

MelonBench: A Benchmark for Coded Entity Resolution over Noisy Web Evidence

Zhishan Lin^{◊♠}, Jingxiao Li^{◊♠}, Yanbing Zhang^{◊♠}, Ziyang Liu[♠], Yuqing Shao[♠],
Yunfei Tong^{◊♠}, Ganghao Liu^{◊♠}, Qian Zhang^{◊♠*}, Zhe Wang^{◊♠*}

[◊]Key Laboratory of Smart Manufacturing in Energy Chemical Process,
Ministry of Education, China

[♠]Department of Computer Science and Engineering,
East China University of Science and Technology, China

Y30241066@mail.ecust.edu.cn, {qianzhang, wangzhe}@ecust.edu.cn

Abstract

Large language models (LLMs) are increasingly deployed as web-browsing agents, yet current benchmarks mostly evaluate explicit queries over clean factual evidence. We introduce MelonBench, a 665-item benchmark for coded entity resolution over noisy open-web user-generated content, instantiated with mid-to-long-tail public figures in Chinese entertainment blind-item discourse. Solving MelonBench requires agents to decode obfuscated clues, retrieve fragmented evidence, and identify the target entity. The benchmark is constructed through a five-stage semi-automatic pipeline with atomic event stratification, adversarial faithfulness auditing, and expert verification. Across 30 systems, the best non-search model reaches 25.1% accuracy and the best search-enabled system reaches 49.9%. Controlled comparisons of the same models with and without web search confirm consistent gains from web search, but also substantial remaining headroom. Diagnostic experiments identify semantic de-obfuscation and open-ended entity recall as key bottlenecks, and show that reasoning-centric search can outperform report-oriented agents. An 80-item finance pilot provides preliminary evidence of transfer beyond entertainment within a Chinese-language setting. MelonBench offers a challenging initial testbed for search agents under intentional obfuscation and noisy evidence. The dataset is publicly available at <https://huggingface.co/datasets/janehaha/MelonBench>.

1 Introduction

The rapid advancement of large language models (LLMs) enables agentic AI systems to autonomously browse the web (Schick et al., 2023; Yao et al., 2023; Qin et al., 2024). By accessing external information beyond static pre-training corpora, such systems can better handle tasks that

require retrieving long-tail knowledge (Wang et al., 2025b; Fan et al., 2024).

To assess how well LLMs use web browsing for complex information seeking, a growing number of benchmarks have been proposed (Mialon et al., 2024; Yoran et al., 2024; Liu et al., 2024; Zhou et al., 2024; Deng et al., 2023; Wei et al., 2025; Zhou et al., 2025). Among them, BrowseComp (Wei et al., 2025) and BrowseComp-ZH (Zhou et al., 2025) require multi-hop retrieval and cross-document synthesis, enabling fine-grained evaluation of search planning and reasoning. However, most existing benchmarks assume explicitly specified queries and relatively clean, authoritative evidence sources, as shown in Figure 1(a). They overlook a critical challenge in real-world web search: the ability to robustly retrieve and reason under intentional semantic obfuscation and noisy user-generated content (UGC).

In many real-world web environments, relevant evidence is fragmented, subjective UGC, with clues obfuscated through domain-specific slang and metaphors. A representative example appears in Chinese online entertainment communities, where **blind items** refer to entities through coded aliases. We use this setting as the first concrete instantiation of coded entity resolution, while not treating it as a culture-neutral proxy for all obfuscated search tasks. This requires agents to decode culturally grounded, semantically vague expressions, formulate effective search queries, and integrate fragmented, non-authoritative evidence that may be subjective or conflicting, making this a challenging end-to-end setting for evaluating search agents under obfuscation and noisy evidence.

To evaluate this setting, we introduce MelonBench, a 665-item benchmark for coded entity resolution. Unlike traditional entity linking, MelonBench provides no explicit mentions or candidate sets. Instead, systems must decode obfuscated clues, retrieve relevant web evidence,

*Corresponding authors.

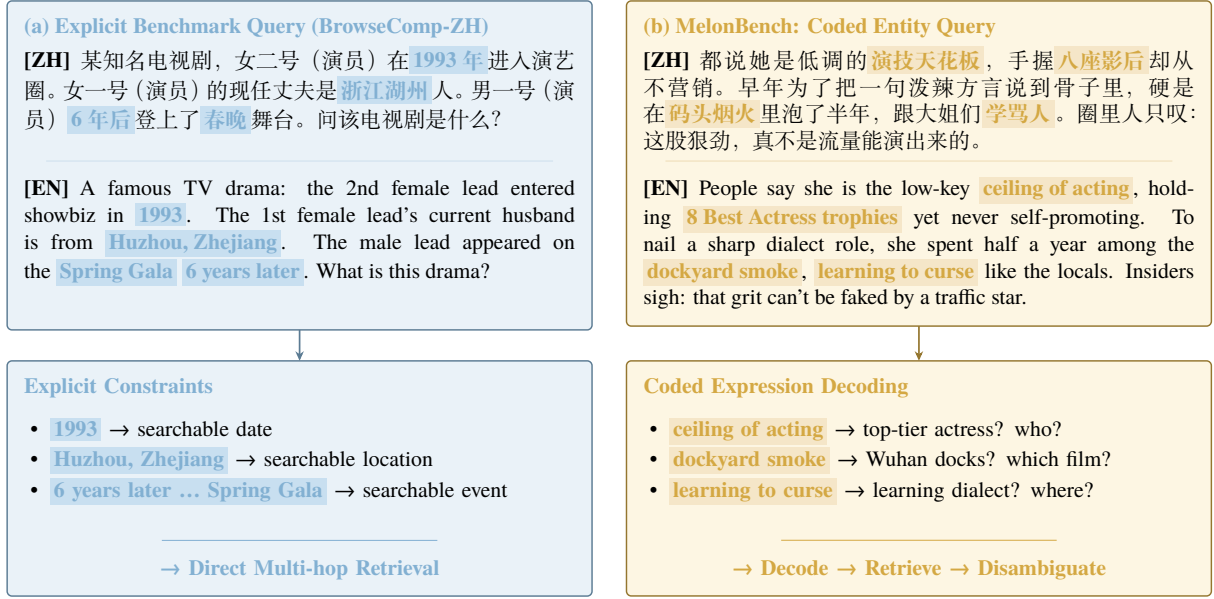


Figure 1: Comparison of explicit and coded benchmark queries. (a) BrowseComp-ZH uses directly searchable constraints for multi-hop retrieval. (b) MelonBench uses entertainment-industry jargon and cultural metaphors requiring decoding before retrieval.

and disambiguate the target entity, forming an end-to-end **decode** → **retrieve** → **disambiguate** pipeline (Figure 1(b)), with five additional translated examples in Appendix B. The benchmark is built through a semi-automatic pipeline combining atomic event stratification, adversarial faithfulness auditing, and expert verification. A later small finance pilot examines cross-domain adaptability, while broader validation remains future work.

We evaluate 30 systems spanning open-source, closed-source, and deployed AI-search products, and conduct ablation diagnostics on de-obfuscation, search lift, recognition–recall gaps, and reasoning-centric search behavior. Our key findings are:

- MelonBench remains far from solved: the best non-search model reaches only **25.1%**, while the best search-enabled system reaches **49.9%**; comparisons between corresponding search and no-search settings confirm that search helps but leaves substantial room for improvement.
- Semantic de-obfuscation is an independent bottleneck: before retrieval can succeed, systems must translate culturally grounded, obfuscated descriptions into effective search intents.
- Systems show much stronger recognition than recall: multiple-choice evaluation boosts accuracy, indicating that many failures stem from open-ended entity retrieval and production, not candidate verification alone.
- Reasoning-centric search can outperform report-

oriented agents on MelonBench, highlighting the importance of query interpretation and entity disambiguation beyond evidence aggregation.

In summary, our contributions are threefold:

- We introduce MelonBench, a novel benchmark evaluating end-to-end search-agent performance for long-tail knowledge under intentional semantic obfuscation and noisy open-web evidence.
- We design a **semi-automatic construction pipeline** based on atomic event stratification, adversarial faithfulness auditing, and expert verification, enabling controllable difficulty while reducing manual annotation burden without removing expert quality control.
- We conduct **comprehensive experiments** across 30 open-source, closed-source, and deployed AI-search systems, with diagnostics on semantic blur, multiple-choice recognition, search lift, thinking-mode trajectories, and a small cross-domain pilot. Results show persistent bottlenecks in semantic de-obfuscation, noisy evidence retrieval, and open-ended entity recall.

2 Related Work

Web Browsing Benchmarks. A growing line of work evaluates LLM-based web agents on increasingly complex tasks, ranging from general web task completion (Mialon et al., 2024; Liu et al., 2024) and autonomous navigation (Zhou et al., 2024; Deng et al., 2023) to realistic information-

seeking (Yoran et al., 2024). More recent benchmarks such as BrowseComp and BrowseComp-ZH (Wei et al., 2025; Zhou et al., 2025) move beyond traditional retrieval benchmarks (Joshi et al., 2017; Kwiatkowski et al., 2019; Yang et al., 2018; Wei et al., 2024; Mialon et al., 2024) that typically rely on structured, easily retrievable sources, instead challenging systems with multi-hop retrieval requiring cross-document synthesis. However, these settings still assume that key query constraints are explicitly stated and therefore directly searchable, and that relevant evidence mostly comes from relatively clean and authoritative sources. MelonBench differs by evaluating retrieval under **intentional semantic obfuscation** and **noisy open-web user-generated evidence**.

Entity Linking and Disambiguation. Entity linking maps textual mentions to knowledge-base entities (Shen et al., 2015; Wu et al., 2020), with recent generative variants (Cao et al., 2021). In contrast, MelonBench provides no explicit mention span: identifying information is hidden behind coded aliases, metaphors, and community-specific expressions, requiring systems to de-obfuscate the query, retrieve open-web evidence, and resolve the target entity without a predefined candidate set.

Retrieval over Noisy User-Generated Content. Retrieval-augmented generation improves factual grounding with external evidence (Fan et al., 2024; Vu et al., 2024; Lai et al., 2025), while online information studies show that UGC is often fragmented, subjective, and factually inconsistent (Vosoughi et al., 2018; Kumar and Shah, 2018; Sattarov and Choi, 2025). Yet retrieval benchmarks usually assume well-formed queries and clean corpora, and misinformation work focuses on truthfulness rather than entity resolution. MelonBench bridges these lines by jointly testing query-side obfuscation, noisy evidence retrieval, and final entity disambiguation.

3 Task Formulation and Domain Motivation

We formulate MelonBench as open-domain coded entity resolution. Given a blind-item query q that refers to its target through coded, anonymized, or metaphorical cues rather than explicit identifiers, the system must identify the single target entity e^* whose public web evidence satisfies q . Since queries are intentionally obfuscated and targets are

primarily mid-to-long-tail entities, solving the task requires open-web search and evidence synthesis beyond parametric memory. We evaluate entity-level accuracy against expert-verified gold answers.

Chinese entertainment blind-item discourse is the first MelonBench instantiation because it combines four requirements for coded entity resolution: rich organic coded language and obfuscation strategies, fragmented user-generated evidence, a measurable long-tail popularity structure, and publicly verifiable ground truth. We provide an expanded domain-motivation discussion in Appendix A. Our current benchmark focuses on mid-to-long-tail public figures selected under the 100K–1M Weibo-follower criterion and is not intended to be statistically representative of all Chinese entertainers or of the Chinese entertainment domain as a whole. These properties support a realistic yet controllable study of retrieval under semantic obfuscation. We view MelonBench as an initial instantiation of a broader coded-entity-resolution framework. The finance pilot in Section 5.4 provides preliminary evidence of analogous obfuscation patterns and adaptability beyond entertainment, while comprehensive cross-domain and cross-lingual validation remains future work.

4 MelonBench

4.1 Dataset Construction

We construct MelonBench with a five-stage semi-automatic pipeline (Figure 2) following the reverse-construction paradigm (Wei et al., 2025) (Appendix C): each instance starts from a verified target entity and is transformed into a coded blind-item query. The pipeline targets three desiderata: **solvability** from public web evidence, **difficulty** through semantic obfuscation rather than evidence sparsity, and **validity** through unique verifiability.

We construct the initial pool from two Wikipedia lists of Chinese male and female entertainers, then use identifiable Weibo accounts and 100K–1M follower counts only as eligibility filters, yielding 1,124 candidates without additional sampling (Appendix C.1). A sensitivity study supports this range as a practical balance between constructability and difficulty, yielding higher valid-item retention (59.2% vs. 22%/17% outside the range) while avoiding the substantially easier >1M regime (Appendix D.2).

We then retrieve supporting evidence from the broader open web via Baidu (Baidu, Inc., 2026).

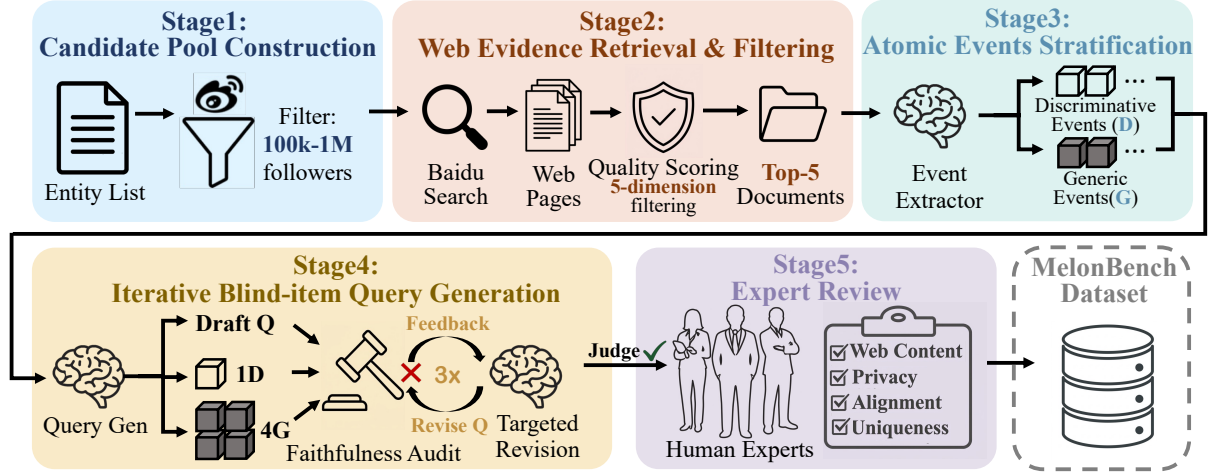


Figure 2: The five-stage semi-automatic construction pipeline. Starting from a filtered candidate pool of mid-to-long-tail entertainers, the pipeline retrieves and scores web evidence, extracts and stratifies atomic events into discriminative (D) and generic (G) categories, generates blind-item queries (Q) through an iterative adversarial audit loop, and concludes with expert review covering source quality, privacy, alignment, and answer uniqueness.

Retrieved pages are scored on relevance, information granularity, non-encyclopedic narrative, drama tension, and wiki-likeness. Those documents below the quality threshold are removed and the top five surviving pages are retained as the evidence basis for subsequent construction (Appendix C.2).

We then generate blind-item queries under **two-layer validity control**.

First, we stratify extracted atomic events into discriminative events, which sharply narrow the candidate space, and generic events, which are true but shared by many entities and primarily introduce ambiguity. Full extraction criteria and prompts are provided in Appendix C.3. Kimi-K2-Instruct then composes each anonymized query by weaving discriminative signal into generic distractors while avoiding directly searchable identifiers. Each draft passes through a faithfulness adversarial audit-revision loop, where an independent judge rejects identifier leakage, omitted clues, or unsupported details. We discard a query as construction-source leakage only when it preserves explicit identifiers or near-verbatim searchable phrases; retrieving original or related public evidence is part of the target search task, not leakage, and thus not grounds for discard. Every retained clue must trace to the selected source events (Appendix C.4).

Second, two domain experts with 10+ years of sustained engagement in Chinese entertainment UGC communities reviewed 771 candidates. It is a verification task rather than a puzzle-solving task: reviewers see the gold entity, retrieved evidence, atomic events, and candidate alternatives, and assess source quality, privacy, query-evidence

alignment, and answer uniqueness (Appendix C.5; see the full review protocol and acceptance criteria). The two experts reached 94.4% agreement ($\kappa=0.714$). Retaining only mutually accepted items yields the final 665-item benchmark.

4.2 Data Statistics

Of the 1,124 candidate entities retained after Stage-1 eligibility filtering (Appendix C.1), 459 (40.8%) are discarded in the subsequent construction stages, yielding the final 665-item benchmark. The main attrition causes are insufficient atomic evidence (26.6%) and expert review rejection (23.1%), showing that automated fact-sufficiency checks and human review provide substantial, non-redundant filtering. Full attrition breakdowns appear in Appendix D.

Figure 3 summarizes the final benchmark statistics. Follower counts are right-skewed and concentrated in the 100K–300K range, confirming that MelonBench targets mid-to-long-tail entities with enough web traces for retrieval but limited parametric memorability. Query lengths cluster around 60–80 characters, consistent with the short-form Chinese social-media style targeted by MelonBench. On average, the pipeline extracts 22 discriminative and 15 generic events per entity, with mean lengths of 63 and 26 characters.

5 Experiments

5.1 Experimental Setup

Models. We evaluate 30 systems spanning open-source models, proprietary models, and deployed

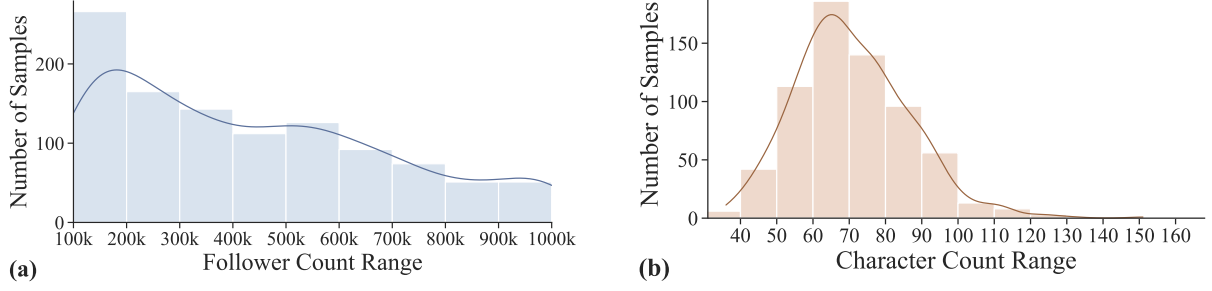


Figure 3: (a) Weibo follower distribution of target entities. (b) Query length distribution.

AI-search products, all listed in Table 1. Detailed model versions and search configurations are provided in Appendix E.1.

Inference and Browsing Setup. All systems receive the same coded query and return a predicted entity with a self-reported confidence score. For API-based models with configurable decoding parameters, we use *temperature* = 0.6, *top_p* = 0.95, and a maximum output length of 8,192 tokens; otherwise, we use the provider’s default settings. Search-enabled systems use their native browsing/search functionality, while no external retrieval is enabled for non-search settings. Commercial search products retain their native retrieval backends to evaluate end-to-end deployed systems. Full inference, decoding, and model-specific configurations are provided in Appendix E.

Evaluation. Each system output is converted into a predicted entity name and confidence score (prompt in Appendix I); for GUI-only AI search products, human annotators extract these fields. All systems use exact/normalized rule-based matching first, with unmatched responses judged by Gemini-3-Flash-Preview (Google, 2025a) for semantic equivalence (prompt in Appendix J.1). A stratified 200-case human audit yields 91.5% agreement, with disagreements confined to judge-positive cases, indicating slight over-permissiveness. Full evaluation details and the judge prompt are provided in Appendix J.2.

Temporal protocol. For data construction and search-augmented systems, evidence collection and evaluation were completed within February–March 2026; because exact reruns of commercial search products on the live web are not guaranteed, these scores are time-stamped baselines.

Metrics. We report **accuracy**, the fraction of correctly resolved entities, and **expected calibration error** (ECE), which measures alignment between predicted confidence and empirical accuracy over five equally spaced bins (Table 1; Appendix E.3).

To characterize external-evidence use, we also report paired search/no-search metrics on controlled subsets. For model m and domain d ,

$$\text{Lift}_{m,d} = \text{Acc}_{m,d}^{\text{search}} - \text{Acc}_{m,d}^{\text{no-search}} \quad (1)$$

and Normalized Search Lift (NSL), the fraction of remaining possible improvement achieved by search, is

$$\text{NSL}_{m,d} = \frac{\text{Acc}_{m,d}^{\text{search}} - \text{Acc}_{m,d}^{\text{no-search}}}{1 - \text{Acc}_{m,d}^{\text{no-search}}} \quad (2)$$

Accuracies lie in $[0, 1]$. With matched settings, no-search approximates the model’s parametric prior, while search measures use of external evidence. Since many deployed search products use proprietary retrieval backends and lack matched base models, Lift and NSL are controlled subset diagnostics rather than full-leaderboard metrics.

5.2 Overall Performance

MelonBench Remains Far From Solved. Table 1 summarizes all evaluated systems on MelonBench. The results reveal that MelonBench remains challenging under both settings. Among non-search configurations, the highest accuracy is only 25.1% (Gemini-3-Pro-Preview), while most models fall below 10%; gpt-oss-120b (0.2%) and Llama4 (0.6%) are near zero. Among search-enabled systems, Kimi (Deep Think) achieves the highest accuracy at 49.9%, leaving roughly half of the benchmark unsolved.

Calibration Reveals Unreliable Confidence. Beyond accuracy, Table 1 shows that many models are poorly calibrated. Several low-accuracy systems exhibit ECE of 66–86%, indicating confident but incorrect predictions. Accuracy and ECE are only moderately correlated (Pearson $r \approx -0.53$), confirming that ECE is not redundant with accuracy. Search generally improves calibration but does not eliminate it, and low ECE should not be

Category	Model	Rea.	Brow.	Acc	ECE (%)
Open-Source	DeepSeek-V3.2	N	N	5.4%	80
	DeepSeek-R1	Y	N	6.0%	79
	Qwen3.5-27B (Non-Thinking)	N	N	1.8%	82
	Qwen3.5-27B (Thinking)	Y	N	1.7%	82
	Qwen3.5-397B-A17B (Non-Thinking)	N	N	4.5%	86
	Qwen3.5-397B-A17B (Thinking)	Y	N	5.3%	82
	Kimi-K2.5 (Non-Thinking)	N	N	7.1%	78
	Kimi-K2.5 (Thinking)	Y	N	<u>18.0%</u>	68
	gpt-oss-120b	Y	N	0.2%	<u>66</u>
	Llama4	N	N	0.6%	79
Closed-Source	Qwen3.5-Plus	Y	N	8.0%	80
	Doubao-Seed-2.0-Pro (Non-Thinking)	N	N	5.9%	84
	Doubao-Seed-2.0-Pro (Thinking)	Y	N	15.6%	71
	GPT-5.2 (Non-Thinking)	N	N	2.6%	41
	GPT-5.2 (Thinking)	Y	N	9.2%	<u>37</u>
	Gemini-3-Flash-Preview	Y	N	17.6%	68
	Gemini-3-Pro-Preview	Y	N	<u>25.1%</u>	51
	Claude-Sonnet-4.6 (Non-Thinking)	N	N	2.4%	42
	Claude-Sonnet-4.6 (Thinking)	Y	N	4.4%	40
AI Search Product	OpenAI DeepResearch	–	Y	48.4%	26
	Grok (Deep Think)	–	Y	23.2%	65
	Doubao DeepResearch	–	Y	49.5%	37
	Doubao (Standard)	–	Y	20.5%	76
	Qwen DeepResearch	–	Y	22.1%	58
	Qwen (Deep Think)	–	Y	34.3%	48
	Kimi (Deep Think)	–	Y	49.9%	37
	Yuanbao (Standard)	–	Y	9.8%	<u>25</u>
	Yuanbao (Deep Think)	–	Y	20.0%	68
	DeepSeek (Standard)	–	Y	14.1%	70
	DeepSeek (Deep Think)	–	Y	20.0%	59

Table 1: Overall performance of all evaluated systems on MeIonBench. Underlined scores denote the best-performing system within each category, while boldfaced scores denote the best result overall across all systems.

interpreted as competence, since conservative systems may be calibrated while remaining inaccurate. We therefore use ECE as a complementary diagnostic of uncertainty awareness rather than as the primary ranking criterion.

Training Coverage May Confound Raw Accuracy. Among non-search models, Kimi-K2.5 and Doubao-Seed-2.0-Pro tend to outperform several other systems. We hypothesize that this pattern is partly explained by differences in Chinese-language and Chinese entertainment/UGC training coverage. This suggests that raw accuracy may conflate coded-entity-resolution capability with parametric prior, where higher scores may reflect prior exposure to relevant entity-clue associations, whereas lower scores may reflect limited domain coverage. In a fixed-evidence control on all 665 queries, we disable browsing for four representative models and provide either the five construction evidence pages (ORACLE-5) or those pages plus 15 same-domain distractors (NOISY-20). Accuracy rises from 5.4-18.0% without evidence to 96.84-98.50% under ORACLE-5 and remains 94.74-98.05% under NOISY-20 (Appendix E.5).

Pair	No-search	Search	Lift	NSL
K2.5-Think → DT	18.0%	49.9%	+31.9	38.9%
Q3.5-Plus → DT	8.0%	34.3%	+26.3	28.6%
Q3.5-Plus → DR	8.0%	22.1%	+14.1	15.3%
DS-V3.2 → Std	5.4%	14.1%	+8.7	9.2%
DS-R1 → DT	6.0%	20.0%	+14.0	14.9%

Table 2: Matched search/no-search diagnostics for systems where both settings are available. K2.5 = Kimi-K2.5; Q3.5 = Qwen3.5; DS = DeepSeek; DT = Deep Think; DR = DeepResearch; Std = Standard search.

Thus, the fixed-evidence control indicates that evidence acquisition is a major differentiating factor across models: once the same relevant evidence is provided, cross-model gaps largely disappear. Nevertheless, prior knowledge may inevitably have a modest effect. Accordingly, we interpret raw leaderboard accuracy as a mixed signal of prior exposure, de-obfuscation ability, retrieval effectiveness, and entity production, and use matched search/no-search diagnostics to better assess the contribution of external evidence.

Search Helps but Remains Unsaturated. Because parametric prior can confound raw accuracy, Table 1 alone cannot cleanly isolate the ef-

fect of web search. We therefore compute Search Lift and Normalized Search Lift (NSL) only for systems with matched search and no-search settings (Table 2); since many deployed search products do not expose their underlying models or matched no-search counterparts, these metrics serve as controlled subset diagnostics rather than full-leaderboard measures. Within this subset, low no-search accuracy and consistent search gains show that prior knowledge alone is insufficient and many MelonBench questions require external evidence. Moreover, stronger search-enabled systems obtain larger Search Lift and NSL relative to their own no-search baselines, suggesting that raw accuracy, despite prior-exposure confounds and limited matched coverage, can still partially reflect end-to-end search capability beyond memorization. Nevertheless, current search-agent pipelines remain far from saturated: even the best search-enabled system reaches only 49.9%, and the largest NSL recovers only part of the remaining headroom. These results reveal persistent bottlenecks in the decode–retrieve–disambiguate pipeline, where success depends not merely on retrieval access but also on search planning, obfuscated-clue interpretation, fragmented-evidence integration, and final entity disambiguation, as further supported by the stronger performance of multi-step or reasoning-centric search modes.

Effect of Reasoning. Enabling thinking mode generally improves accuracy, with the most pronounced gains in Kimi-K2.5 (+10.9%), Doubao-Seed-2.0-Pro (+9.7%), and GPT-5.2 (+6.6%). In contrast, Qwen3.5-72B and Claude-Sonnet-4.6 show only marginal improvement, suggesting that reasoning alone is insufficient without adequate domain knowledge or external evidence.

Analysis of 20 search-augmented paired cases (Appendix K) shows that thinking mode moves beyond surface-form matching toward hypothesis-driven search, rewriting obfuscated clues into searchable hypotheses, revising search directions when early results are incomplete, and verifying candidates against the full clue set. This suggests thinking mode helps by improving query reformulation, hypothesis revision, and constraint-based elimination, rather than by producing longer responses.

Reasoning-centric pipelines can outperform report-oriented ones. Qwen’s Deep Think mode (34.3%) outperforms its DeepResearch variant (22.1%), and Kimi (Deep Think) achieves the

overall best accuracy (49.9%). This may reflect a design mismatch: DeepResearch systems are optimized for synthesizing long-form reports, whereas MelonBench demands deep semantic reasoning over culturally loaded expressions, favoring reasoning-centric modes.

Human performance remains near floor. To contextualize these results, we first conduct a non-expert human evaluation in which 10 participants attempt 100 random items with web search and no time limit, reaching 7% average accuracy. We then evaluate two experts with over 10 years of experience in Chinese entertainment UGC under the same unrestricted web-search setting (Appendix E.4). Their 19% average accuracy is higher than the 7% non-expert result but still low in absolute terms, reflecting the substantially greater difficulty of open-ended entity resolution than the evidence-conditioned verification performed by a separate expert pair in Appendix C.5. Post-task interviews suggest that even for experts deeply familiar with Chinese entertainment, the sampled targets are relatively long-tail and unfamiliar; solving also requires cross-checking fragmented, noisy, and conflicting clues, causing heavy cognitive load.

Human failure does not imply item unanswerability. We analyze the 50 items attempted by the domain experts and compare them with Kimi (Deep Think). Among these items, 6 were solved by at least one non-expert, another 7 were solved by an expert but not by the evaluated non-experts, and 37 were solved by neither evaluated human group. Kimi correctly resolves 6/6, 5/7, and 12/37 items in these three mutually exclusive groups, respectively. In particular, its 32.4% accuracy on the 37 human-unsolved items shows that failure by the evaluated humans is not sufficient evidence that an item is unanswerable. Together with the de-obfuscation ablation in Section 5.3, these results suggest that low human accuracy largely reflects the substantial decoding and retrieval burden of the task rather than a lack of evidence-grounded answers. Detailed statistics are provided in Appendix E.4.

5.3 Ablation Studies and Analysis

We conduct 6 diagnostic experiments to disentangle the main sources of difficulty in coded entity resolution. Complementing the human-conditioned analysis above, these experiments examine whether low end-to-end accuracy arises from intrinsic answerability limitations or from specific stages of the resolution pipeline. The event composition abla-

tion is based on 100 benchmark instances and evaluated using Kimi (Deep Think) only. The semantic de-obfuscation, question-format, and generation-source ablations are conducted on a 204-question subset with 4 representative settings: DeepSeek-V3.2 and Kimi-K2.5-Thinking, each with and without web search. Unless otherwise specified, each ablation changes only one factor while keeping the rest of the evaluation setup fixed.

Semantic de-obfuscation is an independent bottleneck. Figure 4(a) compares three progressively de-obfuscated query formats, from coded (original jargon) through plain (de-obfuscated, entity-masked) to atomic (explicit event statements). Performance improves consistently as semantic ambiguity is removed, with Kimi (Deep Think) rising from 50.0% to 66.2% to 96.6%. Without web search, all models remain below 18.1% across all three conditions, indicating that parametric knowledge alone is far from sufficient even when queries are made explicit. The near-ceiling 96.6% accuracy under explicit atomic statements further indicates that, for this 204-item diagnostic subset, most failures in the original coded setting arise before or during query interpretation and retrieval rather than from an absence of recoverable target information. Qualitative error analysis of a 50-item diagnostic sample (Appendix G.1) also reveals three recurring pre-retrieval errors, namely metaphor flattening, scene-to-event mapping failures, and nickname/generic-event drift, where coded clues fail to be converted into effective search intents. These results show that semantic de-obfuscation is itself a major bottleneck, independent of downstream retrieval.

Models exhibit strong recognition but critically weak recall. Figure 4(b) contrasts open-ended and five-option multiple-choice formats, with the gold answer paired with the four highest-voted incorrect entity predictions across the evaluated systems. Kimi (Deep Think) improves from 50.0% to 76.5%, while DeepSeek-V3.2 jumps from 15.2% to 55.4% (nearly 4 $\times$). Similar gains appear across all tested configurations, indicating that systems can recognize the correct entity among candidates but fail to recall it independently from noisy, fragmented open-web evidence. This suggests that MelonBench is not difficult simply because evidence is sparse or entities are obscure, but rather because the main bottleneck lies in the end-to-end process of retrieving, consolidating, and producing the correct entity name without external scaffold-

ing. We retain the open-ended format as default, since it better captures the real-world challenge of coded entity resolution under noisy evidence.

Event composition ablation. We further examine how the balance between discriminative and generic events affects difficulty. Figure 5 shows that discriminative events are necessary for solvability, while generic events act as effective distractors only when discriminative signal is limited. With no discriminative event, accuracy never exceeds 46.31%. With two discriminative events, the task becomes much easier, with accuracy consistently above 58%. Under the most informative regime for diagnostic evaluation, fixing one discriminative event and increasing generic events from 0 to 4 reduces accuracy from 66.5% to 47.75%. This supports our default choice of one discriminative with four generic events, which preserves answerability while maintaining realistic ambiguity.

Generation-source ablation. To rule out same-source bias from using Kimi-K2-Instruct (Kimi Team, 2025) as the default generator, we regenerate all 204 questions using Gemini-3-Flash-Preview. Results (Figure 4(c)) show that Kimi (Deep Think) maintains strong performance (68.6%), confirming that its advantage stems from genuinely superior retrieval and reasoning rather than distributional overlap. Meanwhile, all models achieve higher accuracy on Gemini-generated questions, suggesting that Gemini tends to produce more straightforward paraphrases that inadvertently lower difficulty.

We further report a popularity-stratified analysis in Appendix F, confirming that the benchmark targets a challenging mid-to-long-tail regime where resolution accuracy generally decreases for less prominent entities, while calibration does not simply track the same popularity trend. We additionally report a retrieval-free LLM-side augmentation study across several recent models in Appendix H, which shows model-dependent gains but suggests that prompt-side augmentation alone does not remove the core difficulty of coded entity resolution.

Taken together, these ablations show that MelonBench is challenging not because any single component fails in isolation, but because success requires the entire chain of de-obfuscation, retrieval, evidence integration, and open-ended entity recall. Together with the fixed-evidence control in Appendix E.5, these results suggest that clue decoding and evidence acquisition are major sources of end-to-end performance variation,

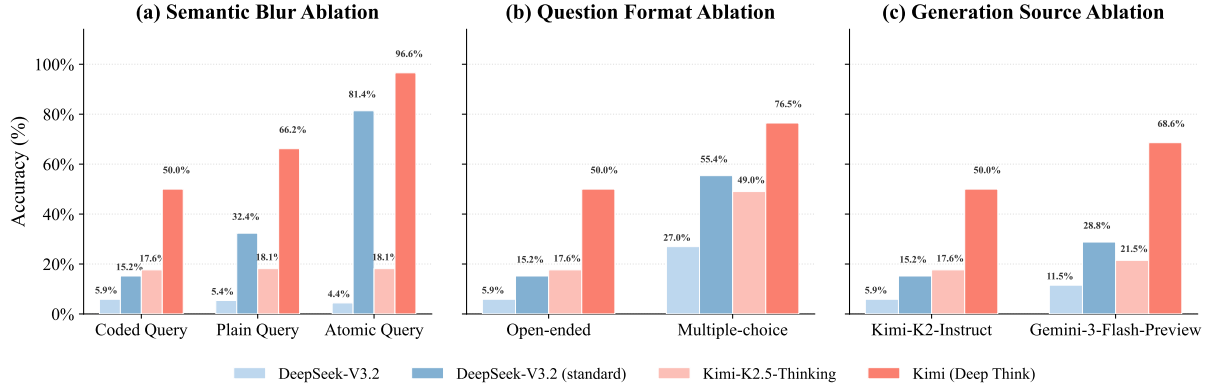


Figure 4: Three diagnostic ablations on MelonBench. (a) Semantic blur ablation (b) Question format ablation (c) Generation source ablation.

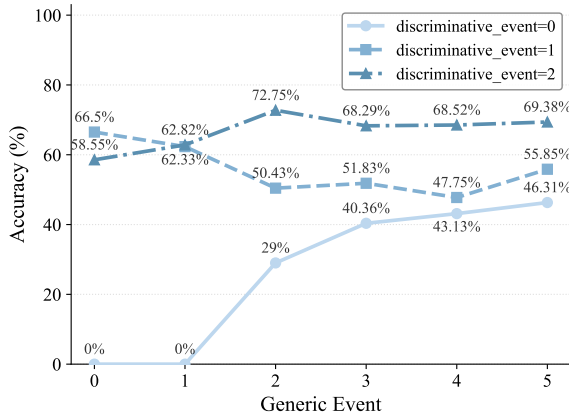


Figure 5: Accuracy under different event compositions. The x-axis denotes generic event count, and each line fixes the discriminative event count.

while parametric prior plays a more limited role. The two most critical bottlenecks are converting coded language into searchable intent and recalling the target entity without candidate assistance. The qualitative trajectories in Appendix K further connect these bottlenecks to thinking-mode behaviors such as query reformulation, search revision, and constraint-based candidate elimination.

5.4 Cross-Domain Pilot

To reduce entertainment-specific exposure bias, we construct an 80-item finance pilot using the same pipeline as MelonBench. It targets A-share listed firms and preserves the core task structure of obfuscated clues, fragmented open-web evidence, and entity-level answer matching. We adapt only domain-specific components, including entity pool, scoring dimensions, and prompt templates, to reflect finance-specific evidence.

Across shared model settings, no-search accuracy remains low while search consistently improves it. The best result reaches only 40.0%,

indicating the pilot is also unsaturated. These results provide preliminary evidence that the framework may transfer beyond entertainment within a Chinese-language setting, while broader cross-domain and cross-lingual validation remains future work. Details and full results appear in Appendix L.

6 Conclusion

We introduced MelonBench, a 665-item benchmark for coded entity resolution over noisy open-web user-generated content. Built through a 5-stage pipeline with adversarial auditing, atomic event stratification, and expert verification, MelonBench evaluates end-to-end search-agent performance under intentional semantic obfuscation. Across 30 evaluated systems, the benchmark remains far from solved: the strongest non-search model reaches only 25.1%, while the best search-enabled system reaches 49.9%. Controlled comparisons between search and no-search models show that enabling search improves accuracy, yet search-enabled performance remains far from saturated, indicating bottlenecks in the full decode–retrieve–disambiguate pipeline. Our ablations further identify semantic de-obfuscation and open-ended entity recall as key failure points. They show that reasoning-centric search can outperform report-oriented agents, suggesting that success on MelonBench depends not only on evidence aggregation but also on precise query interpretation and focused disambiguation. A small finance pilot suggests preliminary adaptability beyond entertainment within a Chinese-language setting, while broader cross-domain, cross-cultural, and cross-lingual validation remains future work. The dataset and associated evaluation resources are publicly available on [Hugging Face](#).

Limitations

This work has the following limitations:

Domain, Cultural, and Language Specificity.

MelonBench is instantiated primarily in Chinese entertainment blind-item discourse, and the finance pilot is also Chinese-language. Thus, our results do not establish generality across languages, cultures, domains, or other coded discourse communities. Such extensions are non-trivial because community-dependent components—including entity priors, evidence sources, obfuscation conventions, leakage-control rules, and expert validation—must be re-instantiated. The finance pilot provides only preliminary evidence of cross-domain feasibility; broader cross-domain, cross-cultural, and cross-lingual validation remains future work.

Scope of Entity Popularity. MelonBench focuses on public figures with 100K–1M Weibo followers, a range chosen to balance sufficient public evidence for construction and verification against the difficulty needed to probe genuine web-search ability. Although our popularity-stratified analysis supports follower count as a useful difficulty proxy within this range, the benchmark does not characterize behavior across the full popularity spectrum. Extending evaluation to both more obscure and more mainstream entities is therefore important future work.

Construction Pipeline and Reproducibility.

The benchmark relies on Baidu-based retrieval and several LLM-in-the-loop stages for filtering, generation, and auditing, so its data distribution may depend on the tools and models used (Wang et al., 2025a; Yuan et al., 2025). The audit enforces faithfulness and non-leakage but intentionally does not impose a fixed degree of semantic obfuscation; consequently, different generators may produce valid but stylistically different queries. Our generation-source ablation confirms such variation, and we use Kimi-K2 because it better preserves the intended blind-item obfuscation patterns. Moreover, live web pages and commercial search systems change over time. We release cached evidence snapshots, temporal metadata, and an archived local search corpus to support reproducibility, but exact reruns of closed commercial search products cannot be guaranteed; their reported scores should therefore be treated as time-stamped baselines.

Residual Evaluation Ambiguity. As a reverse-designed, open-ended benchmark grounded in noisy web evidence, MelonBench cannot guarantee perfect answer uniqueness. Multi-stage adversarial auditing and expert review substantially reduce ambiguity, but a small number of cases may still admit partially plausible alternatives. We therefore provide a multiple-choice variant as a complementary format for reducing open-ended ambiguity and separating recognition from recall. In addition, human validation indicates that the fallback semantic-equivalence judge is slightly over-permissive in some boundary cases, which may mildly overestimate absolute accuracy.

Ethical Considerations

MelonBench uses publicly accessible web evidence about public figures. Blind-item discourse may contain rumors or conflicting claims, so we use it only to model the *style* of semantic obfuscation rather than assuming its claims are reliable. Benchmark instances are instead grounded in publicly accessible evidence associated with a verified target identity. Expert review removes items involving invasion of privacy, explicit content, politically sensitive material, or plausible alternative answers. The two domain experts participated voluntarily as part of the research collaboration.

We also recognize that evaluating coded-reference resolution could improve systems’ ability to decode obfuscated online identities and therefore be misused for deanonymization. To mitigate this risk, the released benchmark is restricted to public figures already documented in public sources, and redistributed evidence snapshots come only from publicly accessible pages. MelonBench is intended strictly for evaluating retrieval and reasoning over noisy web evidence, and we discourage its use for real-world deanonymization or rumor verification beyond public-figure, public-interest contexts.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 62476087 and 62306115, and the Natural Science Foundation of Shanghai under the 2024 Shanghai Action Plan for Science, Technology and Innovation (No. 24ZR1416800).

References

- Anthropic. 2026. [Claude Sonnet 4.6](#).
- Baidu, Inc. 2026. [Baidu search engine](#).
- ByteDance. 2026a. [Doubao chat](#).
- ByteDance. 2026b. [Seed-2.0-pro](#).
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. [Autoregressive entity retrieval](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- DeepSeek-AI. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- DeepSeek-AI. 2025b. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *CoRR*, abs/2512.02556.
- DeepSeek-AI. 2026a. [Deepseek chat](#).
- DeepSeek-AI. 2026b. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *CoRR*, abs/2606.19348.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. [Mind2web: Towards a generalist agent for the web](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. [A survey on RAG meeting llms: Towards retrieval-augmented large language models](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM.
- Google. 2025a. [Gemini 3 Flash: Frontier intelligence built for speed](#).
- Google. 2025b. [A new era of intelligence with Gemini 3](#).
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.
- Kimi Team. 2025. [Kimi K2: open agentic intelligence](#). *CoRR*, abs/2507.20534.
- Kimi Team. 2026a. [Kimi chat](#).
- Kimi Team. 2026b. [Kimi-K2.6](#).
- Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. 2026. [Kimi k2. 5: Visual agentic intelligence](#). *arXiv preprint arXiv:2602.02276*.
- Srijan Kumar and Neil Shah. 2018. [False information on web and social media: A survey](#). *CoRR*, abs/1804.08559.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: a benchmark for question answering research](#). *Trans. Assoc. Comput. Linguistics*, 7:452–466.
- Hanyu Lai, Xiao Liu, Hao Yu, Yifan Xu, Iat Long Iong, Shuntian Yao, Aohan Zeng, Zhengxiao Du, Yuxiao Dong, and Jie Tang. 2025. [Webglm: Towards an efficient and reliable web-enhanced question-answering system](#). *ACM Trans. Inf. Syst.*, 43(5):122:1–122:43.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. [Agentbench: Evaluating llms as agents](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Meta. 2025. [Llama-4-maverick-17b-128e-instruct](#).
- Grégoire Mialon, Clémentine Fourier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. [GAIA: a benchmark for general AI assistants](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *CoRR*, abs/2508.10925.
- OpenAI. 2025a. [Introducing deep research](#).
- OpenAI. 2025b. [Introducing GPT-5.2](#).
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. [Toolllm: Facilitating large language models to master 16000+ real-world apis](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Qwen Team. 2026a. [Qwen chat](#).
- Qwen Team. 2026b. [Qwen3.5-27b](#).
- Qwen Team. 2026c. [Qwen3.5-397b-a17b](#).

- Qwen Team. 2026d. [Qwen3.5: Towards native multi-modal agents](#).
- Qwen Team. 2026e. [Qwen3.6-plus: Towards real world agents](#).
- Qwen Team. 2026f. [Qwen3.7: The agent frontier](#).
- Otabek Sattarov and Jaeyoung Choi. 2025. [Detection of rumors and their sources in social networks: A comprehensive survey](#). *IEEE Trans. Big Data*, 11(3):1528–1547.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Wei Shen, Jianyong Wang, and Jiawei Han. 2015. [Entity linking with a knowledge base: Issues, techniques, and solutions](#). *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460.
- Tencent. 2026. [Yuanbao chat](#).
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. 2024. [Freshllms: Refreshing large language models with search engine augmentation](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 13697–13720. Association for Computational Linguistics.
- Baiqiang Wang, Dongfang Zhao, Nathan R. Tallent, and Luanzheng Guo. 2025a. [On the reproducibility limitations of RAG systems](#). *CoRR*, abs/2509.18869.
- Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, Anningzhe Gao, and Benyou Wang. 2025b. [Huatuo-26m, a large-scale chinese medical QA dataset](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 3828–3848. Association for Computational Linguistics.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. [Measuring short-form factuality in large language models](#). *CoRR*, abs/2411.04368.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. [Browsecomp: A simple yet challenging benchmark for browsing agents](#). *CoRR*, abs/2504.12516.
- Wikipedia contributors. 2020. [Category:chinese male television actors](#).
- Wikipedia contributors. 2024. [Category:chinese television actresses](#).
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. [Scalable zero-shot entity linking with dense entity retrieval](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6397–6407. Association for Computational Linguistics.
- xAI. 2026. [Grok chat](#).
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. 2024. [Assistantbench: Can web agents solve realistic and time-consuming tasks?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8938–8968. Association for Computational Linguistics.
- Peiwen Yuan, Yiwei Li, Shaoxiong Feng, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. [Silencer: From discovery to mitigation of self-bias in llm-as-benchmark-generator](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. 2025. [Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese](#). *CoRR*, abs/2504.19314.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. [Webarena: A realistic web environment for building autonomous agents](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Domain Motivation: Why Chinese Entertainment?

As stated in Section 3, we adopt the Chinese entertainment industry as the initial instantiation domain for MelonBench. Below we elaborate on the four properties that make this domain especially suitable for benchmarking coded entity resolution over noisy, open-web user-generated content. We emphasize that these characteristics are not exclusive to entertainment: analogous obfuscation patterns arise in leaked-product speculation, political insider discourse, and corporate whistleblowing. We begin with entertainment because it offers the most favorable combination of data accessibility, long-tail granularity, and ground-truth verifiability, and view MelonBench as an initial instantiation of a broader evaluation paradigm that may extend to other settings with substantial domain- and language-specific adaptation.

Rich in Coded Language. Chinese entertainment discourse provides a uniquely fertile ground for studying coded entity resolution. Regulatory caution and the playful inventiveness of fan communities have together produced a vast, continuously evolving repertoire of obfuscation strategies—ranging from alphabetic initials and phonetic puns to euphemistic epithets and viral neologisms. This density and diversity make the domain an ideal source from which to draw realistic obfuscation patterns for benchmark construction.

Extreme UGC Fragmentation. Much of the information about entertainment figures is buried in interviews, user-generated posts, comment sections, and social-media threads rather than in well-structured knowledge bases. These sources are inherently fragmented, noisy, and riddled with factual inconsistencies and subjective opinions, closely mirroring the challenging retrieval conditions our benchmark is designed to evaluate.

Quantifiable Long-Tail Distribution. The entertainment industry exhibits a well-defined and externally measurable popularity hierarchy. In particular, Weibo follower counts serve as a reliable, publicly accessible proxy for entity notability, enabling us to precisely target the mid-to-long-tail band—a sweet spot where entities retain sufficient web presence to remain solvable, yet are obscure enough that parametric recall alone does not suffice.

Verifiability and Ethical Accessibility. The entertainment industry generates public records, such as filmographies, award histories, official announcements, and verified social-media profiles, even for mid-to-long-tail figures who receive little mainstream attention. This broad coverage means that both the target identity and the descriptive clues used in queries, whether explicit or implicit, factual or subjective, can be grounded in retrievable online evidence. The public nature of the source material also helps mitigate privacy and ethical concerns.

B Additional Translated Examples

To further illustrate the diversity of obfuscation strategies in MelonBench, we present five additional examples with English translations in Figures 6 and 7, complementing the annotated example in Figure 1. For each entry, we show the original coded query, its English translation, the gold answer, and key decodings that map coded expressions to their factual referents. Discriminative events are marked [D] and generic events [G].

C Details of Dataset Construction

This appendix provides full implementation details of the five-stage construction pipeline summarized in Section 4.1.

C.1 Stage 1: Candidate Pool Construction

We construct the initial entity inventory from two Wikipedia lists of Chinese male and female entertainers (Wikipedia contributors, 2020, 2024) covering actors, hosts, directors, and related public figures. For each listed individual, we identify the corresponding Weibo account and obtain its follower count. Individuals are excluded if no identifiable Weibo account can be found or if their follower count falls outside the predefined 100K-1M range. Weibo is therefore used as an eligibility filter rather than as the source of the initial candidate inventory. Applying these eligibility criteria leaves 1,124 candidate entities, which form the input to the subsequent web-evidence retrieval and benchmark-construction stages. No additional random or manual sampling is performed within this eligible pool. Accordingly, this purposive sampling frame is not intended to be statistically representative of all Chinese entertainers. The detailed motivation and sensitivity analysis for the 100K-1M follower range are provided in Appendix D.2.

C.2 Stage 2: Web Evidence Retrieval and Filtering

The five scoring criteria dimensions are defined as follows:

1. **Relevance Confidence** ($r \in [0, 1]$): The degree to which the document genuinely pertains to the target entity, serving as a soft gating weight.
2. **Information Granularity** ($g \in [1, 5]$): Presence of concrete details such as specific figures, lesser-known locations, or direct quotes, as opposed to vague adjectives.
3. **Drama Tension** ($d \in [1, 5]$): Presence of gossip value, controversy, emotional contrast, or narrative surprise.
4. **Non-Encyclopedic Narrative** ($n \in [1, 5]$): Degree of subjective, informal, or story-driven writing, e.g., interviews, behind-the-scenes accounts, fan commentaries.
5. **Wiki-Likeness** ($w \in [1, 5]$): Extent of formulaic, template-driven encyclopedic prose. This dimension is applied as a *penalty term*.

These individual scores are aggregated into a composite quality score:

$$S = r \cdot \max\left(0, \alpha g + \beta n + \gamma d - \mu w\right) \quad (3)$$

where $\alpha, \beta, \gamma, \mu$ are weight parameters set to 0.4, 0.3, 0.3, 0.2, respectively.

In practice, documents with $S < 3.0$ are discarded. The surviving documents are ranked by S in descending order, and the top-5 are retained as the evidence basis for all subsequent stages. Entities for which no document survives the threshold are excluded from the pipeline entirely.

The complete prompt template used for quality assessment is shown in Figure 8.

C.3 Stage 3: Atomic Event Extraction and Stratification

A key design choice in MelonBench is to stratify extracted events into two categories:

- **Discriminative events:** events that sharply narrow the candidate space and provide strong resolution signals for the target entity (e.g., "His 'plastic Mandarin' sparked complaints, prompting a 'warm-guy' re-recorded version.").

Example 1 (Actress) [ZH] 别急着对号入座，说的是那位电网螺丝钉——精神不稳的老妈早逝，首婚散，二婚守寡，被各路文工团集体退货二十多次后，硬闯中戏窄门，片场拼到导演喊停才停。 [EN] Don't rush to match this—it's about that power-grid cog: her mentally unstable mother died young, first marriage fell apart, widowed in her second. After being turned away by every performing troupe over twenty times, she forced her way through the Central Academy's narrow gate and pushed on set until the director called cut. Answer: 柏寒 (Bai Han) Key decodings: D “电网螺丝钉” (power-grid cog) → Hired as an electricity-meter worker at Beijing Power Bureau after dropping out at 16; “文工团集体退货二十多次” (rejected by every troupe 20+ times) → auditioned for performing-arts troupes over 20 times before entering the Central Academy of Drama G “精神不稳的老妈早逝” (unstable mother died young) → mother suffered from mental illness and passed away; “首婚散，二婚守寡” (first marriage collapsed, widowed in second) → first marriage ended in divorce; second husband died of illness	
Example 2 (Actor) [ZH] 都说京城土著 + 梨园世家是开挂开局，结果他把全部盘缠蒸发在北影厂门口，蹲活等到某部暗线大戏原定男主跑路，哆嗦一句结巴小反派被剪进正片——刚过不惑，零花边，心里只装着戏。 [EN] They say being a Beijing native from a Pear Garden dynasty is playing on easy mode, yet he burned through all his savings at the Beijing Film Studio gate, squatted for work until a spy-thriller's intended lead bailed, stammered his way into a minor villain role that made the final cut—just past forty, zero tabloid gossip, nothing but the craft on his mind. Answer: 曹炳琨 (Cao Bingkun) Key decodings: D “暗线大戏原定男主跑路” (spy-thriller's intended lead bailed) → the espionage drama <i>Lurk</i> (潜伏, 2008); he got the role after the original actor's schedule conflicted; “哆嗦一句结巴小反派被剪进正片” (stammered into a stuttering bit part that made the final cut) → he was nervous on set and stammered; the stammer was kept in the final cut and became the character's trademark trait G “梨园世家” (Pear Garden dynasty) → “Pear Garden” is the classical metonym for Chinese opera; his grandfather was a famous Peking opera performer; “盘缠蒸发在北影厂门口” (savings burned at studio gate) → lost his savings to a scam while waiting for roles outside Beijing Film Studio	
Example 3 (Singer) [ZH] 那位冰刀退役的主儿，嗓子一半磨砂一半烈酒，被三家大厂退货的 demo 在冰窖棚里冻出终极版，十位数播放把东北夜色灌进每个买醉人的耳机，旋律铺陈白山黑水。 [EN] That one who hung up his speed-skating blades—voice half sandpaper, half hard liquor. His demo, rejected by three major labels, got its definitive take in a freezing studio; ten-digit streams poured Northeast nights into every late-night drinker's earbuds, melodies painting the White Mountains and Black Waters. Answer: 侯旭 (Hou Xu) Key decodings: D “冰窖棚里冻出终极版” (definitive take in a freezing studio) → recorded the hit song <i>Escape</i> (逃) at −15°C in a broken-equipment studio, winter 2016; “十位数播放” (ten-digit streams) → surpassed 1 billion streams G “冰刀退役” (hung up his blades) → former professional speed skater who switched to music; “白山黑水” (White Mountains and Black Waters) → classical epithet for Northeast China (Manchuria)	

Figure 6: Translated examples 1–3 from MelonBench, illustrating metaphorical encoding (Example 1), cultural allusion (Example 2), and cross-domain metaphor (Example 3).

- **Generic events:** events that are plausible for many entities in the domain and therefore contribute ambiguity rather than direct identification (e.g., "Acting was mocked by viewers as

causing ‘cringe cancer’.”).

The complete extraction prompt and category assignment instructions are provided in Figure 9.

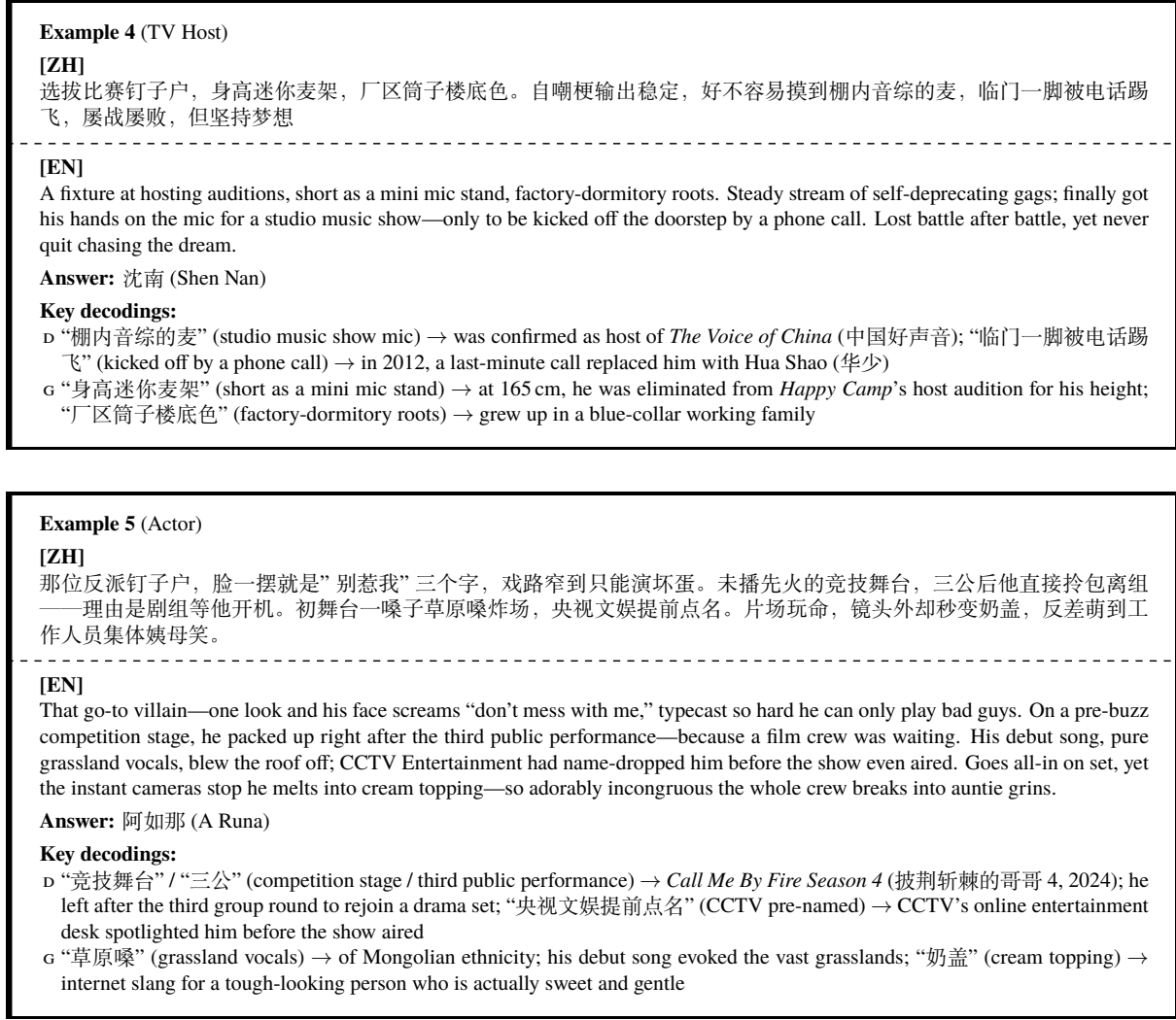


Figure 7: Translated examples 4–5 from MelonBench, illustrating industry insider language (Example 4) and variety-show coded terms with internet slang (Example 5).

C.4 Stage 4: Iterative Blind-Item Query Generation

An independent judge model checks whether the candidate query violates any of the following conditions:

- **Identifier leakage:** the query contains names, aliases, exact titles, dates, or other clues that make retrieval trivial;
- **Factual fabrication:** the query introduces content that is not supported by the selected source events;
- **Coverage failure:** the query omits essential source constraints or fails to faithfully represent the selected event set.

The completion generation prompt and audit prompt are respectively provided in Figure 10 and Figure 11.

In the adversarial revision loop, the generation prompt is augmented with two additional inputs: the previous blind-item query {query} and the structured audit feedback {review_opinion}. An additional constraint instructs the model to make targeted revisions addressing the specific violations identified by the auditor, rather than regenerating from scratch.

C.5 Stage 5: Expert Review Protocol

Stage-5 expert review is designed as an evidence-verification task rather than an open-ended puzzle-solving task. This distinction is important because the human baseline in Section 5.2 asks non-expert participants to solve blind items from scratch, whereas Stage-5 reviewers are shown the gold entity, candidate alternative entities, retrieved evidence pages, extracted atomic events, and the generated blind-item query. Their task is to verify

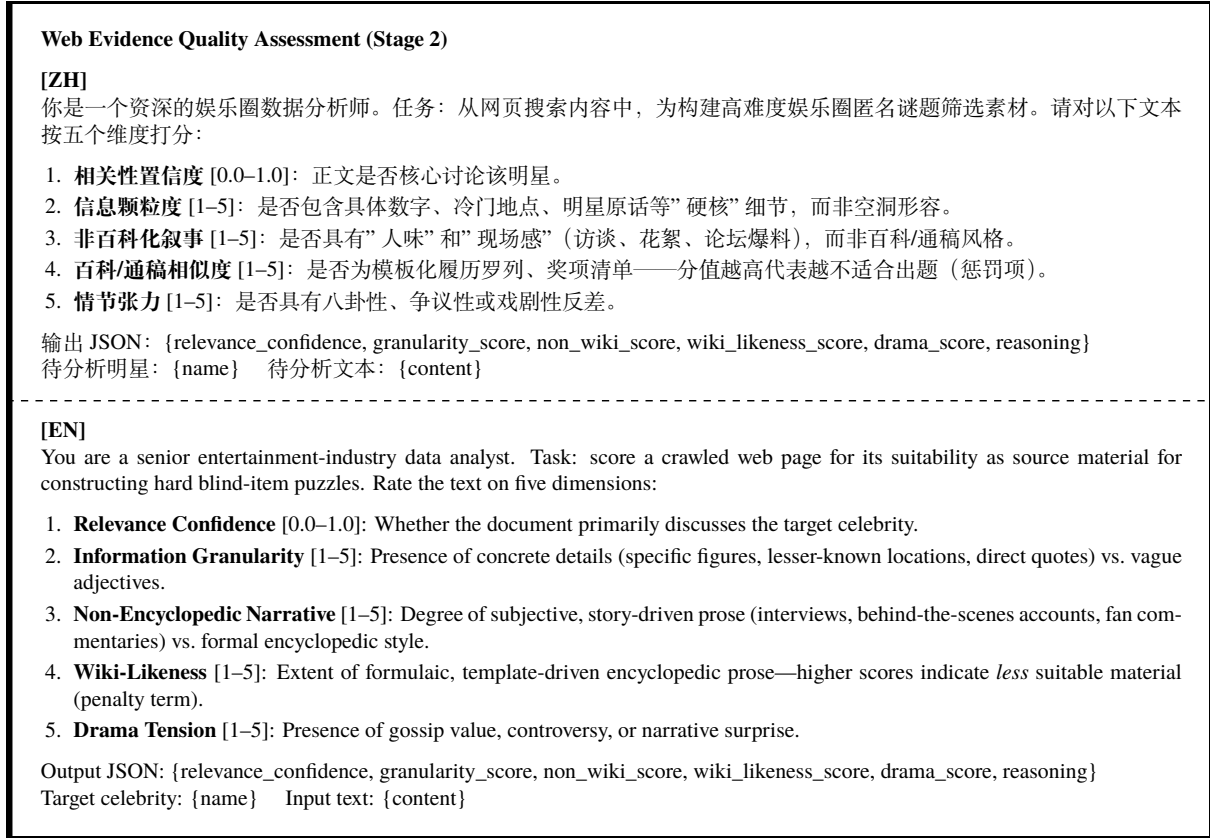


Figure 8: Prompt template for web evidence quality assessment (Stage 2). The evaluator scores each crawled document on five dimensions: relevance confidence, information granularity, non-encyclopedic narrative, wiki-likeness (penalty), and drama tension.

whether the gold entity is supported by the evidence and uniquely satisfies the explicit constraints encoded in the query.

Expert qualifications. We define domain expertise using task-relevant criteria: native-level Chinese proficiency, long-term familiarity with Chinese entertainment-community discourse, knowledge of blind-item slang and obfuscation conventions, and the ability to cross-check celebrity-related claims across heterogeneous web sources. Candidates were screened using authentic entertainment-community slang and blind-item questions drawn from Chinese social-media discourse. We recruited two independent domain experts who satisfied these criteria and had more than 10 years of sustained engagement with Chinese entertainment UGC communities. Here, “10+ years” refers to sustained participation in these online communities rather than professional employment in the entertainment industry. The experts were not involved in the automatic generation or adversarial auditing stages. The Stage-5 reviewers and the experts used for the human base-

line in Section 5.2 were independently recruited, non-overlapping groups, although both were selected using the same expertise criteria. To preserve anonymity under the double-blind review process, we describe their qualifications rather than disclosing identifying information.

Review inputs. For each candidate item, both experts independently received: (i) the gold target entity, (ii) the generated blind-item query, (iii) the extracted discriminative and generic atomic events, (iv) the crawled evidence pages and cached snippets, and (v) a set of candidate alternative answers produced by strong search agents. The experts were asked to make a binary accept/reject decision after checking whether the item satisfied all four criteria below.

All items surviving the automated pipeline undergo a comprehensive four-step human review:

1. **Source Quality Verification:** To prevent low-quality or irrelevant evidence from inducing hallucinated atomic events, reviewers inspect the original crawled web content and discard items with garbled text, encoding artifacts,

Atomic Event Extraction (Stage 3) [ZH] 任务：阅读输入文档，针对目标人物提取所有相关的“完整事件描述”或“属性特征”。 关键定义： 完整事件描述 (Holistic Event)：围绕一个核心主题的完整叙事单元。 聚合原则：【时间】+【核心动作】+【具体细节】+【结果/影响】须合并为一条，严禁将因果关系或从属关系拆解为多条。 唯一性事实 (Unique Facts)：具有高度辨识度、能将该人物与他人区分开的具体事件（含具体作品名、奖项、特殊经历等）。 非唯一性描述 (Non-unique Descriptions)：宽泛的标签、性格评价或模糊经历（如“拍戏很敬业”、“从小家境贫寒”）。 操作准则：(1) 聚合优先——因果链合成一条；(2) 指代消解——将“他/她”替换为人物姓名。 输出 JSON：{unique_facts: [...], non_unique_descriptions: [...]} 目标人物：{name} 输入文档：{content}	[EN] Task: Read the input document and extract all relevant “holistic event descriptions” or “attribute traits” for the target person. Key definitions: Holistic Event: A self-contained narrative unit around one core topic. Aggregation principle: [Time] + [Core action] + [Specific details] + [Outcome/Impact] must be merged into a single entry; causally or hierarchically dependent facts must never be split. Unique Facts: Highly discriminative events that distinguish the person from others (specific works, awards, exceptional experiences). Non-unique Descriptions: Broad labels, personality assessments, or vague experiences (e.g., “very dedicated on set”, “grew up in a poor family”). Guidelines: (1) Aggregate first—merge causal chains into one entry; (2) Resolve coreferences—replace all pronouns with the person’s name. Output JSON: {unique_facts: [...], non_unique_descriptions: [...]} Target person: {name} Input document: {content}
--	---

Figure 9: Prompt template for atomic event extraction (Stage 3). The LLM extracts holistic events from raw text and classifies each as a unique fact or non-unique description.

- inaccessible content, or material irrelevant to the target entity.

2. **Privacy and Sensitivity Screening:** To ensure ethical compliance, reviewers flag and remove items involving invasion of personal privacy, explicit content, or politically sensitive material. Items are retained only when the supporting evidence concerns public figures and publicly accessible, non-sensitive information.

3. **Query–Evidence Alignment Audit:** To guarantee faithfulness, reviewers trace every claim in the blind-item query back to its corresponding source atomic event. Minor wording mismatches are corrected in place when the intended meaning remains fully supported by evidence. Items with unsupported claims, missing selected clues, or fundamental grounding failures are discarded.

4. **Answer Uniqueness Validation:** To reduce evaluation ambiguity, we operationally define answer uniqueness as follows: under the released anonymized query and supporting evidence, exactly one target entity should satisfy all retained atomic constraints. Because
- open-web uniqueness cannot be proven mathematically, we adopt an adversarial human-in-the-loop validation protocol similar in spirit to Zhou et al. (2025). Multiple top-performing AI search agents, including OpenAI DeepResearch, Kimi (Deep Think), Doubao DeepResearch, and other strong browsing systems, independently generate candidate answers for each query. Annotators then manually inspect whether any alternative answer satisfies all constraints embedded in the query using cached evidence and open-web search. If any plausible alternative entity satisfies all retained clues, the item is deemed ambiguous and discarded.

Agreement and conservative filtering. Among 771 Stage-5 candidate items, the two experts achieved 94.4% raw agreement and a Cohen’s κ of 0.714 on binary accept/reject decisions before applying the final filtering rule. To minimize residual ambiguity, we retained only items accepted by both experts with respect to evidence support, query–evidence alignment, privacy/sensitivity screening, and answer uniqueness. This conservative consensus criterion yielded the final 665-item MelonBench benchmark; all other items were dis-



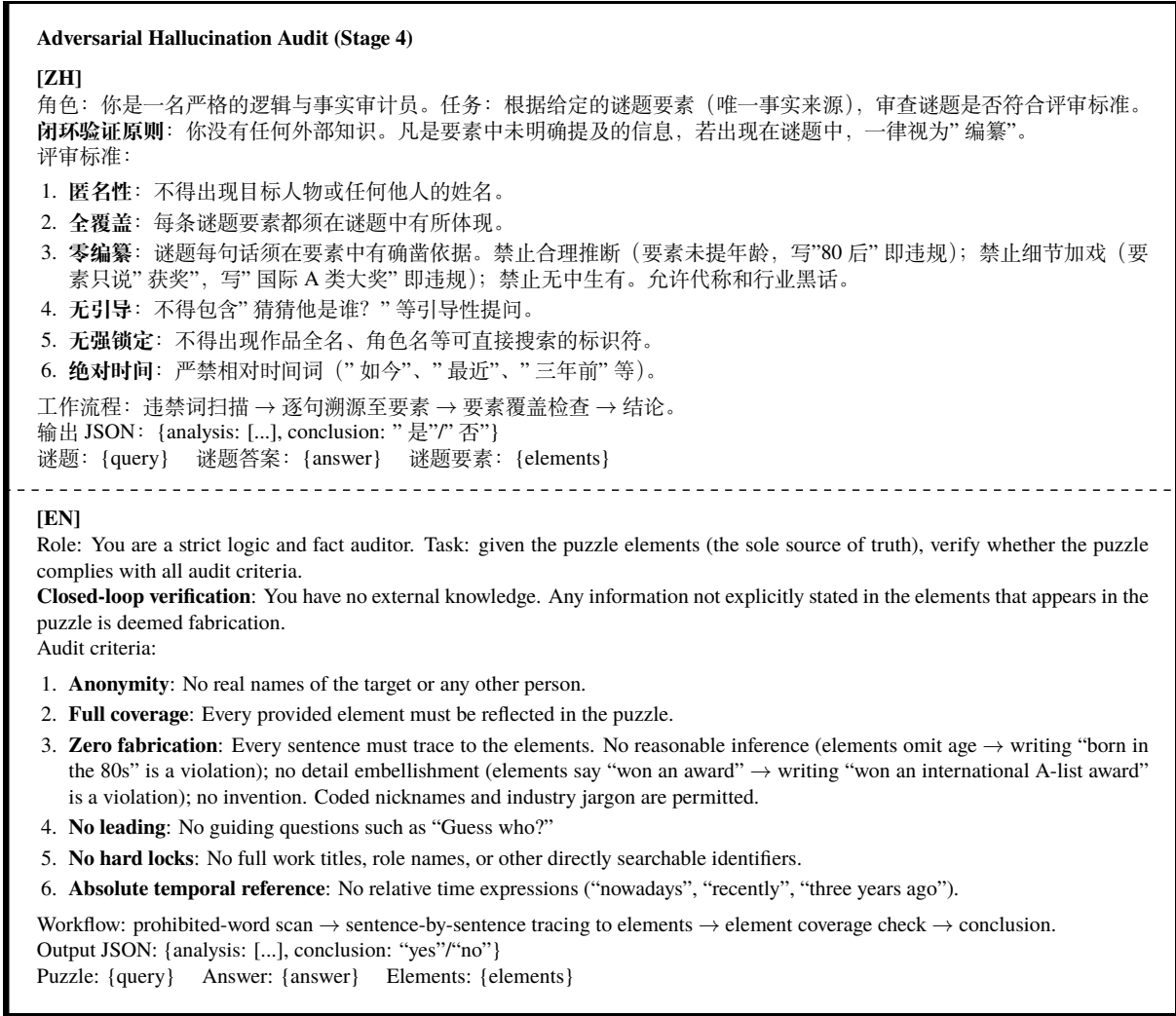


Figure 11: Prompt template for adversarial hallucination audit (Stage 4). Six audit criteria enforce anonymity, full element coverage, zero fabrication, no leading questions, no hard-lock identifiers, and absolute temporal references.

does not converge within 3 iterations—the generated query repeatedly fails one or more audit criteria despite targeted revision.

Two key observations emerge from this analysis. First, the two largest attrition categories—Insufficient Atomic Evidence (26.6%) and Expert Review Rejection (23.1%), together account for nearly half of all discards, confirming that both the automated fact-sufficiency threshold and the human quality gate provide substantial, non-redundant filtering. Second, the remaining four causes are distributed across Stages 2–4, confirming that quality control operates throughout the pipeline rather than relying on any single bottleneck. The balanced attrition profile suggests that no individual stage is overly aggressive or overly permissive, supporting the overall robustness of the construction pipeline.

D.2 Follower-Range Sensitivity Study

To empirically validate the 100K–1M Weibo-follower band, we sample 100 entities below 100K followers and 100 above 1M, then run the full construction pipeline under identical settings.

Valid-item retention drops sharply outside the selected range: only 22% of sub-100K entities and 17% of super-1M entities survive the pipeline, compared with 59.2% within the 100K–1M band. The two groups fail for different reasons. Sub-100K entities frequently lack reliable, discriminative public traces—their web presence is too sparse to extract sufficient atomic events or to support answer verification. Super-1M entities, by contrast, generate abundant search results, but the top-ranked pages are dominated by Weibo profiles, encyclopedic entries, and press-release-style news articles. These structured or templated sources score poorly under

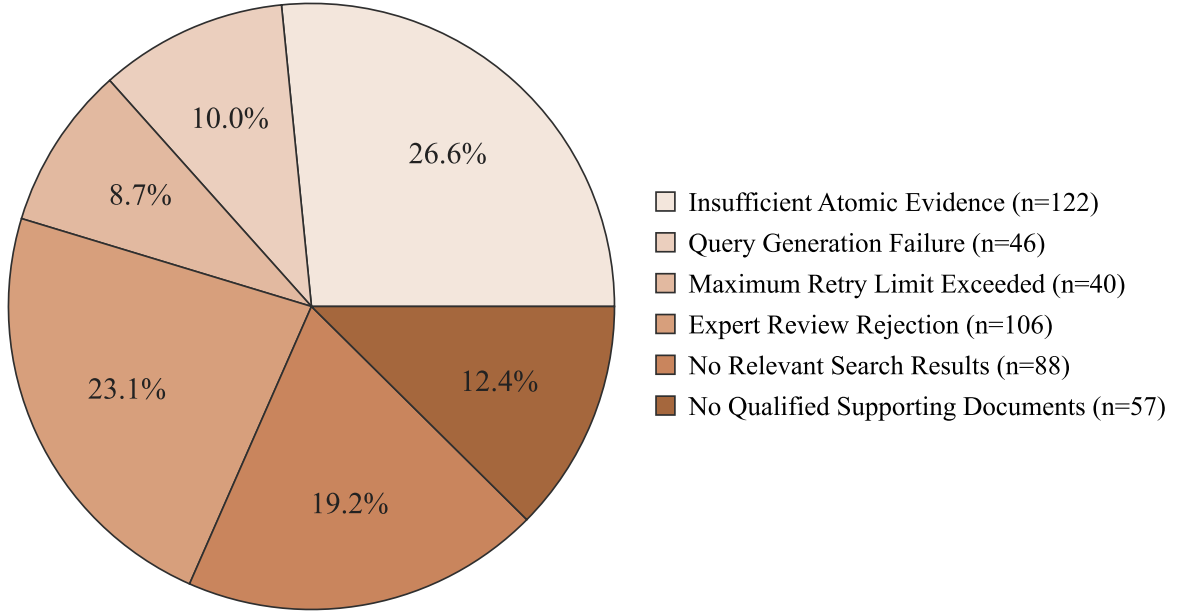


Figure 12: Distribution of pipeline attrition causes ($n = 459$).

the quality framework (Appendix C.2) and rarely provide the fine-grained, narratively expressed evidence needed for query construction.

We further evaluate Kimi Deep Think on the retained valid items from each group. On a 100-item MelonBench sample drawn from the default range, Deep Think achieves 54% accuracy; on surviving sub-100K items it scores 59.09%, and on surviving super-1M items it reaches 82.35%. The elevated accuracy for super-1M entities reflects their stronger representation in model memory and search indices, but it also undermines the intended mid-to-long-tail evaluation setting.

These results indicate that the 100K–1M range balances constructability, verifiability, and task difficulty: below 100K, evidence is too sparse to build valid items reliably; above 1M, evidence is abundant but dominated by templated sources, and the surviving items become too easy. Thus, the 100K–1M band is a task-design choice rather than a population-sampling criterion.

E Details of Experiments

E.1 Full Model List and Configurations

Below we list all evaluated systems with their configuration details. Their aggregate performance is reported in Table 1 in the main paper.

Browsing and search configuration. For systems marked Brow.=Y, we used the system’s native browsing/search functionality in its deployed product setting. Because the search indexes and retrieval pipelines of commercial AI search products are proprietary, we did not standardize the back-end retriever across systems. This choice follows our goal of evaluating end-to-end real-world AI search products. For systems marked Brow.=N, no external browsing or search tool was enabled. The Baidu Search component in the dataset construction pipeline was used only to collect and verify benchmark evidence and was not connected to any evaluated system.

Open-Source Models.

- **DeepSeek-V3.2** (DeepSeek-AI, 2025b): a non-reasoning model developed by DeepSeek.
- **DeepSeek-R1** (DeepSeek-AI, 2025a): a reasoning model developed by DeepSeek.
- **Qwen3.5-27B** (Qwen Team, 2026b): an open-source model from Alibaba, evaluated in both thinking and non-thinking modes.
- **Qwen3.5-397B-A17B** (Qwen Team, 2026c): an MoE model from Alibaba, evaluated in both thinking and non-thinking modes.
- **Kimi-K2.5** (Kimi Team et al., 2026): an open-source model from Moonshot AI, evaluated in both thinking and non-thinking modes.
- **gpt-oss-120b** (OpenAI, 2025): an open-

source reasoning model released by OpenAI.

- **Llama-4-Maverick-Instruct-Basic** (Meta, 2025): a non-reasoning model from Meta.

Closed-Source Models.

- **Qwen3.5-Plus** (Qwen Team, 2026d): a reasoning-enabled proprietary model from Alibaba.
- **Doubao-Seed-2.0-Pro** (ByteDance, 2026b): a proprietary model from ByteDance, evaluated in both thinking and non-thinking modes.
- **GPT-5.2** (OpenAI, 2025b): a proprietary model from OpenAI, evaluated in both thinking and non-thinking modes.
- **Gemini-3-Flash-Preview** (Google, 2025a): a proprietary reasoning model from Google.
- **Gemini-3-Pro-Preview** (Google, 2025b): a proprietary reasoning model from Google.
- **Claude-Sonnet-4.6** (Anthropic, 2026): a proprietary model from Anthropic, evaluated in both thinking and non-thinking modes.

AI Search Products.

- **OpenAI DeepResearch** (OpenAI, 2025a): a multi-step retrieval and report-generation system developed by OpenAI.
- **Grok (Deep Think)** (xAI, 2026): the reasoning-centric search mode of Grok, developed by xAI.
- **Doubao DeepResearch** (ByteDance, 2026a): the deep research mode of Doubao, developed by ByteDance.
- **Doubao (Standard)** (ByteDance, 2026a): the standard search mode of Doubao, developed by ByteDance.
- **Qwen DeepResearch** (Qwen Team, 2026a): the deep research mode of Qwen (Qwen3.5-Plus-basic), developed by Alibaba.
- **Qwen (Deep Think)** (Qwen Team, 2026a): the reasoning-centric search mode of Qwen (Qwen3.5-Plus-basic), developed by Alibaba.
- **Kimi (Deep Think)** (Kimi Team, 2026a): the reasoning-centric search mode of Kimi (Kimi-K2.5-Thinking-basic), developed by Moonshot AI.
- **Yuanbao (Standard)** (Tencent, 2026): the standard search mode of Yuanbao, powered by Tencent’s Hunyuan model.
- **Yuanbao (Deep Think)** (Tencent, 2026): the reasoning-centric search mode of Yuanbao, powered by Tencent’s Hunyuan model.

- **DeepSeek (Standard)** (DeepSeek-AI, 2026a): the standard search mode of DeepSeek (DeepSeek-V3.2-basic).
- **DeepSeek (Deep Think)** (DeepSeek-AI, 2026a): the reasoning-centric search mode of DeepSeek (DeepSeek-R1-basic).

E.2 Hyperparameter Settings

Decoding and Evaluation Details.

- **Non-configurable APIs.** For models whose decoding hyperparameters are not exposed to users (e.g., **GPT-5.2** and **Claude-Sonnet-4.6**), we use the default API settings.
- **Configurable APIs.** For all other models, we follow DeepSeek-AI (2025a) and set the decoding hyperparameters to *temperature* = 0.6 and *top_p* = 0.95.
- **Maximum output length.** We set the maximum generation length to 8,192 tokens for all API-based models to limit excessively long reasoning traces.
- **GUI-based systems.** For AI search products without API access, human annotators query each system through its graphical user interface and record the final outputs for evaluation.

E.3 Evaluation Metrics

Following Zhou et al. (2025), we report two complementary metrics.

Accuracy (Acc). Accuracy measures the fraction of queries for which the predicted entity exactly matches the gold-standard answer:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\hat{e}_i = e_i^*] \quad (4)$$

where $\hat{e}_i$ is the predicted entity and e_i^* is the ground-truth entity for the i -th query.

The grading prompt is provided in Appendix J.

Expected Calibration Error (ECE). ECE quantifies the alignment between a model’s self-reported confidence and its actual performance. We partition predicted confidence scores into $B = 5$ equally spaced bins: $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, and $[0.8, 1.0]$, and compute:

$$\text{ECE} = \sum_{i=1}^B \frac{n_i}{N} |\text{acc}(i) - \text{conf}(i)| \quad (5)$$

where n_i is the number of samples in the i -th bin, $\text{acc}(i)$ is the empirical accuracy of the i -th bin,

and $\text{conf}(i)$ is the average predicted confidence in that bin. A lower ECE indicates better-calibrated confidence estimates.

Since confidence scores are elicited from model outputs rather than model logits, we interpret ECE as a behavioral calibration diagnostic over self-reported confidence, not as a strict probabilistic calibration measure.

E.4 Human Baseline Protocol

Shared setup. Both non-expert and expert evaluations follow the same protocol. Each participant receives only the anonymized coded query; no gold answer, metadata, source evidence, or any other information is provided. Participants use Baidu as their search engine with unlimited queries and no time limit, continuing until they submit an answer or give up.

Non-expert evaluation. Ten participants without domain expertise each attempted 100 randomly sampled items, achieving 7% average accuracy.

Expert evaluation. We recruited a separate pair of domain experts using the same task-relevant expertise criteria as the Stage-5 review protocol in Appendix C.5. These experts were independently recruited and did not participate in dataset construction or verification. Each expert attempted the same 50-item subset of the same 100 items used in the non-expert evaluation. We use a reduced subset because each item required extensive cross-referencing of fragmented web evidence, making the task highly time-consuming for careful expert analysis.

Unlike Stage-5 evidence verification, this evaluation requires experts to identify the target entity from scratch: they must interpret the obfuscated clues, formulate effective search queries, retrieve and reconcile fragmented web evidence, and produce an answer without being shown the gold entity or candidate alternatives. Their average accuracy was 19%. Post-task interviews suggest that the difficulty reflects the benchmark’s intended long-tail and noisy-evidence setting. Even experts who closely follow Chinese entertainment found the sampled targets relatively unfamiliar. They also had to cross-check fragmented, noisy, and sometimes conflicting clues across open-web sources, making the task cognitively demanding.

Human–Model Solvability Analysis Since the expert evaluation covers only a 50-item subset of

Human outcome	#items	Kimi(Deep Think)
Solved by non-experts	6	100.0% (6/6)
Solved by ≥ 1 expert	13	84.6% (11/13)
Not solved by either expert	37	32.4% (12/37)
Overall	50	46.0% (23/50)

Table 3: Model performance conditioned on human success on the 50-item expert-evaluated subset.

the 100 non-expert-evaluated items, the human study does not provide a benchmark-wide partition of all 665 items by human solvability. We therefore use the overlapping 50 items for a diagnostic analysis of whether low human accuracy indicates unanswerability. Table 3 reports Kimi (Deep Think) performance conditioned on observed human success. Of the 13 items solved by at least one expert, Kimi correctly resolves 11 (84.6%). The 6 items solved by non-experts are a subset of these 13 expert-solved items, and Kimi resolves all 6. Among the remaining 37 items not solved by either evaluated expert, Kimi still correctly resolves 12 (32.4%). Overall, Kimi solves 23/50 (46.0%) items on this subset.

This result highlights an important asymmetry in interpreting human performance. Human success provides direct evidence that an item is solvable under the evaluation setting, whereas failure by a finite set of human participants does not establish that the item is unanswerable. In particular, the 12 model successes among the 37 expert-unsolved items provide concrete counterexamples to equating observed human failure with intrinsic unsolvability. We therefore use the human study as a difficulty diagnostic rather than as the operational criterion for benchmark validity; the latter is established during construction through evidence grounding and adversarial answer-uniqueness verification (Appendix C.5).

E.5 Fixed-Evidence Control

To separate evidence acquisition from parametric knowledge, we evaluate four representative model configurations on all 665 queries with browsing disabled and identical per-query evidence. In ORACLE-5, models receive the five source pages retained during benchmark construction; in NOISY-20, these pages are augmented with 15 same-domain distractors. Because the ORACLE-5 pages were selected during reverse construction with access to the target entity, this experiment is intended as a controlled evidence-sufficiency diag-

Model	No-search	ORACLE-5	NOISY-20
Kimi-K2.5 (Think)	18.0	98.35	98.05
Kimi-K2.5	7.1	98.50	96.54
DeepSeek-V3.2	5.4	96.84	95.64
DeepSeek-R1	6.0	97.74	94.74

Table 4: Fixed-evidence accuracy (%) on all 665 MelonBench queries. Browsing is disabled in all conditions.

nostic rather than a realistic retrieval setting. All other inference and evaluation settings follow the main experiments.

Providing the five construction source pages raises accuracy from 5.4–18.0% to 96.84–98.50% and reduces the cross-model spread from 12.6 to 1.66 points. With 15 additional distractors, accuracy remains 94.74–98.05% and the spread is 3.31 points, corresponding to only a 1.62-point average decrease from ORACLE-5. These results show that cross-model differences largely disappear once relevant evidence is held fixed, supporting evidence acquisition as a major source of end-to-end variation.

F Popularity-Stratified Analysis

Figure 14 shows that performance declines as entity popularity decreases, indicating that less prominent figures are harder to resolve. Since most entities in MelonBench fall in the 100K–300K follower range, the benchmark mainly targets a challenging mid-to-long-tail regime rather than popular head entities. The ECE trend in Figure 14(b) further suggests that calibration does not simply mirror popularity-stratified accuracy, reinforcing that ECE captures a complementary uncertainty-awareness failure mode.

G Semantic Blur Ablation

The de-obfuscation prompt is detailed in Figure 13.

G.1 Qualitative Error Analysis

To complement the semantic blur ablation, we conduct a brief qualitative inspection of failed cases from a 50-item diagnostic sample drawn from the ablation subset. This analysis focuses on whether models can convert coded expressions into effective search intents before retrieval and evidence aggregation. We identify three recurring failure modes.

Metaphor flattening. Models often flatten metaphorical clues into generic evaluative descriptions. For example, expressions such as “acting ceiling” are frequently treated as broad praise for strong acting ability, rather than as discriminative signals that should guide the model toward a narrower set of entities, award profiles, or representative works. This leads to overly broad search queries and retrieval of plausible but incorrect candidates.

Scene-to-event mapping failures. Models also struggle to map scene-like or stylistic descriptions to concrete works, locations, dialect training, or behind-the-scenes events. Clues such as “dockyard smoke” or “spicy diction” require an intermediate decoding step: the model must infer the relevant production context before searching. Failed cases often show the model either searching these phrases too literally or replacing them with vague paraphrases, which prevents retrieval from reaching the intended evidence.

Nickname and generic-event drift. A third failure mode arises when ambiguous nicknames, aliases, or generic events dominate the query. In such cases, models tend to drift toward salient or plausible entities that share similar surface attributes, while missing the specific discriminative clue needed to identify the gold answer. This is especially common when generic events are easier to retrieve than the obfuscated signal.

These qualitative patterns are consistent with the quantitative semantic blur ablation: many failures occur before downstream evidence aggregation, because the coded clue is not converted into the right search intent. They therefore further support our conclusion that coded-expression decoding is an independent bottleneck in MelonBench.

H LLM-side Augmentation Study

Model selection and versioning. The LLM-side augmentation study is intended as a supplementary diagnostic experiment rather than an extension of the main leaderboard. The full benchmark results in Table 1 were obtained with a fixed model set available at the time of the main evaluation, whereas this additional 100-item experiment was conducted later using three more recent model endpoints: Kimi-K2.6 (Thinking), DeepSeek-V4-Pro, and Qwen3.7-Max. Because both the model versions and evaluation subset differ from the main

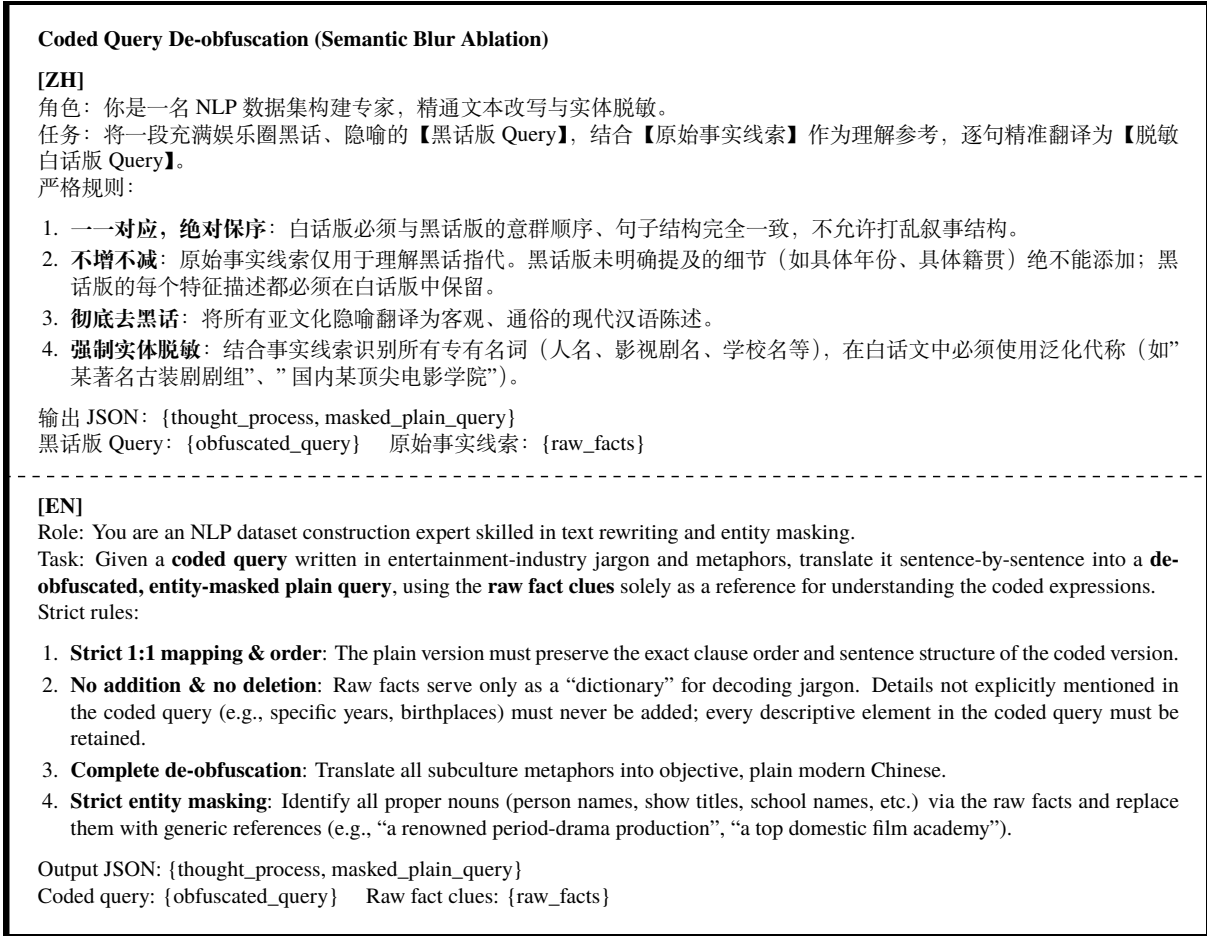


Figure 13: Prompt template for coded query de-obfuscation (semantic blur ablation). Four strict rules enforce order preservation, no addition/deletion, complete de-obfuscation, and entity masking.

table, these results should not be interpreted as directly comparable leaderboard entries. Instead, they provide a broader diagnostic of retrieval-free augmentation methods under a consistent setup within each model.

Setup. To examine whether retrieval-free LLM-side augmentation can alleviate the difficulty of MelonBench, we evaluate two common augmentation strategies, few-shot prompting and self-refinement, on a fixed 100-item subset. We evaluate Kimi-K2.6 (Thinking), DeepSeek-V4-Pro, and Qwen3.7-Max (Qwen Team, 2026f), with reasoning/thinking enabled where supported. All models are evaluated on the same subset using the same fixed few-shot demonstrations. Few-shot demonstrations are selected from correctly solved held-out cases that are disjoint from the evaluated subset, rather than manually written examples. For self-refinement, the model receives only the original prompt and its initial answer in a second call, without gold labels, judge feedback, or external

evidence.

Table 5 shows that retrieval-free augmentation yields heterogeneous and model-dependent gains across the evaluated models. Kimi-K2.6 improves from 17.0% in the zero-shot setting to 20.0% with few-shot prompting and reaches 21.0% when few-shot prompting is combined with self-refinement. DeepSeek-V4-Pro benefits more substantially from few-shot prompting, improving from 21.0% to 31.0%, while adding self-refinement reduces accuracy to 28.0%. Qwen3.7-Max exhibits a different pattern: its zero-shot and self-refinement accuracies are both 7.0%, few-shot prompting alone decreases performance to 3.0%, while combining few-shot prompting with self-refinement reaches 10.0%. These results indicate that augmentation does not yield uniformly monotonic improvements across model families or augmentation strategies. Given the 100-item subset, small differences should be interpreted cautiously. Overall, even the strongest retrieval-free augmented result reaches only 31.0%, suggesting that prompt-side augmentation alone

Model	Zero-shot	Self-refine	Few-shot	Few-shot + Self-refine
Kimi-K2.6 (Thinking)	17.0%	18.0%	20.0%	21.0%
DeepSeek-V4-Pro	21.0%	25.0%	31.0%	28.0%
Qwen3.7-Max	7.0%	7.0%	3.0%	10.0%

Table 5: Retrieval-free LLM-side augmentation results on a 100-item Me1onBench subset. Bold indicates the best augmentation setting within each model

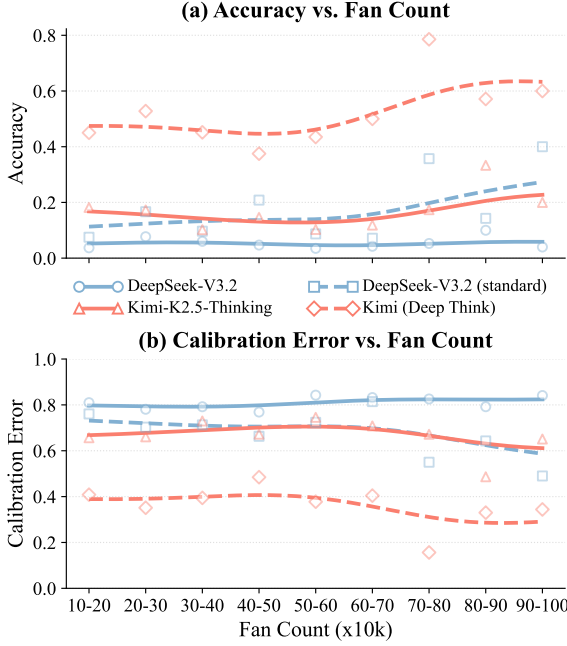


Figure 14: Accuracy and calibration error across fan-count bins. Panel (a) reports accuracy, and panel (b) reports expected calibration error (ECE).

does not remove the core difficulty of Me1onBench.

I Blind-Item Inference

This prompt is used at evaluation time: each blind-item query is wrapped in this template and sent to the model under test. The model must output a structured JSON response containing its reasoning (explanation), a concise entity prediction (exact_answer), and a self-reported confidence score (confidence). The complete prompt is shown in Figure 15.

J Eval Judge Prompt

J.1 Judge Prompt

This prompt is used in the fallback evaluation stage. When exact string matching fails, the question, model answer, and reference answer are sent to a judge model to determine whether the predicted core entity is semantically equivalent to the gold answer. The judge allows valid name variants (e.g.,

aliases or stage names) and returns a structured Boolean decision. The complete prompt is shown in Figure 16.

J.2 Human Validation of Judge Decisions

To assess the reliability of the fallback semantic-equivalence judge, we conduct a stratified human audit of 200 judge decisions. The audited set consists of 100 judge-positive cases, where the judge marks the model prediction as semantically equivalent to the reference answer, and 100 judge-negative cases, where the judge marks the prediction as incorrect. This stratified design allows us to inspect both potential false positives and potential false negatives of the judge.

The audit is conducted by one domain expert with over 10 years of active experience in Chinese entertainment user-generated-content communities. The expert is shown the same inputs as the judge, namely the question, the model’s structured JSON answer, and the reference answer, but is blind to both the judge decision and the model identity. The expert then independently determines whether the model answer should be counted as correct.

As shown in Table 6, human validation agrees with 183 out of 200 audited judge decisions. Among the 100 judge-positive cases, 17 are judged by the human expert to be false positives. These false positives are concentrated in 11 unique questions, suggesting that judge errors mainly arise from a small number of ambiguous or boundary cases rather than being broadly distributed across the benchmark. In contrast, all 100 judge-negative cases are confirmed by the human expert as incorrect model answers, and we observe no false negatives in the audited judge-negative set.

These results indicate that the fallback judge is modestly over-permissive in a small number of ambiguous cases. Consequently, residual judge bias may slightly overestimate absolute accuracy. However, because the observed disagreements are localized and the judge is only used after rule-based exact or normalized matching fails, the audit supports the overall reliability of the evaluation pipeline

Blind-Item Inference Prompt [ZH] {query} 猜测上述的描述指的是谁? 你的回复应遵循以下 json 格式, 且仅输出 json: <pre>{ "explanation": "{你对最终答案的解释}", "exact_answer": "{你简洁的最终答案}", "confidence": "{你对答案的置信度得分在"0%" 到"100%" 之间}" }</pre>	
[EN] {query} Based on the above description, who is this person? Your reply must follow the JSON format below and output only JSON: <pre>{ "explanation": "{Your explanation for the final answer}", "exact_answer": "{Your concise final answer}", "confidence": "{Your confidence score between "0%" and "100%}" }</pre>	

Figure 15: Prompt template for blind-item inference. The model receives a coded blind-item query and must return a structured JSON with its explanation, predicted entity, and confidence score.

Judge decision	Sampled	Human-agreed	Human-disagreed	Agreement
Judge-positive	100	83	17	83.0%
Judge-negative	100	100	0	100.0%
Overall stratified audit	200	183	17	91.5%

Table 6: Human validation of the fallback semantic-equivalence judge.

while clarifying the main direction of remaining judge error.

K Qualitative Trajectory Analysis of Thinking Mode

To complement the quantitative results in Section 5.2, we conduct a post-hoc qualitative trajectory analysis using Kimi-K2.6 in a search-augmented setting. We use Kimi-K2.6 for this analysis because Kimi-K2.5, the model used in the main quantitative experiments, was no longer accessible to us at the time of trajectory inspection. Since the purpose of this analysis is explanatory rather than comparative, we treat Kimi-K2.6 as a representative reasoning-oriented search system and use it only to inspect qualitative behavioral patterns.

We analyze 20 paired cases in which the non-thinking search configuration fails while the corresponding thinking search configuration succeeds. This analysis is intended to explain successful flip cases rather than to claim that thinking mode universally improves performance across all models and instances. Importantly, we focus only on ob-

servable retrieval and answer trajectories rather than hidden reasoning traces. For each paired case, we compare the issued search queries, retrieved evidence, intermediate candidate entities, and final verification behavior. We then manually group recurring differences between the two modes into three trajectory-level patterns: obfuscated clue decoding, search trajectory control, and fragmented evidence integration.

These cases suggest that thinking mode improves performance in retrieval-augmented coded entity resolution not merely by producing longer responses, but by changing how the model uses search. In successful trajectories, thinking mode first maps obfuscated or metaphorical clues into searchable hypotheses, then iteratively revises the search direction as new evidence is obtained, and finally eliminates candidates that satisfy only part of the clue set. This behavior is particularly important for MelonBench because each query combines culturally grounded obfuscation with fragmented open-web evidence; shallow keyword matching can easily retrieve plausible but incorrect candidates.

The trajectory analysis also clarifies the limita-

Answer Evaluation Judge Prompt

[ZH]

请你扮演一位资深的娱乐圈专家。你的任务是作为评测官，结合【问题】，判断【学生答案】中给出的核心人物是否与【参考答案】一致。

请严格遵循以下评判规则：

1. **仅关注核心答案：**你只需要判断【学生答案】最终给出的核心人物（如 exact_answer）是否与【参考答案】一致，无需评判其解释过程。如果学生答案未给出具体确切的人物（如回答“无法确定具体人物”），请直接判为错误（False）。
2. **忽略题目严谨性：**无需考虑【问题】的描述是否严谨或是否能唯一指向该人物，只要【学生答案】的核心结论命中了【参考答案】的人物（注意：本名、艺名视为同一人），即可判为正确。
3. **多答案处理机制：**如果【学生答案】给出了多个名字，你需要结合【问题】判断这些人物是否确实在题目描述中有所体现。只有当【参考答案】包含在这些名字中，且其他名字也是对题干合理分析得出的，才可判为正确。
4. **绝对的格式要求：**请仅使用 Python Dict 的格式返回你的评判结果，不要包含任何额外的解释。

输出格式：

```
{{"reason": "< 简短理由 >", "judge": True/False}}
```

Few-shot 样例（6 条，覆盖边界情况）：

1. 学生答“白鹿”，参考答案“白梦妍” → True（本名/艺名为同一人）
2. 学生答“赵丽颖”，参考答案“白庆琳” → False（不同人物）
3. 学生答“富大龙”（非 JSON 格式），参考答案“阎青妤” → False（不同人物）
4. 学生答“无法确定具体人物”，参考答案“苑冉” → False（未给出确切人物）
5. 学生答“吴京”，参考答案“吴京安” → False（名字相似但非同一人）
6. 学生答“朱一未和白昕怡”，参考答案“白昕怡” → True（多答案中包含参考答案且分析合理）

接下来，请开始你的评判任务：

【问题】：{question}

【学生答案】：{student_answer}

【参考答案】：{reference_answer}

[EN]

You are a senior entertainment industry expert. Your task is to serve as an evaluation judge: given the [Question], determine whether the core person identified in the [Student Answer] matches the [Reference Answer].

Strictly follow these evaluation rules:

1. **Focus only on the core answer:** Judge only whether the final person (e.g., exact_answer) matches the reference. If the student fails to name a specific person (e.g., “cannot determine”), mark it as incorrect (False).
2. **Ignore question rigor:** Do not consider whether the question uniquely identifies the person. As long as the student’s core conclusion matches the reference (note: birth names and stage names count as the same person), mark it correct.
3. **Multi-answer handling:** If the student names multiple people, verify that all named individuals are reasonably supported by the question. Only mark correct if the reference answer is among them and the other names are justified.
4. **Strict format:** Return only a Python dict with no extra explanation.

Output format:

```
{{"reason": "<brief justification>", "judge": True/False}}
```

Few-shot examples (6, covering edge cases):

1. Student: “Bai Lu”, Reference: “Bai Mengyan” → True (stage name = birth name, same person)
2. Student: “Zhao Liying”, Reference: “Bai Qinglin” → False (different people)
3. Student: “Fu Dalong” (non-JSON), Reference: “Yan Qingyu” → False (different people)
4. Student: “Cannot determine”, Reference: “Yuan Ran” → False (no specific person given)
5. Student: “Wu Jing”, Reference: “Wu Jing’an” → False (similar names, different people)
6. Student: “Zhu Yiwei and Bai Xinyi”, Reference: “Bai Xinyi” → True (reference included among justified multi-answers)

Now begin your evaluation:

[Question]: {question}

[Student Answer]: {student_answer}

[Reference Answer]: {reference_answer}

Figure 16: Prompt template for answer evaluation. The judge model receives the question, model answer, and reference answer, and determines whether the predicted entity is semantically equivalent to the gold answer.

tion of reasoning alone. Thinking mode is most effective when there is enough domain knowledge or retrieved evidence to support meaningful hypoth-

esis revision. When the model lacks the cultural background needed to interpret the clue, or when retrieved evidence remains too generic to distin-

Mechanism	Observed pattern	Representative paired case
Obfuscated clue decoding and implicit constraint expansion	The thinking trajectory rewrites literal or metaphorical clues into searchable semantic hypotheses and infers implicit constraints that are not directly expressed in the surface form.	For <i>Lv Jianzhong / Tank</i> , the non-thinking trajectory remains close to literal phrases such as “heart red light.” In contrast, the thinking trajectory interprets the health-related clue as pointing to hereditary heart disease and links the “Three Kingdoms love song” clue to <i>San Guo Lian</i> . This reformulation turns vague surface expressions into searchable constraints that support entity resolution.
Search trajectory control	The thinking trajectory revises the search direction when early results are generic, incomplete, or conflicting, rather than persisting with the initial query formulation.	For <i>Chu Junchen</i> , the non-thinking trajectory stays around broad queries such as “3D director” and “military camp,” which retrieve generic or weakly related evidence. The thinking trajectory redirects the search toward <i>The Taking of Tiger Mountain 3D</i> and actor-training evidence, allowing the model to connect the clue set to a more specific candidate.
Fragmented evidence integration	The thinking trajectory verifies the full clue set before committing to an answer, instead of stopping at a candidate that only partially matches the query.	For <i>Gu Yawei</i> , the non-thinking trajectory stops at a partially matching singer identity. The thinking trajectory checks additional fragmented constraints, including “50 pairs of glasses,” imperfect Cantonese, and the intern role in <i>The Queen of News 2</i> , before selecting the final entity.

Table 7: Qualitative trajectory patterns from 20 Kimi-K2.6 search-augmented flip cases, comparing observable retrieval and answer behavior between non-thinking and thinking configurations.

guish candidates, additional reasoning steps may not yield a correct answer. This is consistent with the heterogeneous gains observed in Section 5.2: some models benefit substantially from thinking mode, while others show only marginal improvement.

Overall, the qualitative evidence supports the interpretation that successful thinking-mode behavior relies on three capabilities: semantic query reformulation, hypothesis revision, and constraint-based candidate elimination. These capabilities complement the quantitative ablations in Section 5.3, which show that MelonBench difficulty arises from the coupled chain of de-obfuscation, retrieval, evidence integration, and open-ended entity recall.

L Finance Cross-Domain Pilot

L.1 Motivation and Scope

To provide preliminary evidence that the coded-entity-resolution framework is not limited to Chinese entertainment, we construct a small finance pilot. The goal of this pilot is not to introduce a full finance benchmark, but to test whether the core evaluation structure of MelonBench can be instantiated in a substantially different domain.

The finance pilot follows the same high-level generate–verify pipeline as MelonBench. Instead of using entertainment figures as target entities, it

uses A-share listed firms. Each item describes a target firm through obfuscated or indirect clues and requires the system to identify the correct company from fragmented open-web evidence. The answer is evaluated at the entity level, allowing common aliases such as the official Chinese short name, full legal name, and stock-code-associated company name when they unambiguously refer to the same firm.

We emphasize that this is a small-scale pilot. It contains 80 items and is intended to support the broader applicability of the proposed evaluation paradigm, rather than to serve as a comprehensive cross-domain benchmark. Accordingly, we report it separately from the main MelonBench leaderboard.

L.2 Construction Details

The finance pilot preserves the five-stage construction structure used in MelonBench: candidate pool construction, web evidence retrieval and filtering, atomic event stratification, blind-item-style query generation, and manual verification. The domain-specific adaptations are summarized below.

Candidate pool. The candidate entities are A-share listed firms. We use listed companies because they provide public identifiability, sufficient open-web evidence, and entity-level verifiability. Each target company is associated with a unique stock

code, which helps disambiguate firms with similar short names. The stock code is used for verification and alias normalization, but is not exposed in the query.

To avoid reducing the task to recognition of highly memorized market icons, we further screen for mid-to-long-tail firms. A candidate is preferred when it satisfies three conditions. First, it must be unambiguously resolvable as a listed firm through its official short name, full legal name, and stock-code-associated name. Second, it must have enough public evidence to support retrieval and verification, including both formal disclosure sources and noisier open-web discussion. Third, it should have relatively low prior salience: we avoid companies whose identity is likely to be recovered from a single famous event, dominant industry position, stock-code association, widely repeated slogan, or landmark transaction. At the same time, we do not select entities that are so obscure that the evidence cannot be reliably retrieved or cross-checked. The intended candidate region is therefore “searchable but not memorized”: firms with public traces, adjacent competitors, and non-trivial operating histories, but without one-shot identifying fame.

Evidence retrieval and filtering. For each candidate firm, we retrieve open-web evidence using search queries based on the company name, stock abbreviation, business segments, major events, and public disclosures. Retrieved pages may include financial news, company announcements, exchange disclosures, investor-relations pages, industry reports, and discussion-oriented web pages. In the finance setting, we also allow noisy sources such as Xueqiu, stock forums, Zhihu finance discussions, finance blogs, and secondary reposted news, but only when the relevant clue can be traced back to or corroborated by company announcements, annual reports, exchange Q&A, or authoritative financial journalism.

The filtering prompt replaces entertainment-specific source criteria with finance-adapted criteria. Each page is scored along six dimensions: (i) **relevance confidence**, measuring whether the page is truly about the target company rather than a same-name or only tangential mention; (ii) **granularity**, measuring whether it contains concrete, verifiable business details rather than generic evaluations; (iii) **non-wiki style**, measuring whether it provides discussion, analysis, or event context beyond a template company profile; (iv) **reliability**,

measuring whether the facts are disclosure-grade or can be corroborated by authoritative sources; (v) **narrative tension**, measuring whether the page contains an event, role change, route shift, or operational consequence useful for blind-item construction; and (vi) **wiki-likeness**, a penalty dimension for pages dominated by boilerplate introductions, official PR, stock summaries, or announcement-like restatements.

Scored by those dimensions, noisy or UGC-like pages are not rejected by default; their value lies in providing indirect phrasing and market-discussion context, but the factual ground truth must remain publicly retrievable and independently verifiable.

Atomic event stratification. As in MelonBench, we extract atomic events from the retained evidence and divide them into discriminative and generic events. The finance extraction prompt requires each fact to be directly related to the target company and explicitly stated or strongly supported by the retained text. Fragmentary statements are merged into complete business-event units rather than treated as isolated snippets. A useful event unit typically combines a business role or actor, an action, an object or operating scene, and a result or role shift, such as a change in customer type, product route, supply-chain position, commercialization stage, production deployment, or post-transaction business consequence.

Discriminative events, stored as *unique facts*, are not simply the most famous facts about a company. They are selected for low-prior but retrievable distinctiveness: they should help narrow the candidate space after web search, while remaining unlikely to be solved from parametric memory alone. We therefore prefer business relationship changes, route switching, operating-position changes, implicit consequences of capacity, certification, or customer-structure changes, and rumors or discussion threads later confirmed by disclosure. We avoid using as primary discriminators facts whose uniqueness depends mainly on company names, stock codes, exchange boards, exact financial figures, formal project or acquisition names, famous counterparties, or one-shot headlines.

Generic events, stored as *non-unique descriptions*, are broad but true company-profile clues that can plausibly apply to multiple nearby firms. Examples include deep involvement in a niche sector, heavy R&D investment, reliance on major downstream customers, visible market controversy,

growth volatility, or exposure to an industry cycle. These clues are useful not because they identify the answer directly, but because they create a plausible first-pass candidate set. Each final item combines one discriminative event with several generic or weakly discriminative descriptions, preserving the same “mislead first, then gradually resolve” structure used in MelonBench.

Query generation and verification. The query-generation stage converts selected atomic events into an obfuscated natural-language query. Following the cross-domain finance prompt, each generated item uses one unique fact and four non-unique facts. The non-unique facts are placed early to construct a broad but plausible company portrait, while the unique fact is split into two or three semantic fragments and dispersed across the query. The query must not directly reveal the company name, stock code, exchange-board identity, formal project name, official transaction label, exact year, exact financial number, or other searchable anchor.

The main difficulty is intended to come from semantic de-obfuscation rather than from memorizing a rare keyword. Therefore, the unique event is rewritten through controlled transformations such as consequence-first phrasing, relationship rewriting, object generalization, causal inversion, and market-expectation reversal. The query should allow a reader to recover a search intent such as “company profile + operating change + consequence”, but should not expose the complete event skeleton in a single sentence. In particular, we avoid placing the core action, key object, and result of the unique fact together, because that would reduce the task to searching a lightly anonymized announcement.

Each draft is checked for identifier leakage, unsupported clue introduction, answer ambiguity, and excessive searchability. The verification stage also checks whether every sentence is grounded in the selected facts, whether all selected clues are reflected in the query, whether the unique fact has been properly fragmented, and whether the opening clues create a sufficiently broad candidate set. Finally, each retained item is manually verified for evidence grounding, public retrievability, and answer uniqueness.

L.3 Representative Example

Figure 17 shows one representative item from the finance pilot. The example illustrates that the pilot preserves the same core difficulty structure as

MelonBench: the target entity is hidden, the clues must be decoded before retrieval, and the answer must be resolved from fragmented public evidence.

L.4 Evaluation Protocol

The finance pilot uses the same open-ended entity-resolution format as MelonBench. For each question, the model must output the predicted target company. A response is counted as correct if it unambiguously identifies the gold company, including accepted aliases such as the official short name, full legal name, or a stock-code-associated name. Ambiguous answers, industry-level answers, parent-company/subsidiary confusions, and answers naming only a product or business segment are marked incorrect unless they uniquely identify the target listed firm.

We evaluate both no-search and search-augmented settings. No-search settings test whether the model can resolve the entity from parametric knowledge and internal reasoning alone. Search-augmented settings test whether the model can decode the indirect clues into effective search intents and integrate fragmented open-web evidence to identify the target company.

L.5 Results

Table 8 reports the full pilot results. Overall, no-search configurations remain weak, while search-augmented configurations generally perform better. The best-performing setting, Kimi (Deep Think), achieves 40.0% accuracy, corresponding to 32 correct answers out of 80. This indicates that the finance pilot is also far from saturated.

The average accuracy of the four no-search settings is 19.06%, while the average accuracy of the six search-augmented settings is 27.92%. This descriptive comparison suggests that search remains important outside the entertainment domain as well. However, search alone does not solve the task: even the strongest search-augmented setting answers fewer than half of the questions correctly. The qualitative pattern is therefore consistent with the main motivation of MelonBench: successful performance requires not only access to web evidence, but also decoding indirect clues, forming useful search intents, and resolving the correct entity from noisy or fragmented evidence.

L.6 Model-Version Caveat

The finance pilot was conducted after the main MelonBench evaluation. Due to model availability,

[ZH]

这家老牌油墨厂长期被贴上区域集中、产供销一体的标签，听起来只是包装赛道里常见的老兵。真正让它换牌的，不是继续铺多口味，而是把原本被视作多元副线的盘子几乎清空——圈里再翻账本，才发现它只剩一张王牌攥在手里。

[EN]

This old-line ink factory has long been tagged with regional concentration and production-supply-sales integration, sounding like just a common old soldier in the packaging track. What really made it change its card was not continuing to spread many flavors, but almost clearing out the plate that had originally been seen as diversified side lines—when the circle flipped through the account book again, only then did it discover that it had only one ace card clenched in its hand.

Figure 17: A representative item from the finance pilot.

Model / Setting	Think	Search	Pilot Acc.
DeepSeek-V4-Pro (DeepSeek-AI, 2026b)	Y	N	26.25%
DeepSeek-V4-Flash (DeepSeek-AI, 2026b)	Y	N	18.75%
Kimi-K2.6(Non-Thinking) (Kimi Team, 2026b)	N	N	8.75%
Kimi-K2.6(Thinking) (Kimi Team, 2026b)	Y	N	22.50%
DeepSeek-V4-Pro + search	Y	Y	31.25%
DeepSeek-V4-Flash + search	Y	Y	33.75%
Kimi(standard)	N	Y	17.50%
Kimi(Deep Think)	Y	Y	40.00%
Qwen3.6(Deep Think) (Qwen Team, 2026e)	Y	Y	25.00%
Qwen DeepResearch	—	Y	20.00%

Table 8: Full results on the finance cross-domain pilot. “—” indicates that the corresponding product setting does not expose a separate thinking-mode toggle and boldfaced scores denote the best result overall across all systems.

API updates, and product-version changes, some model or product versions used in this pilot are not identical to those used in the main MelonBench leaderboard. In particular, the pilot table includes pilot-time settings such as DeepSeek-V4-Pro, DeepSeek-V4-Flash, and Kimi-K2.6, which should not be treated as exact counterparts of the versions reported in the main results.

For this reason, we do not merge these results into the main benchmark leaderboard, and we do not use them to revise the main ranking conclusions. Instead, the finance pilot is reported as a separate cross-domain study. Its purpose is to provide preliminary empirical evidence that the same coded-entity-resolution framework can be instantiated in a non-entertainment domain.

L.7 Limitations

The finance pilot has three important limitations. First, it is small in scale: 80 items are sufficient for

a pilot study, but not for drawing comprehensive conclusions about finance-domain entity resolution. Second, it covers only one additional domain. Although finance differs substantially from entertainment in entity type, evidence sources, terminology, and verification procedures, broader validation across additional domains remains necessary. Third, the pilot uses model/product versions available at the time of evaluation, which are not fully identical to those in the main MelonBench evaluation. The results should therefore be interpreted as preliminary evidence of framework adaptability, not as a direct extension of the main leaderboard.

Despite these limitations, the finance pilot suggests that the proposed evaluation is not inherently tied to Chinese entertainment. Its core structure of obfuscated clues, fragmented open-web evidence, and open-ended entity resolution can be adapted to A-share listed firms, providing a feasible path toward broader cross-domain benchmarks.