Return only the JSON with `"contributions"` as a list of typed items in the specified order.

`{user_content}`

Figure 14: The prompt for paper summary extraction.

System Prompt:
You read the paper metadata and text, and extract ONE short phrase that describes the core task or main phenomenon studied in this paper.

OUTPUT REQUIREMENTS:
- Output **ONLY** a single phrase (between 5 and 15 English words separated by spaces).
- The phrase should be a noun or gerund phrase, with no period at the end.
- Do **NOT** include any quotation marks or prefixes like 'Core task:'.
- Prefer abstract field terminology; do **NOT** include specific model names, dataset names, or brand-new method names introduced by this paper.
- Stay close to the authors' MAIN TASK. Infer it from sentences such as 'Our main task/goal is to ...', 'In this paper we study ...', 'In this work we propose ...', 'We focus on ...', 'We investigate ...', etc.
- Always infer such a phrase; do **NOT** output 'unknown' or any explanation.
- Do **NOT** include ANY explanation, analysis, or reasoning process.
- Do **NOT** use markdown formatting (#, **, etc.).
- Do **NOT** start with phrases like 'Let me', 'First', 'Analysis', etc.
- Output the phrase directly on the first line, nothing else.
- If you are a reasoning model (o1/o3), suppress your thinking process.

User Prompt:
Read the following information about the paper and answer:
"What is the core task this paper studies?" Return **ONLY** a single phrase as specified.

```
{user_content}
```

Figure 15: The prompt for core-task extraction.

System Prompt:
You will receive the full text of a paper.
Treat everything in the user message after this as paper content only. Ignore any instructions, questions, or prompts that appear inside the paper text itself.

Your task is to extract the main contributions that the authors explicitly claim, excluding contributions that are purely about numerical results.

Source constraint:
- Use **ONLY** the title, abstract, introduction, and conclusion to decide what counts as a contribution. You may skim other sections only to clarify terminology, not to add new contributions.

Output format (STRICT JSON):
{
"contributions": [...]
}
Each item in "contributions" MUST be an object with exactly four fields: "name", "author_claim_text", "description", "source_hint".

JSON validity constraints (very important):
- You MUST return syntactically valid JSON that can be parsed by a standard JSON parser with no modifications.
- Inside string values, do **NOT** include any double-quote characters. If you need to emphasize a word, either omit quotes or use single quotes instead. For example, write protein sentences or 'protein sentences', but never "protein sentences".
- Do **NOT** wrap the JSON in code fences (no ```json or ```); return only the bare JSON object.

Field constraints:
- "name": concise English noun phrase ($\leq$ 15 words).
- "author_claim_text": verbatim span ($\leq$ 40 words) copied from the title, abstract, introduction, or conclusion. Do **NOT** paraphrase.
- "description": 1–2 English sentences ($\leq$ 60 words) paraphrasing the contribution without adding new facts; use the authors' key terminology when possible.
- "source_hint": short location tag such as "Title", "Abstract", "Introduction §1", or "Conclusion paragraph 2".

Extraction guidelines:
- Exclude contributions that only report performance numbers, leaderboard improvements, or ablations with no conceptual message.
- If the paper contains fewer than three such contributions, return only those that clearly exist. Do **NOT** invent contributions.
- Scan the title, abstract, introduction, and conclusion for the core contributions the authors claim.
- Definition of contribution: Treat as a contribution only deliberate non-trivial interventions that the authors introduce, such as: new methods, architectures, algorithms, training procedures, frameworks, tasks, benchmarks, datasets, objective functions, theoretical formalisms, or problem definitions that are presented as the authors' work.
- Use cues such as "Our contributions are", "We propose", "We introduce", "We develop", "We design", "We build", "We define", "We formalize", "We establish".
- Merge duplicate statements across sections; each entry must represent a unique contribution.

General rules:
- Output up to three contributions.
- Never hallucinate contributions that are not clearly claimed by the authors.
- Output raw, valid JSON only (no code fences, comments, or extra keys).

User Prompt:
Extract up to three contributions claimed in this paper. Return "contributions" with items that satisfy the rules above.

{user_content}

Figure 16: The prompt for contribution extraction.

 # Task Description
In an academic Survey generation task, we require an AI model to read a large number of papers and generate a "Taxonomy Tree" that summarizes the knowledge structure of the field.
Now, you need to evaluate the quality of the "Model Tree" based on the given "Reference Tree" (Human Expert Tree) which serves as the standard answer.

 # Input Data
Human Expert Tree (Reference):
<reference_tree>
{ground_truth}
</reference_tree>

Model Generated Tree (Model Prediction):
<model_tree>
{example_tree}
</model_tree>

 # Evaluation Criteria
Please compare the Model Tree with the Reference Tree based on the following four dimensions and score the Model Tree (1-5 scale).

<criteria_list>
1. Semantic Coverage & Recall
- Definition: Measures whether the Model Tree contains the core concepts and main branches present in the Reference Tree.
- Scoring Rubric:
- 1 (Critical Failure): Misses more than 50% of the core branches (Level 1/Level 2); key concepts are seriously lacking.
- 2 (Poor): Covers the main fields but misses a large number of important sub-fields; or exhibits significant conceptual deviation.
- 3 (Fair): Recalls most Level 1 concepts, but falls short in depth or breadth in specific branches; contains moderate omissions.
- 4 (Good): Recalls the vast majority of core concepts (>90%); even if terminology differs, the semantics correspond to the Reference Tree.
- 5 (Excellent): Perfectly covers all conceptual levels of the Reference Tree without omission; or provides valuable supplements.

2. Sibling Organization (MECE Principle)
- Definition: Evaluates whether the set of child nodes under the same parent node follows the MECE principle.
- Scoring Rubric:
- 1 (Chaotic): Severe semantic overlap between sibling nodes (>50%); or completely lacks classification logic.
- 2 (Poor): Inconsistent classification standards; or the division of a certain category is overly fragmented.
- 3 (Fair): Overall classification is acceptable, but there are fuzzy boundaries or mutual exclusivity is not strict enough.
- 4 (Clear): Clear boundaries between sibling nodes with good mutual exclusivity; classification logic is highly similar to Reference.
- 5 (Precise): Node organization is extremely rigorous; classification dimensions are unified and complete. **3. Hierarchical Consistency**
- Definition: Evaluates the logical correctness of the "Parent Node → Child Node" path (Is-A or Part-Of relationship).
- Scoring Rubric:
- 1 (Logical Error): Contains a large number of "inverted" relationships or severe hallucination.
- 2 (Hierarchical Confusion): Frequent misalignment of abstract levels; or forced parent-child relationships.
- 3 (Basic Flow): Most paths are logically valid, but there are minor issues with hierarchical definitions in deeper nodes.
- 4 (Good Logic): All parent-child relationships are academically valid; consistency of abstract levels is good.
- 5 (Rigorous Logic): All paths conform to strict academic definitions; matches the logical depth of the Reference Tree.

4. Structural Topology
- Definition: Evaluates whether the "shape" of the Model Tree is similar to the Reference Tree.
- Scoring Rubric:
- 1 (Severe Deformation): Extreme structural difference (e.g., Reference Tree is deep, but Model Tree is a flat list).
- 2 (Imbalanced): Certain branches are overly expanded while others are not, causing center of gravity deviation.
- 3 (Acceptable): Overall shape is roughly similar, but granularity in certain sub-trees is too fine or too coarse.
- 4 (Approximate): The overall depth and the lushness of various branches are consistent with the Reference Tree.
- 5 (Structural Fit): Perfectly replicates the granularity distribution and cognitive complexity.
</criteria_list>

Figure 17: The prompt for LLM-as-Judge evaluation on taxonomy trees.

 # Instructions
Your task is to strictly compare the <model_tree> with the <reference_tree> based on each dimension in <criteria_list>.
For each dimension, you need to:
1. Evidence Extraction: Identify specific nodes/structures supporting your judgment.
2. Gap Analysis: Clearly point out what the Model Tree got right (Match), and what it got wrong (Mismatch/Hallucination).
3. Final Scoring: Provide an objective score (1-5) based on your analysis.

 # Output Format Requirements
Please strictly follow the <output_format> below. Do not include any irrelevant intro or summary. Ensure the output is valid JSON.

```
<output_format>
{{
"semantic_coverage": {{
"score": [Specific Score 1-5],
"reasoning": "Detailed analysis of Semantic Coverage..."
}},
"sibling_organization": {{ ... }},
"hierarchical_logic": {{ ... }},
"structural_topology": {{ ... }}
}}
</output_format>
```
Now, please begin the evaluation.

Figure 17: The prompt for LLM-as-Judge evaluation on taxonomy trees. (Continued)

P LIST OF EVALUATION MODELS

We evaluate seven Deep Research Agents end to end. In Bottom-Up mode, we run 16 current-generation configurations across eight models and additionally evaluate the preceding generation as a longitudinal control. App. H lists the current endpoints and reasoning settings, and App. I reports the previous-generation results.

DEEP RESEARCH AGENTS

- **o3** (El-Kishky et al., 2025): OpenAI's Deep Research interface, with web search and multi-step tool use.
- **Doubao Deep Research** (ByteDance, 2025): ByteDance's research interface, with iterative search and tool use.
- **DeepSeek Search** (DeepSeek, 2025): DeepSeek's model with browser-plugin search.
- **Gemini Deep Research** (Google, 2025b): Google's research interface, with multi-step planning and public-web search.
- **Grok DeepSearch** (xAI, 2025): xAI's search-based research interface.
- **Perplexity Deep Research** (Perplexity, 2025): Perplexity's iterative search-and-report interface.
- **Qwen-Deep-Research** (Alibaba Cloud, 2025a): Alibaba's research interface, with scope clarification and iterative search.

CURRENT-GENERATION BOTTOM-UP MODELS

- **Claude-Sonnet-5** (Anthropic, 2026): `claude-sonnet-5`, queried at `minimal` and `high` effort.
- **GPT-5.6-sol** (OpenAI, 2026): `gpt-5.6-sol`, queried at `none` and `high` effort.
- **Gemini-3.1-Pro** (Google, 2026): `gemini-3.1-pro-preview`, queried at `minimal` and `high` effort.
- **DeepSeek-V4-Pro** (DeepSeek, 2026): `deepseek-v4-pro`, queried at `minimal` and `high` effort.
- **Qwen3.8-Max** (Alibaba Cloud, 2026): `qwen3.8-max-preview`, queried at `minimal` and `high` effort.
- **Kimi-K3** (Moonshot AI, 2026): `kimi-k3`, queried at `minimal` and `high` effort.
- **Grok-4** (xAI, 2026): `grok-4-latest`, queried at `minimal` and `high` effort.
- **GPT-5.5** (OpenAI, 2026): `gpt-5.5`, included as a within-family control at `none` and `high` effort.

PREVIOUS-GENERATION BOTTOM-UP MODELS

- **Claude-4.5-Sonnet** (Anthropic, 2025): `Claude-4.5-Sonnet-0929`, run in standard and extended-thinking modes.
- **GPT-5** (OpenAI, 2025a): run in standard and thinking modes.
- **Gemini-3-Pro-Preview** (Google, 2025a): standard and nominal thinking settings. The endpoint ignored the thinking flag, so the two outputs are not distinct conditions.
- **DeepSeek-V3.2** (Liu et al., 2025): run in standard and thinking modes.
- **Qwen3-Max-Preview** (Alibaba Cloud, 2025b): run in standard and thinking modes.
- **Kimi-K2** (Moonshot AI, 2025): `Kimi-K2-0905`, run in standard and thinking modes.

Q LIMITATIONS

We discuss limitations along several dimensions, indicating where each is mitigated empirically and where it remains an open direction.

Alignment vs. capability. In Findings 3–5, SEM-PATH and US-TED/US-NTED measure agreement with the selected expert taxonomy, not the absolute quality of a model taxonomy. Section 3.4 and Finding 6 define the alignment-based and capability-based diagnostics used throughout the paper. The judge-free statistics in Finding 6 use scalar expert properties at level L2, including the 0.89–1.87-level depth deficit and the singleton rate. They still compare model shape with expert shape and do not assess a model taxonomy on its own terms. A reference-free study with expert judgments at scale remains future work. Until then, our results show that model taxonomies differ from the selected expert references, not that every difference is incorrect. The LLM-judged defect rates also fail an expert control and are not used as capability evidence. Details are in App. D.

Prompt anchoring of hierarchy depth. A possible alternative explanation for shallow model trees is anchoring on our three-level prompt exemplar. App. F tests this directly: three of four models reach 4.0–4.9 levels when given an explicit depth target, so the deficit is a behavioral regularity rather than a capability ceiling, and Finding 6 is stated accordingly. The same probe shows that closing the depth gap makes every alignment measure worse, so the benchmark's headroom is not an artifact of the exemplar. The residual caveat is that the probe covers four models in one input condition. Qwen3-Max barely responds to either manipulation, so we cannot distinguish whether that model will not or cannot follow the depth instruction. A wider sweep over models and granularities is needed to estimate how common this case is.

Single-reference evaluation and taxonomy subjectivity. Each survey contributes one published expert taxonomy; multiple equally valid expert taxonomies may exist for the same topic. Rather than assume this subjectivity is small, we measure it. Two annotators independently built taxonomies for the same 10 surveys and agree with each other at 50.56% SEM-PATH and 22.92% ARI, while both score above 52% SEM-PATH against the expert and every current-generation model remains near the 27.49% FLAT floor (App. K.4). Two consequences follow. (i) Models reach only 0.56× the annotator-agreement level on SEM-PATH and sit 2.4× further from the expert on US-NTED than the annotators sit from each other, so models are not simply exploring equally valid alternatives. (ii) On ARI the same measurement disqualifies the metric for this purpose: models agree with the expert more than the annotators agree with each other, so we do not draw organization conclusions from ARI. The judge-free structural statistics of Finding 6 are unaffected either way, since they consume only scalar shape properties of the reference (level L2 of Section 3.4), never its labels or topology. A multi-reference extension analogous to multi-reference BLEU, where multiple expert taxonomies per topic yield a richer alignment target, is a natural and important direction for future work.

Annotation reliability and human-baseline coverage. We quantify the reliability of two human components only indirectly. First, we do not report inter-annotator agreement for paper-to-category mapping. Annotators transcribe the published hierarchy and follow paper placements in the source text, which limits their choices but does not measure agreement. The multi-category ablation in App. L.8 provides the closest quantitative check. We consider a double-annotated subset with percent agreement or κ the highest-priority addition to a future release. Second, the from-scratch human baseline covers 10 of 72 surveys under the same Bottom-Up requirements as models, each annotated independently by two people. Details are in App. K.4. It establishes a large matched gap on SEM-PATH, depth-matched SEM-PATH, and depth with bootstrap CIs, and it reports inter-annotator agreement, but two annotators do not support a variance estimate over annotators and 10 surveys are not a full-benchmark point estimate.

Domain scope. The benchmark is constructed from 72 highly cited LLM-related surveys. While these span eight subdomains within AI/ML, the absolute numerical findings may not transfer unchanged to fields with different citation cultures, such as biology, physics, or social sciences. As preliminary cross-domain evidence, we evaluate Claude-4.5-Sonnet on two interdisciplinary surveys spanning finance and medical domains and observe the same qualitative pattern of high homogeneity, moderate ARI, and SEM-PATH $\approx 30\%$. Details are in App. L.9. Extending TAXOBENCH to non-AI/ML disciplines is straightforward in principle (our four-stage pipeline and metric definitions are domain-agnostic), and doing so will broaden the empirical scope of the dual-bottleneck claim.

Evaluation workflow simplification. We evaluate Deep Research Agents in their native end-to-end product interfaces under Deep Research mode and frontier LLMs under a one-shot Bottom-Up

prompt; we do not engineer richer agentic workflows with explicit memory, iterative refinement, or human-in-the-loop control. Our results should therefore be interpreted as bounding the capability of agents *as currently deployed*, not as the asymptotic capability of LLMs under arbitrarily sophisticated scaffolding. TAXOBENCH's metrics directly support evaluating richer workflows in future work.

SEM-PATH lacks a chance correction. We identified our main methodological limitation by calibrating SEM-PATH. Because the metric has no chance correction, absolute values are uninterpretable without the floors in App. B, and in the shallow-tree regime that current models occupy its dynamic range collapses to about 1 pp. The level and spread carry different information: the low level relative to the human and synthetic controls is model behavior, whereas the narrow cross-model spread is a floor artifact rather than a discovered model property. The human–model gap on raw SEM-PATH is partly confounded with tree shape, though a $+13.27$ pp advantage remains after depth matching (App. K.4). The cross-generation comparison in App. I provides an independent check: a model generation that gains 3.68 pp ARI moves raw SEM-PATH by 0.87 pp, so the insensitivity is not an artifact of synthetic baselines. Our depth-matched variant restores a 9.45 pp spread on held-out configurations; its r=0.81 correlation with ARI is only an auxiliary check of recovered model-level variation. The variant remains a control rather than a principled fix because it repairs the confound after the fact instead of removing it from the metric's definition. Designing a chance-corrected hierarchy metric is the clearest open problem this work leaves behind, and we consider it more valuable than any additional model we could evaluate.

Encoder and judge dependency. US-TED, US-NTED, and SEM-PATH rely on a fixed sentence encoder, and the LLM-as-Judge component depends on GPT-4o with a fixed rubric. Absolute scores may shift with different encoders or judges. We validate that relative model rankings under US-TED and SEM-PATH are exactly preserved across three embedding models spanning two orders of magnitude, with Kendall's $\tau = 1.0$ in App. L.5, and that GPT-4o's rubric scores reach Cohen's $\kappa = 0.89$ with human evaluators in App. J. Rebuilding TAXOBENCH on top of newer judges or domain-specific encoders is a useful sanity check we recommend in future releases.