K-BROWSECOMP: A Web Browsing Agent Benchmark Grounded in Korean Contexts

Anonymous ACL submission

Abstract

Frontier model evaluations are shifting from foundational capabilities (e.g., instruction following and reasoning) toward compositional, agentic ones, but Korean agentic benchmarks remain scarce. We introduce K-BROWSECOMP, a web-browsing agent benchmark grounded in Korean contexts, consisting of 400 problems. The 300-problem K-BROWSECOMP-VERIFIED subset is manually constructed and validated by native Korean speakers. On this subset, frontier LLMs, including GPT-5.5, DeepSeek-V4-Pro, and GLM-5.1, reach only 30.00–45.67%, a substantial drop from BrowseComp, while Korean LLMs released through Korea's Proprietary AI Foundation Model program obtain only 0.00–10.33%. We further construct a 100-problem synthetic split using hard few-shot exemplars and failure-mode-targeted generation to exploit the asymmetry between solving and creating web browsing problems. On the models evaluated on both splits, accuracy on the synthetic split remains low (0.00–17.00%), indicating comparable difficulty to the human-written set. We publicly release our data and code.

1 Introduction

While leading frontier models from the US and China are shifting their evaluation focus from foundational capabilities (*e.g.*, instruction following, reasoning, and tool calling) toward compositional agentic evaluation (Singh et al., 2025; DeepSeek-AI, 2026), the Korean AI community remains largely anchored to static benchmarks (Korean Ministry of Science and ICT, 2025; Choi et al., 2026; Park et al., 2026; NAVER Cloud HyperCLOVA X Team, 2026). Korean agentic benchmarks are still virtually nonexistent, leaving the community without a standardized way to measure progress. Developing such benchmarks is important for two reasons.

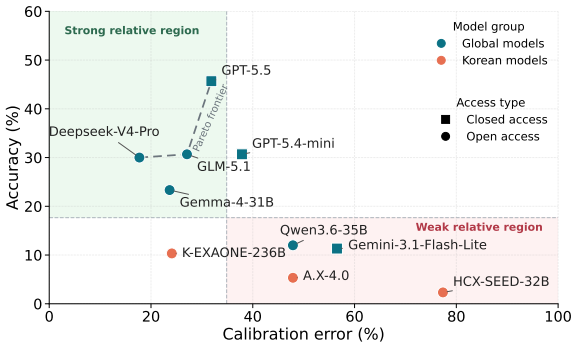

Figure 1: **Accuracy and calibration error of evaluated models on K-BROWSECOMP-VERIFIED.** Higher accuracy and lower calibration error indicate better performance. The shaded quadrants are defined by the median accuracy and calibration error across models. The dashed line marks the Pareto frontier.

- For **Korean developers and users**, language usage and population size place Korea at a structural disadvantage relative to larger language communities, raising AI sovereignty concerns for queries requiring Korean local and cultural knowledge (Kim et al., 2024; Son et al., 2025).

- Second, Korean agentic benchmarks serve the **broader research community**. Frontier models increasingly saturate existing benchmarks, blurring the boundary between in-distribution and out-of-distribution evaluation (Dong et al., 2024; Spiesberger et al., 2026). Agentic benchmarks grounded in linguistically and culturally distinct contexts can therefore provide a principled testbed for broader generalization (Romanou et al., 2025; Whitehouse et al., 2026).

Based on this motivation, we take a first step toward building Korean agentic benchmarks by proposing K-BROWSECOMP, a browsing agent benchmark grounded in Korean contexts. We focus on browsing agents for two reasons. First, browsing agents are uniquely dependent on local and cultural knowledge: their core function is to retrieve

region-specific information from the web, making the gap between Korean and English-centric contexts especially consequential for end users. Second, browsing agents are inherently compositional, jointly exercising instruction following, tool calling, and multi-turn interaction, and thus provide a comprehensive testbed for Korean agentic models.

K-BROWSECOMP consists of 400 problems in total, including the K-BROWSECOMP-VERIFIED subset of 300 manually crafted problems, which are validated by native Korean speakers, and a 100-problem synthetic diagnostic split. As shown in Figure 1 and Table 2, the strongest evaluated models achieve only 30.00–45.67% on K-BROWSECOMP-VERIFIED. Specifically, `GPT-5.5` and `DeepSeek-V4-Pro` obtain 45.67% and 30.00%, respectively, which is a substantial drop from their reported performance on the original BrowseComp (Wei et al., 2025) (84.4% and 83.4%, respectively). Moreover, Korean LLMs released through the first round of government-funded "Proprietary AI Foundation Model Project" (Korean Ministry of Science and ICT, 2025) obtain substantially lower scores ranging from 0.00% to 10.33%. Our trajectory analysis shows that these gaps do not reduce to a single bottleneck. Some models terminate search prematurely or fail to emit stable tool-call trajectories, while others retrieve relevant Korean web evidence but fail to preserve candidate sets, constraints, role bindings, or final-answer state across sources. This indicates that K-BROWSECOMP can serve as a diagnostic target for the Korean ecosystem to develop browsing agents tailored to Korean users.

Beyond K-BROWSECOMP-VERIFIED subset, the core technical contribution of this paper lies in our methodology for constructing the remaining 100 problems with LLMs. A defining property of browsing tasks is their *information asymmetry*: solving a problem can be difficult, while verifying a candidate answer is comparatively easier once the relevant evidence path is known (Wei et al., 2025). Motivated by this asymmetry, we examine an analogous question on the construction side: *while solving such problems is hard, is it also hard to create them*?

To this end, we employ a web browsing LLM agent as a proposer in the same format as our 17 human annotators and study whether it can produce challenging and well-defined problems. When naively instructed, the machine-generated problems were either solvable by frontier models or ill-defined. However, when we (i) provide hard human-written problems as few-shot exemplars or (ii) instruct the agent to target categorized failure modes identified in our analysis, the generated problems become substantially higher in quality and difficulty. Specifically, the models evaluated on both splits achieve 0.00–17.00% accuracy on the 100 machine-generated problems, compared with 0.00–23.33% on the corresponding subset of K-BROWSECOMP-VERIFIED. These results indicate that our generation method produces a challenging diagnostic split.

2 Related Work

Web browsing agents. Prior work established the foundation of evidence retrieval and reasoning, the core components of open-domain QA (Kwiatkowski et al., 2019; Yang et al., 2018; Trivedi et al., 2022; Press et al., 2023). However, their evidence sources are typically predefined (*e.g.*, Wikipedia), or LLMs operate over a fixed number of short turns, which became insufficient to evaluate frontier LLMs. Hence, recent web browsing benchmarks test LLMs as agents that use tools (*e.g.*, web search) over a longer horizon across multiple websites. For instance, BrowseComp (Wei et al., 2025) proposed a set of challenging questions that even humans with access to a web browser can't solve within 2 hours, and BrowseComp-ZH (Zhou et al., 2025) extends the paradigm to utilizing multiple Chinese websites. In our work, we propose K-BROWSECOMP, which necessitates accessing multiple Korean websites. Beyond simply curating the question in Korean language, we consider a broad range of components that reflect the needs of Korean users (*e.g.*, search conventions, local entities, semi-structured pages in Korean websites, and culturally grounded clues) that aren't covered by existing benchmarks.

Korean and regional language evaluation. Korean LLM evaluation has mainly focused on static language understanding, factual knowledge, and reasoning. KorQuAD and KoBEST measure reading comprehension and core NLU skills, while KMMLU scales multitask knowledge evaluation for Korean (Lim et al., 2019; Jang et al., 2022; Son et al., 2025). CLIcK (Kim et al., 2024) evaluates cultural and linguistic intelligence in Korean, and recent regional or multilingual benchmarks such as INCLUDE (Romanou et al., 2025) and MENLO (Whitehouse et al., 2026) emphasize lo-

Type: Parallel **Category:** Entertainment / Media
Question: 2025년에 발매된 앨범 4번 트랙에 있는 이 노래는 띄어쓰기를 제외하고 5글자의 제목을 가졌다. 이 노래의 가사에는 '괜찮아'라는 단어가 7번 등장한다. 이 노래의 제목은 무엇인가?

Question: This song, which appears as the fourth track on an album released in 2025, has a five-character title excluding spaces. The word "괜찮아" ("it's okay") appears seven times in the lyrics. What is the title of the song?

Gold: 편의점 그녀 **Gold:** Convenience store girl

Type: Multi-Hop **Category:** Literature / Books / Language
Question: 한국교육과정평가원에서 2018년 6월 시행한 고3 모의평가 국어 과목 29번에서 다루는 작품에 대하여, 해당 작품의 시인이 생전 열 번째로 간행한 시집에서 열 세 번째로 실려있는 시의 제목은?

Question: In Question 29 of the Korean Language section of the June 2018 mock CSAT for high school seniors administered, what is the title of the poem that appears as the thirteenth poem in the tenth poetry collection published during the poet's lifetime of the poet whose work is discussed in the question?

Gold: 위로 솟구치는 꽃들 **Gold:** Flowers Rising Upward

Figure 2: **Examples of K-BROWSECOMP problems.** The left example requires parallel-branching (*i.e.*, gathering information from multiple websites) while the right example requires multi-hop reasoning (*i.e.*, sequentially traversing through websites).

cally grounded knowledge and native-like norms. These resources are important for Korean language evaluation, but they usually do not require agents to search the web, maintain evidence state, or synthesize information across pages. K-BROWSECOMP fills this gap by converting Korean cultural, institutional, educational, geographic, and media knowledge into agentic browsing tasks. This design evaluates whether a model can act as a useful Korean web agent, beyond answering questions from static benchmark distributions.

Synthetic task generation. A growing body of work uses LLMs to scale instruction data and evaluation data, including self-instruction, synthetic task generation, and adversarial example construction (Wang et al., 2023; Xu et al., 2024; Nie et al., 2020). At the same time, benchmark quality remains a central concern because model-generated tasks can be underspecified, too easy, or contaminated by highly indexed sources. Recent work on data contamination and benchmark generalization shows that evaluation results can overstate model ability when tasks are memorized or only require shallow generalization (Dong et al., 2024). K-BROWSECOMP contributes a complementary construction strategy for web-browsing tasks. It uses human-verified Korean browsing questions to identify recurring failure modes, then asks a browsing agent to generate new questions that target those modes. This pipeline uses the verifiability structure of browsing questions to scale Korean agentic evaluation while keeping the answer unique and publicly supported.

3 K-BROWSECOMP

K-BROWSECOMP is a Korean browsing benchmark designed to evaluate whether web browsing agents can retrieve hard-to-find public information grounded in Korean contexts to solve user queries. As shown in Figure 2, each question is designed to have a single short answer that is stable over time and supported by publicly accessible web evidence. The same figure illustrates the benchmark's two reasoning formats. *Parallel-branching* questions require intersecting multiple independent constraints to identify a unique answer, while *multi-hop* questions require using an intermediate finding to retrieve subsequent evidence. In the following sections, we describe the construction and validation of the K-BROWSECOMP-VERIFIED subset (§3.1) and the 100 LLM-generated questions (§3.2).

3.1 K-BROWSECOMP-VERIFIED

Annotator guidelines. We ask annotators to follow three rules when constructing K-BROWSECOMP-VERIFIED. First, each question should be grounded in Korean contexts and supported by public textual web evidence. Second, each question should be difficult to answer through direct search, but easy to verify once the answer was found. Third, each question should require either multi-hop reasoning or parallel constraint satisfaction, with at least four steps or constraints. We prohibit annotators from using LLMs to create questions, and submitting questions that relies on private, paid, downloaded, or non-textual sources. Also, we require each item to have a unique answer (*i.e.*, no questions that have multiple plausible answers) and temporally stable answer (*i.e.*, answers that don't change over time). Further details are explained in Appendix A.

Validation procedure. All finalized items are manually verified by the authors. For each item, we check whether the gold answer, intermediate entities, and cited sources are recoverable from public web evidence. Items with inaccessible, insufficient, or inconsistent evidence are returned to the original annotators for revision. We also review each item for natural wording, temporal stability, and answer uniqueness. When our baselines produce a con-

Mode	Name	Definition
F0	Incomplete trajectory or malformed output	The model produces an incomplete trajectory, malformed output, or no valid final answer.
F1	Ineffective initial search direction	The model fails to choose a useful initial search strategy.
F2	Search-access structure failure	The model fails to access evidence hidden behind difficult page structures.
F3	Cross-source hopping failure	The model fails to connect evidence across weakly linked sources or entity contexts.
F4	Semi-structured parsing failure	The model misreads tables, lists, rankings, databases, or institutional pages.
F5	Search-result selection failure	The model retrieves relevant evidence but selects the wrong source or candidate.
F6	Sparse entity normalization failure	The model fails to resolve rare names, aliases, spelling variants, or historical names.
F7	Constraint-tracking failure	The model finds partial candidates but fails to satisfy all constraints.
F8	Intermediate reasoning failure	The model fails at date arithmetic, ordering, counting, comparison, or filtering.

Table 1: **Failure-mode taxonomy.** We manually identify and label reasons why models fail on K-BROWSECOMP-VERIFIED. These are later used as the source to construct the synthetic subset of K-BROWSECOMP.

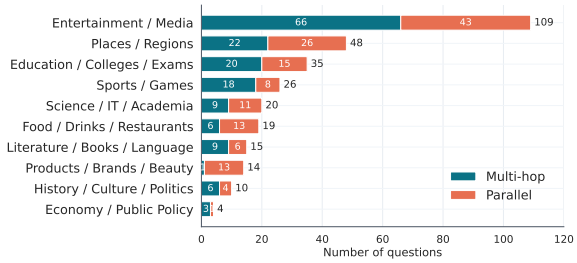

Figure 3: **Category distribution of K-BROWSECOMP-VERIFIED.** Bars show the number of questions in each category, decomposed by question type. Numbers inside bars indicate the counts of *multi-hop* and *parallel-branching* questions, and numbers at the end of bars indicate category totals.

crete answer that differs from the gold answer, the case is manually inspected to determine whether the model finds a plausible alternative answer. An item is retained only when the problem statement, gold answer, expected trajectory, source URLs, and checklist values are mutually consistent, and when the answer is unique.

Dataset statistics. The verified subset consists of **300** questions, where 160 are *multi-hop* questions (53.3%) and 140 are *parallel-branching* questions (46.7%). Figure 3 illustrates the category and reasoning-format distribution. The largest category is *Entertainment and Media*, with 109 questions (36.33%). The second largest category is *Transportation, Places, and Regions*, with 48 questions (16.00%). The remaining questions cover a range of categories, including *education, sports, science, literature, products, history*, and *public policy*.

Trajectory-level failure taxonomy. Each K-BROWSECOMP item requires an extended information-seeking process: forming useful search queries, retrieving candidate evidence, selecting relevant results, interpreting semi-structured source metadata, combining information across multiple steps or parallel branches, and deriving the final answer. To identify where errors occur, we define a trajectory-level failure-mode taxonomy, summarized in Table 1. The taxonomy is constructed through manual inspection of recurring model errors. In Appendix C, we provide examples of each failure mode.

3.2 Generating synthetic web-browsing problems with browsing agents

Generation pipeline. To generate synthetic problems, K-BROWSECOMP-VERIFIED provides both seed pages and a failure taxonomy. Its gold source pages are used to collect related Korean web pages, while observed baseline failures define the nine *failure modes* in Table 1. For generation, we target F1–F8 and exclude F0, since malformed or incomplete trajectories are not a reusable content-level failure for question construction. Each seed page is provided to Claude Code (`claude-opus-4.7` at maximum effort) and is tasked with drafting a single adversarial question. Specifically, it opens the web page and constructs a Korean question *backwards* (i.e., as opposed to opening a web page after reading a question) so that reaching it requires multi-hop or parallel-constraint retrieval, as in the verified set. The question withholds the answer and any paraphrase of it, the source URL and domain, and the page's most identifying entity name, and targets a designated failure mode. The agent refines the question over four iterations (draft → test → revise) until it satisfies the filtering rules.

Filtering rules. Each candidate question passes three sequential filters, and failing any one returns the agent to revision. The first tests *searchability*: the agent issues several web queries and checks whether the gold answer already appears among the results; if so, the question is too easy and is rewritten to target a less exposed detail. The second tests *well-formedness*: a reference solver receives the full source page together with the ques-

tion and must recover the answer, confirming that it is uniquely and faithfully extractable from the page rather than ambiguous or absent. The third tests *adversarial difficulty*: a search-only solver, granted web search but no direct access to the page, attempts the question, and the candidate is retained only if every target model (`gpt-5.4-mini` and `gemini-3-flash-preview`) fails, where failure includes both an incorrect answer and an explicit abstention. Acceptance further requires that each model's failure be attributable to one of the eight failure modes, F1–F8; the agent labels the realized mode honestly, re-assigning it from the intended one when the observed failure differs, for instance recording an unreachable-page failure as F2 rather than disguising it as a parsing failure.

Dataset statistics. Among the 268 candidate questions we generated, 100 problems were not filtered out (37.3% yield). Among the accepted questions, 55, 32, 10, and 3 are obtained on the first through fourth iteration, respectively. Each accepted question is labeled with the failure mode(s) it elicits, distributed as F4 (59), F7 (21), F2 (14), F8 (13), F3 (13), F6 (8), F5 (6), and F1 (1), with semi-structured parsing and constraint accumulation being the most frequently exploited weaknesses. Rejections (*i.e.*, problems that were filtered out) are dominated by seed pages that are already fully indexed: for the majority of the 168 rejected items the gold answer surfaces directly in search (*i.e.*, *searchability*). Also, 66 are solvable by either `gemini-3-flash-preview` or `gpt-5.4-mini`. In the subsequent sections, we use the 100 problems that were not filtered.

Synthetic split diagnostics. We analyze the generated problems as a complementary stress split. The synthetic split preserves the reasoning-format balance of K-BROWSECOMP-VERIFIED, with 53.0% multi-hop and 47.0% parallel examples. Its main differences appear in category composition and question length. The share of *Entertainment and Media* drops from 36.3% to 9.0%, while the share of *Science, IT, and Academia* rises from 6.7% to 33.0%. The questions are also longer, with mean length increasing from 174.46 to 248.40 characters. We further embed all questions with a multilingual sentence-transformer model and test whether the two splits can be separated from question text alone. A simple classifier separates the two splits well, with an ROC AUC of 0.8873. These results suggest that the synthetic split preserves the reasoning format, but differs in domain and surface profile. We therefore report it separately from K-BROWSECOMP-VERIFIED in Section 5.2 and provide further analyses in Appendix D.

4 Experimental Setup

Baselines. We evaluate proprietary models and open-weight models on K-BROWSECOMP-VERIFIED. The proprietary baselines include `GPT-5.5`, `GPT-5.4-mini`, and `Gemini-3.1-Flash-Lite` (OpenAI, 2026; Google, 2026a). The open-weight baselines include `DeepSeek-V4-Pro`, `GLM-5.1`, `Qwen3.6-35B-A3B`, and `Gemma-4-31B-it` (DeepSeek-AI, 2026; Z.ai, 2026; Qwen Team, 2026; Google, 2026b). We also evaluate Korean open-weight models, including `K-EXAONE-236B-A23B`, `A.X-4.0`, `HyperCLOVAX-SEED-Think-32B`, and `Kanana-2-30B-A3B-Thinking-2601` (LG AI Research, 2026; SK Telecom, 2025; NAVER Cloud HyperCLOVA X Team, 2026; Kakao Corp., 2026).

Evaluation protocol. We build our browsing-agent evaluation on the `search_evals` framework (Perplexity Research, 2025). We use the *deep-research agent* and set *Perplexity Search* as the search backend for evaluation. For comparability across models, each agent is given a budget of 10 search calls per question, which is the default setting of the `search_evals` framework. Each model is evaluated once on each of the 300 questions in the verified subset. We use `GPT-5.4-mini` to extract the final answer from each model response and match it against the gold answer, following BrowseComp (Wei et al., 2025). We report the single-run accuracy, which corresponds to pass@1 in this setting.

5 Experimental Results

5.1 Main results

Table 2 reports the performance of closed and open models on K-BROWSECOMP-VERIFIED. `GPT-5.5` achieves the highest accuracy, with 45.67%. Among the remaining models, `GPT-5.4-mini` and `GLM-5.1` both reach 30.67%, while `Deepseek-V4-Pro` obtains 30.00%. These results show that the benchmark remains challenging even for recent high-performing models.

Smaller open-weight models show a wide performance range. `Gemma-4-31B-IT` reaches

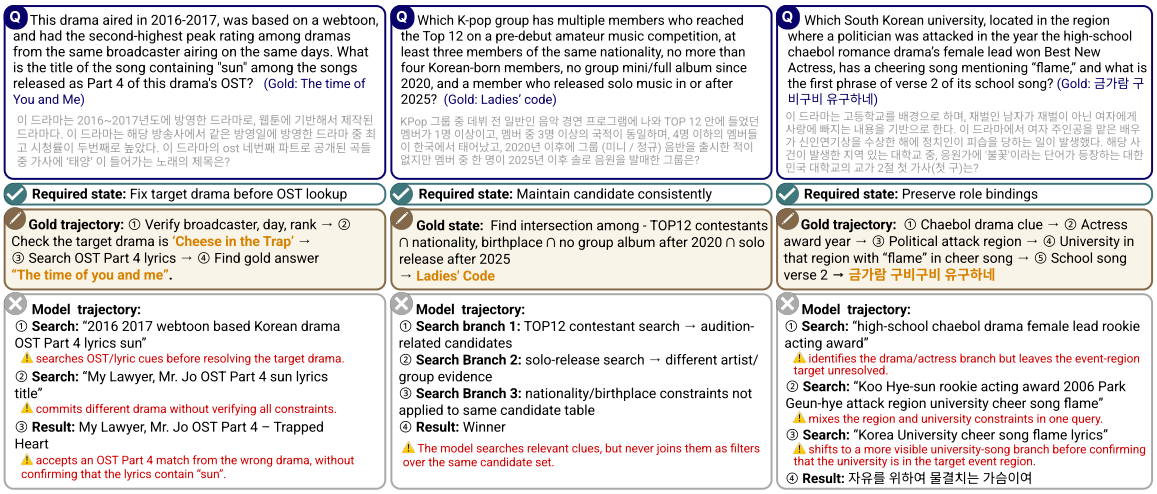

Figure 4: **Representative trajectory-level failures in K-BROWSECOMP.** Each panel contrasts the required intermediate state with the model trajectory. The examples illustrate three recurring post-retrieval failures: *candidate capture*, *unmerged evidence branches*, and *misbound evidence chains*.

		Verified subset		Synthetic subset
Model	**Access**	**Pass@1 Acc.**	**Calib. Err.**	**Pass@1 Acc.**
GPT-5.5	Closed	45.67	31.86	–
GPT-5.4-mini	Closed	30.67	37.88	–
DeepSeek-V4-Pro	Open	30.00	17.72	–
GLM-5.1	Open	30.67	27.07	–
Qwen3.6-35B-A3B	Open	12.00	47.89	15.00
Gemini-3.1-Flash-Lite	Closed	11.33	56.55	11.00
Gemma-4-31B-it	Open	23.33	23.66	17.00
K-EXAONE-236B-A23B	Open	10.33	24.09	13.00
A.X-4.0	Open	5.33	47.89	1.00
HCX-SEED-Think-32B	Open	2.33	77.37	2.00
Kanana-2-30B-A3B-Think	Open	0.00	–	0.00

Table 2: **Performance on K-BROWSECOMP-VERIFIED and the synthetic split.** Pass@1 accuracy (%) and calibration error (%) are computed on the 300-question K-BROWSECOMP-VERIFIED subset. Synthetic accuracy is computed on the 100-question diagnostic synthetic split for the completed subset of models, and is not pooled with the verified score. All runs use the same external retrieval pipeline with the Perplexity Search API.

23.33%, outperforming `Qwen3.6-35B-A3B` at 12.00%. Korean open models score lower in this setting, with `K-EXAONE-236B-A23B` at 10.33%, `A.X-4.0` at 5.33%, and `HyperCLOVAX-SEED-Think-32B` at 2.33%.

Following BrowseComp-ZH (Zhou et al., 2025), we report expected calibration error over five equal-width confidence bins, computed as the weighted average gap between mean confidence and empirical accuracy. Lower values indicate better alignment between confidence and correctness.

5.2 Results on Synthetic Splits

The last column of Table 2 reports performance on the 100-question synthetic split. Due to evaluation cost, we evaluate this split on a focused subset of completed baselines, using the same browsing-agent harness and grading protocol as in Section 4. Accuracy remains low across all evaluated models, ranging from 0.00% to 17.00%, which confirms that the synthetic split is not trivially easier than the verified subset. The model ranking changes slightly: `Gemma-4-31B-it` remains the strongest evaluated model on the synthetic split, while `Qwen3.6-35B-A3B` and `K-EXAONE-236B-A23B` move above `Gemini-3.1-Flash-Lite`.

6 Analysis

6.1 Trajectory-Level Failure Patterns

Whereas Table 1 catalogs failure modes at the level of individual reasoning or search steps, here we examine how a subset of these modes manifest as recurring *trajectory-level dynamics*: patterns in which step-level errors compose across many steps after partially relevant evidence has already been retrieved. Many errors arise after models have already found partially relevant evidence, and these errors often come from losing track of candidates, constraints, and evidence across multiple steps. An illustration is shown in Figure 4.

F5+F7: Search-result selection & constraint-tracking failure (candidate capture). This failure occurs when the model commits to a plausible entity before all upstream constraints have been verified. After this early commitment, later searches become confirmatory: the model searches within the local evidence space of the selected candidate and treats unresolved constraints as satisfied. The final answer can appear locally supported, while

still violating the full question. In the left panel of Figure 4, the model should first identify the target drama before searching its OST lyrics, but instead starts from OST and lyric cues and anchors on *My Lawyer, Mr. Jo*. It then returns *Trapped Heart*, an OST Part 4 song from the wrong drama, illustrating how premature candidate commitment can mislead. See Appendix C.1 for additional examples.

F7: Constraint-tracking failure (unmerged evidence branches). In this pattern, the model searches for several relevant constraints, but each query creates a *separate evidence branch*. The branches are never converted into filters over a shared candidate set. The trajectory may appear systematic because it touches many clues, but the candidate table remains inconsistent across steps.

The middle panel of Figure 4 shows this pattern. The question asks for a K-pop group satisfying parallel constraints on pre-debut competition history, member nationality, and album release status. The gold state is the intersection of these constraints, which yields *Ladies' Code*. However, the model searches these clues separately, applies them to different artists or groups, and returns *Winner*, even though it violates the album-release constraint. This error shows that the retrieved clues were not used to derive the answer consistently.

F3: Cross-source hopping failure (misbound evidence chains). In this pattern, the model follows a plausible search sequence, but binds an intermediate result to the wrong role. This is most visible when the question changes entity type across steps. If the region, institution, song, and lyric roles are not preserved, a high-visibility entity can replace the intended target. The right panel of Figure 4 illustrates this failure. The gold path must preserve a chain from the drama to the actress's award year, the political attack region, a university in that region, and the first phrase of verse 2 of its school song. The model identifies part of the drama and actress branch, but leaves the event region unresolved. It then folds the award, event, university, and cheer-song clues into a broad query, shifts to a visible university-song branch, and returns a plausible phrase from the wrong institution.

F0: Incomplete trajectory or malformed output (answer finalization). Certain models begin a reasonable browsing trajectory but do not stabilize it into a well-formed final answer. This includes incomplete trajectories, unstable intermediate conclusions, excessive uncertainty, or final responses that do not follow the expected answer format. These cases show that browsing performance depends not only on retrieval and reasoning, but also on controlled finalization. The model must know when the evidence is sufficient, when more search is needed, and how to produce a concise exact answer.

Model	Correct n / avg. calls	Wrong n / avg. calls	Δ calls
GPT-5.5	143 / 7.08	157 / 9.30	+2.22
GPT-5.4-mini	92 / 7.88	208 / 9.51	+1.63
Gemini-3.1-Flash-Lite	34 / 5.26	266 / 6.73	+1.47
DeepSeek-V4-Pro	90 / 7.47	210 / 9.80	+2.33
GLM-5.1	92 / 8.23	208 / 9.38	+1.15
Gemma-4-31B-it	70 / 5.20	230 / 8.10	+2.90
Qwen3.6-35B-A3B	36 / 8.11	264 / 9.69	+1.58
K-EXAONE-236B-A23B	31 / 8.35	269 / 9.45	+1.10
A.X-4.0	16 / 2.38	284 / 1.43	-0.95
HyperCLOVAX-SEED-Think-32B	7 / 6.71	293 / 6.84	+0.13

Table 3: **Search-call usage by outcome.** Each cell reports n and average search calls on K-BROWSECOMP-VERIFIED. Δ is the wrong-trial average minus the correct-trial average. All runs use a 10-call budget.

Takeaway. Together, these patterns show that many browsing failures emerge after the required evidence has been retrieved, and that F0, F3, F5, and F7 are the dominant failure modes once retrieval has succeeded. This suggests that progress on K-BROWSECOMP depends on stronger mechanisms for maintaining candidates, constraints, role bindings, and final-answer state across multiple turns.

6.2 Search Effort and Failure Persistence

One might query if the reason behind low performances is due to the insufficient budget allocated for interacting with the search API. In this subsection, we conduct an analysis that shows this is not the case. Table 3 provides a complementary trajectory-level view by comparing search-call usage on correct and incorrect trials. For most models, incorrect trials use more search calls than correct trials and often approach the 10-call budget. This pattern is especially clear for `GPT-5.5` (7.08 vs. 9.30), `DeepSeek-V4-Pro` (7.47 vs. 9.80), `Gemma-4-31B-it` (5.20 vs. 8.10), and `Qwen3.6-35B-A3B` (8.11 vs. 9.69).

The higher search usage on failed trials suggests that many errors are not simply due to insufficient retrieval. Models fail even after evidence has been retrieved, because they do not reliably merge constraints, preserve entity roles, or verify candidates before committing to an answer. `A.X-4.0` is an exception, using few searches on both correct

and incorrect trials, which suggests shallow search or premature stopping. `HyperCLOVAX-SEED-Think-32B` shows little separation between correct and incorrect trials, consistent with failures in trajectory completion or final-answer stabilization. Overall, search volume is a weak predictor of success. A stronger signal is whether the model can maintain candidate, constraint, and role state across the searches it performs.

6.3 Failures in Korean Open-Weight Models

A benchmark should not merely report low scores but inform development. The Korean open-weight models we evaluate underperform global counterparts by a wide margin (Table 2); we provide a model-level diagnosis to guide future iterations.

`A.X-4.0`. Built on Qwen2.5 (Qwen et al., 2025), the model commits to plausible Korean web snippets before all constraints are verified, consistent with its unusually low search usage in Table 3. Korean-targeted post-training has not, under this evaluation setting, translated into trajectory-level state maintenance (Appendix E).

`K-EXAONE-236B-A23B`. Despite its scale, the model frequently loses the intermediate entity chain after relevant retrieval, with later queries drifting from the target entity (Appendix E). Parameter count alone does not appear to substitute for long-horizon state tracking.

`HyperCLOVAX-SEED-Think-32B`. The model initiates reasonable browsing trajectories but rarely converges to a well-formed final answer. Its search counts on correct and incorrect trials are nearly identical (Table 3), suggesting the bottleneck lies in finalization, not retrieval effort.

`Kanana-2-30B-A3B-Thinking-2601`. The model does not produce completed browsing runs under the search_eval harness, often emitting tool-call objects that violate the protocol. We read this as incomplete adaptation to the tool-use paradigm instead of a content-level limitation.

Takeaway. Across models, Korean open-weight LLMs show no clear advantage on K-BROWSECOMP despite its Korean-centric design, and the gap to closed frontier APIs remains substantial. The bottlenecks we observe are stage-specific (*e.g.*, tool-call protocol, post-retrieval state, and answer finalization) rather than reducible to a single cause. We position K-BROWSECOMP as a shared diagnostic target to encourage prioritization of browsing-agent capabilities alongside static-benchmark performance.

6.4 Trajectory Diagnostics on the Synthetic Split

A trajectory review helps explain why the synthetic split remains difficult. As reported in Section 3.2, accepted synthetic questions concentrate on modes that the pipeline can instantiate reliably, especially semi-structured parsing (F4) and constraint tracking (F7). The completed runs show that these targeted operations also reappear as solving failures.

Models often reach the right source family or intermediate entity, but fail at the page-level or candidate-level state needed for the final answer. In a repository metadata question (Appendix F Figure 14), the model reaches the KOPRI repository source family, but extracts a nearby incorrect file-size value instead of the target PDF metadata field. In a KBO record question (Appendix F Figure 15), the model identifies 안우진 (*An Woo-jin*) from the 2022 Kiwoom Heroes pitching clues, but does not maintain a stable table of his 2026 appearances and opponent AVG values before selecting the final date. These cases show that the main bottleneck is not broad retrieval. The synthetic split instead demonstrates whether models can preserve the correct source or candidate state long enough to extract an exact value or enforce the final comparison.

7 Conclusion

We introduced K-BROWSECOMP, a Korean web-browsing agent benchmark with a 300-question human-verified subset and a 100-question synthetic split. Even strong frontier models remain far from solving it, and Korean open-weight models lag substantially behind global counterparts. The synthetic split is similarly difficult, indicating that failure-mode-targeted generation, paired with verification and filtering, produces useful diagnostic items. Our analysis shows that many failures occur after models retrieve relevant Korean web evidence: they often fail to maintain candidates, constraints, source pointers, or final-answer state across the trajectory. Progress on Korean browsing therefore requires stronger trajectory-level state maintenance, not only broader language coverage or larger model scale. We release K-BROWSECOMP as a target for building reliable web-browsing agents that function on Korean web environments.

Limitations

While K-BROWSECOMP provides a first step toward evaluating Korean web-browsing agents, several limitations remain. First, the verified set is still modest in scale and uneven in domain coverage: it contains 300 human-written items, with a substantial share concentrated in entertainment/media and place-related queries. Thus, the benchmark may not fully represent the breadth of Korean web-search use cases. Second, model performance is measured under a single browsing harness, search backend, search-call budget, and pass@1 setting, which may conflate model capability with retrieval coverage, tool-interface reliability, and evaluation protocol choices. Third, although the synthetic split is useful as a diagnostic stress test, it differs from the verified set in surface form and domain composition, and was evaluated only on a subset of models due to cost. Finally, because web evidence and search rankings change over time, continued revalidation will be necessary to preserve answer uniqueness, accessibility, and reproducibility.

References

Eunbi Choi, Kibong Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Hyunjik Jo, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Haeju Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Heuiyeen Yeen, Hwan Chang, Stanley Jungkyu Choi, and 46 others. 2026. K-EXAONE technical report. *arXiv preprint arXiv:2601.01739*.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Google. 2026a. Gemini 3.1 Flash-Lite. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-1-flash-lite. Accessed: 2026-05-24.

Google. 2026b. google/gemma-4-31B-it. https://huggingface.co/google/gemma-4-31B-it. Accessed: 2026-05-24.

Myeongjun Jang, Dohyung Kim, Deuk Sin Kwon, and Eric Davis. 2022. KoBEST: Korean balanced evaluation of significant tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3697–3708, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kakao Corp. 2026. kanana-2-30b-a3b-thinking-2601. https://huggingface.co/kakaocorp/kanana-2-30b-a3b-thinking-2601. Accessed: 2026-05-24.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Korean Ministry of Science and ICT. 2025. Call for proposals: Sovereign AI foundation model project.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

LG AI Research. 2026. K-EXAONE-236B-A23B. https://huggingface.co/LGAI-EXAONE/K-EXAONE-236B-A23B. Accessed: 2026-05-24.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean QA dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

NAVER Cloud HyperCLOVA X Team. 2026. HyperCLOVA X 32b think. *arXiv preprint arXiv:2601.03286*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.5. Accessed: 2026-05-14.

Sungrae Park, Sanghoon Kim, Jungho Cho, Gyoungjin Gim, Dawoon Jung, Mikyoung Cha, Eunhae Choo, Taekgyu Hong, Minbyul Jeong, SeHwan Joo, Minsoo Khang, Eunwon Kim, Minjeong Kim, Sujeong Kim, Yunsu Kim, Hyeonju Lee, Seunghyun Lee, Sukyung Lee, Siyoung Park, and 18 others. 2026. Solar open technical report. *arXiv preprint arXiv:2601.07022*.

Perplexity Research. 2025. search_evals: An evaluation framework for AI-first web search. https://github.com/perplexityai/search_evals.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All. https://qwen.ai/blog?id=qwen3.6-35b-a3b. Accessed: 2026-05-24.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya K Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, volume 2025, pages 83291–83322.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2025. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.

SK Telecom. 2025. A.X-4.0. https://huggingface.co/skt/A.X-4.0. Accessed: 2026-05-24.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Ari Spiesberger, Juan J Vazquez, Nicky Pochinkov, Tomáš Gavenčiak, Peli Grietzer, Gavin Leech, and Nandi Schoots. 2026. Soft contamination means benchmarks test shallow generalization. *arXiv preprint arXiv:2602.12413*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. ♫ MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2504.12516*.

Chenxi Whitehouse, Sebastian Ruder, Tony Zhiyang Lin, Oksana Kurylo, Haruka Takagi, Janice Lam, Nicolò Busetto, and Denise Diaz. 2026. MENLO: From preferences to proficiency – evaluating and modeling native-like quality across 47 languages.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *International Conference on Learning Representations*, volume 2024, pages 30745–30766.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Z.ai. 2026. GLM-5.1: Towards Long-Horizon Tasks. https://z.ai/blog/glm-5.1. Accessed: 2026-05-24.

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. *arXiv preprint arXiv:2504.19314*.

A Dataset Construction

Contributor instructions. Contributors were given a written guide that specified the goal, item format, and exclusion rules for K-BROWSECOMP-VERIFIED. Figure 5 shows an excerpt of the instructions used during question construction. They were asked to write difficult fact-seeking questions that require web browsing and evidence synthesis. Each item had to be grounded in Korean contexts and had to have a short answer supported by public web evidence. Contributors were encouraged to test draft questions with web-enabled models to check difficulty, but they were instructed not to use LLMs to generate or revise the questions.

Korean grounding. Each question had to contain Korean-specific information in the question, the answer, or the intended evidence path. This included Korean institutions, local places, public transportation, educational materials, cultural artifacts, media, historical facts, or expressions commonly used in Korea. The grounding requirement was intended to test browsing ability in Korean information environments, including sources whose structure, terminology, and search conventions differ from English web resources.

Reasoning formats. Each question followed one of two formats. In *multi-hop* questions, the solver first identifies an intermediate entity or value, then uses it to retrieve later evidence and reach the final answer. In *parallel-branching* questions, the solver evaluates several independent constraints and identifies the unique entity satisfying all of them. Multi-hop items required at least four reasoning steps. Parallel-branching items required at least four independent constraints. Contributors were asked to balance the two formats across their submitted items.

Difficulty criteria. Questions were designed to be difficult to answer through direct keyword search. Contributors typically began with a verified target fact, selected attributes with a large search space, and then wrote an inverted question that required recovering the target through search. The answer and intermediate steps had to be easy to verify once the correct path was found. At the same time, the answer keyword should not be the title of a standalone document that reveals the answer without following the intended reasoning path.

Evidence requirements. All required evidence had to be available from public web pages. Sources requiring login, payment, private databases, or membership access were not allowed. Items were also excluded when solving required downloading or interpreting PDFs, spreadsheets, images, or other non-textual artifacts. Contributors were discouraged from relying on a single web platform as the sole basis for an item. When possible, intermediate facts were supported by multiple independent sources.

Answer constraints. Each item had to have a unique final answer under the stated conditions. Contributors were asked to add constraints when another answer could plausibly satisfy the question. Items involving time-sensitive facts had to include a reference date in the question. During review, questions were checked for ambiguous wording, unstable clues, and plausible alternative answers.

Submission format. Each item is submitted in a structured JSON format containing the problem statement, gold answer, expected solving trajectory, source URLs, checklist values for key intermediate slots, Korean-specific keywords, and a short rationale. The checklist is used to check whether the evidence path is valid in our validation procedure. The rationale explained why the item required Korean-context browsing and why direct search was unlikely to be sufficient.

Revision and acceptance. Submitted items were manually reviewed for evidence accessibility, source consistency, wording clarity, answer uniqueness, and temporal stability. During this review, we also checked submissions for private personal information, sensitive personal data, and offensive content, and revised or removed items that failed this screen. Items were returned to the original contributor when the evidence was inaccessible, insufficient, inconsistent, or dependent on excluded source types. Items were also revised or removed when a baseline web-enabled model produced a different concrete answer that was judged to be a plausible alternative. An item was accepted only after the problem statement, gold answer, evidence path, source URLs, and checklist values were mutually consistent.

Recruitment and Payment Non-researcher participants of this study were recruited via a social media post and the remaining participants were authors of this paper. Non-researcher participants

문제 제작 시 주의사항 Precautions When Creating Questions

- **독립성 배제:** 정답이 되는 키워드(인물, 사건, 단체 등)은 해당 키워드를 제목으로 하는 단독 문서(위키 백과, 프로필 페이지, 특집 기사 등)가 없어야 합니다. (앞선 추론 단계들을 거치지 않고서는 정답에 도달할 경로가 원천 차단되어야 합니다.)

  한국교육과정평가원에서 2018년 6월 시행한 고3 모의평가 국어 과목 29번에서 다루는 작품에 대하여, 해당 작품의 시인이 생전 열 번째로 간행한 시집에서 열 세 번째로 실려있는 시의 제목은?

  - https://www.yes24.com/product/goods/28193132 / 처음으로 사랑을 들었다 이수익 시인 시집 13번째 시?
- **비정형 데이터 활용 문제 금지:** 직접 다운로드해서 열어봐야 하는 PDF, Excel, 이미지처럼 비정형 데이터를 읽어야 풀 수 있는 문제는 출제 불가능합니다. 모든 정보는 웹에서 텍스트로 찾을 수 있어야 합니다.
- **공개 웹 정보 활용:** 반드시 로그인 없이 접근 가능한 공개된 웹 정보만을 활용해야 합니다. (유료 DB, 회원제 커뮤니티 제외)
- **유일해의 존재:** 동시에 정답이 될 수 있는 다른 키워드가 존재하지 않아야 합니다.
- **정보는 하나의 웹 플랫폼에 의존적이지 않아야 합니다.** (e.g., 나무위키 정보만으로 문제를 구성하는 것 주의)
- 다단계 추론 (Multi-hop), 병렬 추론 (Parallel Branching) 각 문제는 같은 비율로 만들어져야 합니다. (각 7-8개 => 총 15개

Figure 5: **Excerpt from the written instructions provided to contributors for constructing K-BROWSECOMP-VERIFIED questions.** The guide summarizes the main exclusion and validation rules: answer keywords should not be directly revealed by standalone documents, required evidence must come from publicly accessible textual web sources, non-textual artifacts such as PDFs, spreadsheets, and images are excluded, each question must have a unique answer, evidence should not depend on a single web platform, and multi-hop and parallel-branching items should be balanced.

were paid ₩100,000 (approximately USD 60–70) per 10–15 instances crafted, which corresponds to roughly 4 hours of work and exceeds South Korea's 2025 minimum hourly wage of ₩10,030. Author-participants did not receive separate compensation.

Data Consent All participants were aware that the annotations would be part of constructing the K-BROWSECOMP benchmark.

Artifact release and licensing. We will release K-BROWSECOMP, including the verified and synthetic questions, source URLs, expected trajectories, checklist values, and evaluation code, to support research and evaluation. The K-BROWSECOMP dataset and evaluation code will both be released under the MIT License. The license applies only to benchmark items, metadata, and code created by the authors and contributors; it does not grant rights to third-party web pages linked as evidence, whose contents remain governed by their original providers' terms. We provide source URLs for verification but do not redistribute the full contents of linked web pages.

B Dataset Construction Details

K-BROWSECOMP-VERIFIED consists of **300** human-written Korean browsing problems. The problems were written by Korean-speaking contributors, including both researchers and non-researchers, who were instructed to create questions that require web browsing. Each contributor submitted items in a structured JSON format containing the problem statement, gold answer, expected reasoning chain, source URLs, intermediate checklist values, and Korean-specific keywords. An example submission format is shown in Figure 6.

B.1 Category-wise Performance Analysis

Table 4 reports model accuracy across the major categories of K-BROWSECOMP-VERIFIED. The categories are ordered by the number of questions. Performance varies substantially across domains, even for the strongest frontier models.

Across most models, *Entertainment & Media* and *Sports & Games* are among the highest-performing categories. These questions often involve public figures, entertainment programs, music releases, sports records, or widely discussed media content that are relatively well represented on the public web. By contrast, categories such as *Science, IT & Academia*, *Products, Brands & Beauty*, and *Education, Colleges & Exams* remain difficult for nearly all models. These questions frequently require linking sparse Korean entities across multiple sources, resolving institutional terminology, or navigating semi-structured local web pages.

The gap between global frontier models and Korean-specialized models is also substantial.

GPT-5.5 achieves the strongest overall performance at 47.7%, while Korean-specialized models such as K-EXAONE-236B-A23B and A.X-4.0 remain below 11%. This gap persists even in categories that contain strongly Korean-specific cultural or institutional information. The results suggest that current Korean-specialized LLMs still struggle with long-horizon retrieval, multi-step evidence integration, and constraint tracking in realistic Korean web environments.

Several categories additionally exhibit high variance due to sparse but difficult entity structures. For example, *History, Culture & Politics* shows relatively high scores for some models despite its difficulty because the category contains only ten questions and several items involve recoverable public records once the correct evidence chain is identified. In contrast, *Science, IT & Academia* consistently remains difficult because many questions require cross-source reasoning over researcher profiles, academic affiliations, institutional announcements, and Korean-language metadata.

C Failure Modes Details

This appendix provides additional details for the trajectory-level failure patterns discussed in Section 6.1. The main text identifies three representative patterns: candidate capture, unmerged evidence branches, and misbound evidence chains. Here, we describe how these patterns appear in concrete browsing trajectories and provide additional examples in Figure 11. The examples show that many failures occur after partial retrieval has already succeeded, when the model fails to preserve candidate, constraint, or role state across the trajectory.

C.1 Candidate Capture Failure Details

Candidate capture occurs when the model commits to a plausible candidate before the upstream constraints have been fully verified. After this early commitment, later searches are conducted inside the local evidence space of that candidate. The final answer may therefore appear supported by retrieved evidence, but the candidate itself is not licensed by the full question.

Figure 11(b) shows an answer-side variant of this pattern. The question requires preserving the dependency chain from company to brand, H&B store, and award product. The model searches Olive Young award products before fixing the upstream company and brand. This search order leads the trajectory toward a plausible *CLIO/Goodal/Peripera* branch, and the model then collects award products from that branch. These products are locally plausible, but they are not supported by the company–brand constraint required by the question. The resulting output is therefore a set of plausible products, not the unique gold answer, 어성초 흔적 에센스 패드 (*Abib Heartleaf Spot Pad Calming Touch*).

D Synthetic Split Diagnostics

We conduct additional diagnostics to characterize the 100 synthetic problems. The goal is to test whether the synthetic split follows the same distribution as K-BROWSECOMP-VERIFIED, and to identify which factors explain any observed separation.

We first compare the two splits using question embeddings. All 400 questions are embedded with sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, producing 384-dimensional vectors. We then train a cross-validated domain classifier whose task is to predict whether a question comes from K-BROWSECOMP-VERIFIED or from the synthetic split. This classifier reaches ROC AUC 0.8873 $\pm$ 0.0281. This means that the two splits can be separated from question text alone.

We also run a maximum mean discrepancy test in PCA-reduced embedding space. Maximum mean discrepancy measures how far apart two sets of points are as distributions under a kernel. In this analysis, the two point sets are the verified question embeddings and the synthetic question embeddings. The permutation test randomly shuffles the split labels and recomputes the MMD statistic many times. The resulting p-value measures how often a random label split produces an MMD value at least as large as the observed one. We obtain RBF $\mathrm{MMD}^2 = 0.031967$ and $p = 0.004975$, which indicates a statistically clear embedding-level shift.

To interpret the source of this separation, we train additional classifiers using only simple metadata. Question length alone gives ROC AUC 0.7885 $\pm$ 0.0558. Category alone gives ROC AUC 0.7773 $\pm$ 0.0342. The multi-hop or parallel label alone gives ROC AUC 0.4250 $\pm$ 0.0197. Using length, category, and type together gives ROC AUC 0.8277 $\pm$ 0.0275. These results show that much of the embedding separability comes from longer synthetic questions and category rebalancing. The

Model	C1 n=109	C2 n=48	C3 n=35	C4 n=26	C5 n=20	C6 n=19	C7 n=15	C8 n=14	C9 n=10	C10 n=4	Overall n=300
GPT-5.5	54.1	45.8	42.9	53.8	25.0	36.8	40.0	50.0	70.0	25.0	47.7
GPT-5.4-mini	39.4	31.2	20.0	34.6	5.0	31.6	26.7	21.4	20.0	50.0	30.7
GLM-5.1	32.1	35.4	22.9	42.3	10.0	26.3	20.0	28.6	60.0	25.0	30.7
DeepSeek-V4-Pro	34.9	31.2	31.4	30.8	10.0	21.1	20.0	21.4	50.0	25.0	30.0
Gemma-4-31B-it	25.7	22.9	20.0	34.6	5.0	26.3	13.3	14.3	40.0	25.0	23.3
Gemini-3.1-Flash-Lite	12.8	14.6	5.7	3.8	5.0	15.8	0.0	7.1	40.0	25.0	11.3
Qwen3.6-35B-A3B	10.1	14.6	8.6	23.1	5.0	10.5	0.0	7.1	40.0	25.0	12.0
K-EXAONE-236B-A23B	11.0	8.3	8.6	15.4	5.0	10.5	0.0	7.1	40.0	0.0	10.3
A.X-4.0	4.6	6.2	2.9	11.5	5.0	5.3	0.0	0.0	20.0	0.0	5.3
HyperCLOVAX-Seed-Think-32b	2.8	4.2	2.9	0.0	0.0	0.0	0.0	0.0	10.0	0.0	2.3

Category labels. C1: Entertainment & Media; C2: Places & Regions; C3: Education, Colleges & Exams; C4: Sports & Games; C5: Science, IT & Academia; C6: Food, Drinks & Restaurants; C7: Literature, Books & Language; C8: Products, Brands & Beauty; C9: History, Culture & Politics; C10: Economy & Public Policy.

Table 4: **Accuracy by category on K-BROWSECOMP-VERIFIED. Categories are ordered from left to right by the number of questions.**

reasoning-format label contributes little, consistent with the near-identical multi-hop and parallel proportions across the two splits.

E Trajectory-Level Failure Diagnostics for Korean Open-Weight Models

This appendix expands the analysis of Korean open-weight models in Section 6.1. The main results show that their low performance does not come from a single failure type. Instead, the trajectories reveal different bottlenecks in the browsing loop: some models fail to form a useful candidate state after retrieval, while others retrieve relevant evidence but lose the dependency chain needed for later search and final answer selection.

Shallow evidence control in `A.X-4.0`. Figure 12 illustrates a failure in which `A.X-4.0` reaches relevant Korean historical evidence, but does not convert the retrieved snippets into a maintained candidate ledger. The question requires one historical person to satisfy several constraints at once: civil-service examination records, surname-and-clan evidence linked to a modern athlete, inspection-related office history, and a memorial academy in the person's hometown. The model issues a single broad query that concatenates most of these clues. This search surfaces locally relevant names, including Jeong Hwang, Min Yeong-hwan, and Kim Jong-seo, but the model evaluates them as separate snippet-level candidates. The same-origin modern-athlete clue is never used as a binding constraint, so the trajectory does not keep track of which historical candidates remain valid after each check. The final *I don't know* response therefore occurs after partial retrieval has succeeded. The bottleneck is state formation: the model sees relevant evidence, but fails to organize it into a shared candidate table that can support elimination and verification.

Cross-source chain drift in `K-EXAONE-236B-A23B`. Figure 13 shows a later-stage failure. Unlike the `A.X-4.0` case, `K-EXAONE-236B-A23B` initially reaches the right web region. Its first search retrieves the KakaoTalk free-emoticon event and a plausible list of distributed character emoticons. At this point, the model should freeze the event-side state, identify the smaller-animal character, and then search for the creator and official YouTube channel metadata. The trajectory instead moves to creator search before stabilizing the target emoticon. Subsequent queries contain misspelled character names, malformed creator strings, and unrelated YouTube terms. Although the model later finds a plausible creator-like entity and a channel date, it cannot verify the event–emoticon–creator–channel link. This failure shows that successful initial retrieval is not enough. The intermediate entity must remain stable across source boundaries; otherwise, later searches may be locally plausible but no longer licensed by the original question.

Trajectory completion and protocol stability. The remaining Korean open-weight models show different bottlenecks. `HyperCLOVAX-SEED-Think-32B` often begins a reasonable browsing process, but its trajectories do not reliably converge to an exact final answer. These failures are not limited to one content type. They include unresolved constraint checks, unstable candidate states, and final responses that remain uncertain or under-specified. `Kanana-2-30B-A3B-`

Thinking-2601 shows an earlier bottleneck in the evaluation loop. Because it frequently fails to emit valid tool-call objects under our browsing harness, many runs do not become completed search trajectories. This indicates that tool-use protocol reliability remains a prerequisite for evaluating higher-level browsing skills such as evidence selection and cross-source reasoning.

Summary. These examples clarify why Korean open-weight models score poorly on K-BrowseComp. The issue is not only whether a model can retrieve Korean web pages. A successful browsing agent must convert retrieved pages into a persistent trajectory state: candidate ledgers, constraint checks, entity-role bindings, and final-answer commitments. The Korean open-weight models fail at different points in this process. A.X-4.0 often stops at shallow snippet-level evidence, K-EXAONE-236B-A23B loses dependency chains after relevant retrieval, HyperCLOVAX-SEED-Think-32B struggles to stabilize final answers, and Kanana-2-30B-A3B-Thinking-2601 is limited by tool-call protocol failures.

F Synthetic Split Trajectory Diagnostics

This appendix expands the trajectory-level analysis of the synthetic split. The synthetic split is constructed to target failure modes that can be instantiated reliably by the generation pipeline, with semi-structured parsing (F4) and constraint tracking (F7) appearing most frequently among accepted items. The examples in Figures 14 and 15 show how these targeted modes also appear in completed model trajectories. In both cases, the model does not fail because it searches in an unrelated region of the web. It reaches the correct source family or recovers the main intermediate entity, but loses the page-level or candidate-level state needed to produce the exact final answer.

Metadata-field extraction after successful source localization. Figure 14 illustrates a semi-structured parsing failure in a repository metadata question. The question asks for the file size displayed for a PDF attached to a KOPRI repository record. The required state is narrow: the solver must preserve the target repository item and read the file-size field attached to that specific PDF. The model first retrieves the correct topical context, namely the 2015 KOPRI press release about the Araon Antarctic mid-ocean ridge study. It also reaches the relevant repository neighborhood. However, the trajectory does not keep the target item fixed when reading the file metadata. Because the search results contain nearby PDF records with similar filenames and related institutional context, the model selects a neighboring file-size value, 1.56 MB, while the gold value is 698.85 kB.

This case shows a page-level state failure. The model has enough evidence to enter the right source family, but it does not bind the final extraction step to the correct repository record. The error is therefore not a failure of broad topical retrieval. It is a failure to preserve an exact source pointer and extract the requested field from that source. This is the kind of operation that appears simple after the target page is known, but remains difficult for browsing agents because visually or textually adjacent records create plausible distractors.

Record-level ledger failure after entity recovery. Figure 15 shows a constraint-tracking failure in a KBO record question. The first part of the question identifies the target player through 2022 pitching clues: a Kiwoom Heroes right-handed pitcher who led the league in ERA, strikeouts, and innings. The model correctly recovers 안우진 (*An Woo-jin*) as the target player. It then searches for his 2026 appearance records and reaches relevant baseball-record pages and reports about specific appearances. The remaining task is not entity identification, but comparison over a stable game-level ledger. The solver must maintain rows containing the date, appearance, and opponent AVG values for the player's 2026 games, then select the unique maximum.

The model does not maintain this ledger. Instead, it selects a plausible later appearance date from the retrieved evidence without verifying the maximum opponent AVG condition across all appearances. Its final answer is 05.14, while the gold answer is 05.08. This demonstrates a candidate-level state failure. The model has already found the central entity and entered the right record space, but the final comparison requires a structured table over all relevant games. Without such a table, the trajectory can remain locally plausible while failing the global constraint.

Relation to the synthetic generation pipeline. These two cases clarify what the synthetic split contributes beyond the human-written verified set. The generation pipeline does not merely produce

questions with obscure answers. It creates items in which the answer is easy to verify once the correct state is fixed, but difficult to obtain unless the model preserves that state across the trajectory. For F4-style items, the key state is the exact page item and field to be read. For F7-style items, the key state is a candidate or record ledger that must survive later filtering and comparison. The accepted synthetic questions therefore stress operations that are common in real browsing: reading institutional metadata, selecting among neighboring records, maintaining candidate tables, and enforcing final constraints after the main entity has already been found.

Takeaway. The synthetic split should be interpreted as a complementary diagnostic stress split, not as a replacement for K-BROWSECOMP-VERIFIED. Its category and surface profile differ from the verified set, but its trajectory failures reveal the same broader weakness: models often retrieve relevant Korean web evidence without converting it into a persistent state for exact extraction or final selection. Figures 14 and 15 show this at two granularities. The repository example fails at the source-field level, while the KBO example fails at the candidate-ledger level. Together, they explain why accuracy on the synthetic split remains low even when models reach relevant sources.

G AI assistants in research/writing

In subsection 3.2, we used claude-code to generate the synthetic subset of K-BROWSECOMP and in section 4, include multiple LLMs as our baselines in K-BROWSECOMP. For writing, we mainly used AI assistants for revising grammar and fluency.

H Computational experiments

Model Size Among the open-weight baselines we evaluate, several disclose their parameter counts. Three are dense models: Gemma-4-31B-it (31B parameters), HCX-SEED-Think-32B (32B parameters), and A.X-4.0 (72B parameters; continual pretraining of Qwen2.5-72B on Korean data). Five are Mixture-of-Experts (MoE) models, for which we report total / active parameter counts: Qwen3.6-35B-A3B (35B / 3B), Kanana-2-30B-A3B-Think (30B / 3B), K-EXAONE-236B-A23B (236B / 23B), GLM-5.1 (744B / 40B), and DeepSeek-V4-Pro (1.6T / 49B). For the closed-weight baselines (GPT-5.5, GPT-5.4-mini, and Gemini-3.1-Flash-Lite), parameter counts are not publicly disclosed by the providers.

Compute Budget All experiments ran on openrouter API calls except for Korean models that were not supported on the platform. The total API cost for this project was approximately 320 USD.

K-BROWSECOMP Example (Original)

```
{
  "problem": "한국교육과정평가원에서 2018년 6월 시행한 고3 모의평가 국어 과목 29번에서 다루는 작품에 대하여, 해당 작품의 시인이 생전 열 번째로 간행한 시집에서 열 세 번째로 실려있는 시의 제목은?",
  "answer": "위로 솟구치는 꽃들",
  "type": "multi-hop",
  "category": "문학/도서/언어",
  "expected_chain": [
    {"step": 1, "description": "2018년 6월 시행한 고3 모의평가 국어 과목 29번에서 다루는 작품 확인",
      "sources": ["https://url1"]},
    {"step": 2, "description": "이수익 시인이 생전 열 번째로 간행한 시집 - 소개글 읽어보기",
      "sources": ["https://url2"]},
    {"step": 3, "description": "처음으로 사랑을 들었다 시집에서 목차 확인",
      "sources": ["https://url3"]},
    {"step": 4, "description": "13번째 시 식별"}
  ],
  "checklist": [
    {"step": 1, "slot": "다루는 작품", "value": "이수익-결빙의 아버지"},
    {"step": 2, "slot": "10th_collection", "value": "처음으로 사랑을 들었다"},
    {"step": 4, "slot": "final_answer", "value": "위로 솟구치는 꽃들"}
  ],
  "korean_specific_keywords": ["고3 모의평가", "한국 시"]
}
```

K-BROWSECOMP Example (English Translation)

```
{
  "problem": "In Question 29 of the Korean language section of the June 2018 high-school mock CSAT
  administered by KICE, identify the literary work discussed. For the poet of that work, what is the
  title of the thirteenth poem in the poetry collection that was published as the poet's tenth
  collection during his lifetime?",
  "answer": "Flowers Surging Upward",
  "type": "multi-hop",
  "category": "Literature / Books / Language",
  "expected_chain": [
    {"step": 1, "description": "Identify the literary work discussed in Question 29 of the June 2018
    high-school mock CSAT Korean section.",
      "sources": ["https://url1"]},
    {"step": 2, "description": "Find the poetry collection published as Lee Su-ik's tenth collection
    during his lifetime.",
      "sources": ["https://url2"]},
    {"step": 3, "description": "Check the table of contents of First Hearing Love.",
      "sources": ["https://url3"]},
    {"step": 4, "description": "Identify the thirteenth poem in that collection."}
  ],
  "checklist": [
    {"step": 1, "slot": "target_work", "value": "Lee Su-ik, The Frozen Father"},
    {"step": 2, "slot": "tenth_collection", "value": "First Hearing Love"},
    {"step": 4, "slot": "final_answer", "value": "Flowers Surging Upward"}
  ],
  "korean_specific_keywords": ["Korean high-school CSAT", "Korean poetry"]
}
```

Figure 6: **Example contributor submission format used in K-BROWSECOMP-VERIFIED.** The top shows the original Korean item and the bottom shows its English translation. Each item was submitted as a structured JSON object containing the problem statement, gold answer, expected reasoning trajectory, intermediate checklist values, and Korean-specific keywords.

(a) F1. Ineffective initial search direction

Q Question

다음 조건을 모두 만족하는 말은? 가족과 함께 일을 하는 기수와 경기를 뛴 적이 있고, 한 신문의 편집부장에게 경기 전 모습으로 기대를 받았으며, 가장 오래 보유한 마주는 2013년 11월에 처음 등록했고, 단 한 번의 우승 기록을 보유한다.

(Which horse satisfies all of the following conditions: it has raced with a jockey who worked together with family members, was expected to perform well before a race based on its appearance according to a newspaper editor-in-chief, has an owner who was first registered in November 2013 and remained the horse's longest-associated owner, and has exactly one career win?)

Gold: 퍼플망고 *(Purple Mango)*

S Required state

Form a racehorse candidate set before searching ownership and win-record constraints.

G Gold trajectory

① Identify the jockey clue.
② Find the newspaper editor's pre-race comment.
③ Verify the longest-owning owner's registration date.
④ Check that the horse has exactly one win.
→ **퍼플망고** *(Purple Mango)*

M Model trajectory

Raw search queries

```
"편집부장" "경기 전" "말" "기수"
("editor-in-chief" "before the race" horse jockey)
```

```
"편집부장" "예시장" "경마 말"
("editor-in-chief" "paddock" "horse racing")
```

```
"2013년 11월" "마주" "등록" 경주마
("November 2013" horse owner registration
racehorse)
```

① Searches generic editor/race/horse terms.
② Drifts into general paddock or racing-advice pages.
③ Never forms a concrete racehorse candidate set.
④ **Result: I don't know**

! Failure point

The trajectory breaks at the first retrieval step. The initial queries are too broad to surface candidate horses, so the later ownership and win-record checks have no entity to attach to.

Example model behavior: GPT-5.5

(b) F2. Search-access structure failure

Q Question

KSIC 코드 J63391로 분류되며, 상장폐지 이력이 있고, 2026년 5월 기준 중소기업이며, 본사가 강남구에 있는 회사의 공식 유튜브 채널에서 업로드 일자를 오래된 순으로 정렬했을 때 두 번째 영상 제목은?

(What is the title of the second-oldest video on the official YouTube channel of a company that is classified under KSIC code J63391, has a delisting history, is categorized as an SME as of May 2026, and is headquartered in Gangnam-gu?)

Gold: MBC드라마 '마의' OST Part.1_소향_오직 단 하나_소향 인터뷰 *(MBC Drama Horse Doctor OST Part 1_Sohyang_Only One_Sohyang Interview)*

S Required state

Access a channel-level ordered list, then select the second-oldest video.

G Gold trajectory

① Identify the company under KSIC J63391.
② Verify delisting history, SME status, and Gangnam headquarters.
③ Open the official YouTube channel.
④ Sort videos from oldest to newest.
⑤ Read the second video title.
→ **MBC드라마 '마의' OST Part.1_소향_오직 단 하나_소향 인터뷰** *(MBC Drama Horse Doctor OST Part 1_Sohyang_Only One_Sohyang Interview)*

M Model trajectory

Raw search queries

```
KSIC J63391 한국표준산업분류 J63391
(KSIC J63391 Korean Standard Industrial
Classification)
```

```
"J63391" 회사 상장폐지 강남구 중소기업
("J63391" company delisting Gangnam-gu SME)
```

```
"63391" "데이터베이스" "온라인"
("63391" database online)
```

① Searches company-identification evidence.
② Does not reproduce the YouTube channel ordering operation.
③ Picks a plausible company/video result from search.
④ **Result: 에프앤가이드 온라인 기업설명회** *(FnGuide Online Corporate Briefing)*

! Failure point

The answer is hidden behind an access structure: official channel navigation and oldest-first ordering. General web search surfaces plausible videos, but not the required ordered second item.

Example model behavior: GPT-5.5

Figure 7: **Representative search-direction and access-structure failures in K-BrowseComp.** Each panel shows the question, gold answer, required trajectory state, intended gold trajectory, model search queries, and the point where the trajectory first diverges. F1 illustrates an ineffective initial search direction, where broad generic queries fail to surface any concrete racehorse candidates and leave later constraints unattached. F2 illustrates a search-access structure failure, where the required answer depends on reproducing an internal platform operation, namely oldest-first video ordering within an official YouTube channel.

(a) F3. Cross-source hopping failure

Q Question

퍼듀대학교 공식 한인회 웹사이트를 개발한 초대 멤버 중, 석박통합과정으로 서울대학교에 입학한 사람이 서울대학교에서 소속된 학과 이름은?

(Among the founding members who developed the official Purdue Korean Association website, what was the name of the department affiliated with the person who entered Seoul National University through an integrated M.S./Ph.D. program?)

Gold: 협동과정 인공지능 전공 *(Interdisciplinary Program in Artificial Intelligence)*

S Required state

Preserve the chain: PKA website developer → founding member → SNU integrated Ph.D. entrant → Korean department name.

G Gold trajectory

① Locate Purdue Korean Association website/member evidence.
② Identify the founding web developer.
③ Connect the person to SNU integrated graduate admission.
④ Read the Korean program affiliation.
→ **협동과정 인공지능 전공** *(Interdisciplinary Program in Artificial Intelligence)*

M Model trajectory

Raw search queries

```
퍼듀대학교 공식 한인회 웹사이트 개발 초대 멤버 석박통합과정
서울대학교
(Purdue Korean Association website founding
developer integrated M.S./Ph.D. Seoul National
University)
```

```
"Purdue Korean Association" "website" "Seoul National
University"
```

```
site:purdueka.org "서울대학교" OR "SNU" "석박통합"
(site:purdueka.org "Seoul National University" OR
"SNU" integrated Ph.D.)
```

```
"purdueka.org" "Developed" "member" "SNU"
```

① Reaches relevant Purdue/SNU search regions.
② Does not bind the PKA developer to the SNU page.
③ Normalizes the affiliation to a more generic department.
④ **Result: 컴퓨터공학부** *(Department of Computer Science and Engineering)*

! Failure point

The trajectory loses the cross-source link between the Purdue-side person and the SNU-side program. The final answer is plausible in the SNU context, but not the department attached to the target person.

Example model behavior: GPT-5.5

(b) F4. Semi-structured parsing failure

Q Question

정가가 17,000원 이상이고 시/에세이 분류에 해당하며, 성별이 다른 저자 2명이 있고, 종이책 ISBN 맨 뒤 네 자릿수의 합이 19인 책의 목차에서 가장 긴 챕터 제목은?

(What is the longest chapter title in the table of contents of a book that satisfies all of the following conditions: the list price is at least 17,000 KRW, the category is poetry/essay, the book has two authors of different genders, and the sum of the last four digits of the print ISBN is 19?)

Gold: 착한 여자는 천국에 가고 나쁜 여자는 어디에나 가지만 어리석은 여자는 군부대로 강연을 간다

(Good women go to heaven, bad women go everywhere, but foolish women go to military bases to give lectures)

S Required state

Extract and join book-page fields: price, category, authors, ISBN, and table of contents.

G Gold trajectory

① Filter books by category and price.
② Verify two authors with different genders.
③ Sum the last four digits of the print ISBN.
④ Parse the table of contents.
⑤ Select the longest chapter title.

M Model trajectory

Raw search queries

```
"종이책 ISBN" "시/에세이" "목차" "17,000원" "저자"
("print ISBN" "poetry/essay" "table of contents"
"17,000 KRW" author)
```

```
site:aladin.co.kr/shop/wproduct.aspx "시/에세이" "정가"
"ISBN" "목차" "17,000원" "지은이" "옮긴이"
(site:aladin.co.kr/shop/wproduct.aspx
"poetry/essay" price ISBN "table of contents"
"17,000 KRW" author translator)
```

```
"시/에세이" "17,000원" "ISBN" "목차" "지음" "공저"
("poetry/essay" "17,000 KRW" ISBN "table of
contents" co-author)
```

① Searches the correct book domains.
② Encounters many partially matching book pages.
③ Does not extract all required fields from one page.
④ **Result: I don't know**

! Failure point

The failure begins after domain selection. The model reaches semi-structured book pages, but cannot reliably parse and combine the relevant slots into one candidate.

Example model behavior: GPT-5.5

Figure 8: **Representative cross-source linking and semi-structured parsing failures in K-BrowseComp.** Each panel shows the question, gold answer, required trajectory state, intended gold trajectory, model search queries, and the point where the trajectory first diverges. F3 illustrates a cross-source hopping failure, where the model reaches both the Purdue-side and SNU-side evidence regions but fails to preserve the person-level link across sources. F4 illustrates a semi-structured parsing failure, where the model reaches relevant book pages but fails to extract and combine price, category, author, ISBN, and table-of-contents fields from the same candidate.

(a) F5. Search-result selection failure

Q Question

엔데믹 이후 신규 개통된 철도 노선 중 2026년에 연결되는 두 역을 찾고, 2025년 기준 이용객이 더 적은 역을 고른다. 그 역과 가장 가까운 산, 같은 지역 내 높이가 가장 비슷한 다른 산, 그 산 이름의 유래 동물을 확인하여 2026년 이후 가장 먼저 돌아오는 해당 동물의 해를 답하라.

(Among railway lines newly opened after the endemic period, identify the two stations connected in 2026 and select the station with lower ridership as of 2025. Then find the mountain closest to that station, identify another mountain in the same region with the most similar elevation, determine the animal associated with the origin of that mountain's name, and answer the first year after 2026 corresponding to that animal in the zodiac cycle.)

Gold: 2033년 *(2033)*

S Required state

Select the correct station branch before continuing to mountain, animal-origin, and zodiac-year steps.

G Gold trajectory

① Identify the newly opened line and two stations connected in 2026.
② Select the lower-ridership station as of 2025.
③ Find the nearest mountain.
④ Find the closest-height mountain in the same region.
⑤ Resolve the animal origin and next year.
→ **2033년** *(2033)*

M Model trajectory

Raw search queries

```
2026년 연결되는 두 역 신규 개통 철도 노선 2024 개통 역 이용객 2025
(newly opened railway line stations connected in 2026 ridership 2025)
```

```
서화성역 가장 가까운 산
(nearest mountain to Seo-Hwaseong Station)
```

```
서화성역 주변 산 고도 화성시 남양읍
(mountain elevation near Seo-Hwaseong Station Hwaseong Namyang-eup)
```

① Retrieves station-level evidence.
② Moves into the Seo-Hwaseong station branch.
③ Selects the wrong downstream mountain/animal chain.
④ **Result: 2036년** *(2036)*

! Failure point

The model finds relevant station evidence, but chooses the wrong downstream candidate from nearby search results. The final year follows a coherent but incorrect entity chain.

Example model behavior: GPT-5.5

(b) F6. Sparse entity normalization failure

Q Question

중앙대학교 명예교수이며, 법인설립허가 내부결재일이 1956년인 문화체육관광부 소관의 사단법인 회장을 역임했고, YTN 라디오 등에서 4차 산업혁명과 미래 교육 관련 인터뷰 및 강연을 한 인물은?

(Which person satisfies all of the following conditions: they are an emeritus professor at Chung-Ang University, served as president of an incorporated association under the Ministry of Culture, Sports and Tourism whose internal approval date for establishment permission was in 1956, and gave interviews and lectures on the Fourth Industrial Revolution and future education on YTN Radio and related media?)

Gold: 남영준 *(Nam Young-jun)*

S Required state

Resolve a sparse Korean person entity across university, association, and media-interview clues.

G Gold trajectory

① Verify the Chung-Ang emeritus-professor clue.
② Connect the association to Korea Library Association.
③ Check chair history and 1956 incorporation clue.
④ Match YTN/interview topics to the same person.
→ **남영준** *(Nam Young-jun)*

M Model trajectory

Raw search queries

```
중앙대학교 명예교수 "4차 산업혁명" "미래 교육" "YTN 라디오"
(Chung-Ang University emeritus professor "Fourth Industrial Revolution" "future education" "YTN Radio")
```

```
"이성규" "중앙대학교" "명예교수" "4차 산업혁명"
("Lee Seong-gyu" "Chung-Ang University" emeritus professor "Fourth Industrial Revolution")
```

```
site:mcst.go.kr 사단법인 법인설립허가 1956 "한국도서관협회" 회장 중앙대학교 명예교수
(site:mcst.go.kr incorporated association establishment approval 1956 "Korea Library Association" president Chung-Ang University emeritus professor)
```

① Enters the correct institution and topic area.
② Tests nearby scholar names.
③ Confuses a related library-science figure.
④ **Result: 남태우** *(Nam Tae-woo)*

! Failure point

The model stays in the right disciplinary neighborhood but normalizes the clues to a nearby person. The final name is locally plausible, yet it does not satisfy the full entity chain.

Example model behavior: GPT-5.5

Figure 9: **Representative search-result selection and entity-normalization failures in K-BrowseComp.** Each panel shows the question, gold answer, required trajectory state, intended gold trajectory, model search queries, and the point where the trajectory first diverges. F5 illustrates a search-result selection failure, where the model retrieves relevant station evidence but follows the wrong downstream mountain and zodiac branch. F6 illustrates a sparse entity normalization failure, where the model stays in the correct disciplinary neighborhood but maps the clues to a nearby person instead of the target entity.

(a) F7. Constraint-tracking failure

Q Question

골든디스크어워즈 디지털 음원 부문에서 3곡 이상 수상했고, 발매곡 가사에 멤버 이름이 들어가며, 멤버 중 한 명이 고등학교 배경 드라마에 출연했고, 멤버 국적이 2개 이상인 K-pop 그룹은?

(Which K-pop group satisfies all of the following conditions: the group won awards for at least three songs in the Digital Song Division of the Golden Disc Awards, has a released song whose lyrics contain the name of one of the members, has at least one member who appeared in a drama set in a high school, and consists of members from at least two different nationalities?)

Gold: 미쓰에이 *(miss A)*

S Required state

Maintain one candidate ledger across all parallel constraints.

G Gold state

Golden Disc digital-song winners with $\geq$ 3 awarded songs
- $\cap$ lyrics mentioning a member name
- $\cap$ member appeared in a high-school drama
- $\cap$ multinational membership

$\rightarrow$ **미쓰에이** *(miss A)*

M Model trajectory

Raw search queries

```
Golden Disc Awards Digital Song Bonsang winners
groups 3 songs TWICE BLACKPINK
```

```
(G)I-DLE Golden Disc Awards Digital Song Bonsang
Tomboy Queencard Fate
```

```
(G)I-DLE My Bag lyrics Miyeon Minnie Soyeon Yuqi
Shuhua names
```

① Finds a strong partial candidate: (G)I-DLE.
② Verifies award-history evidence for that candidate.
③ Searches lyric evidence inside the same branch.
④ Does not enforce all remaining constraints.
⑤ **Result: (여자)아이들** *((G)I-DLE)*

! Failure point

The model commits to a candidate that satisfies salient early constraints. Later searches support that candidate locally, but the full constraint intersection is never checked.

Example model behavior: GPT-5.5

(b) F8. Intermediate reasoning failure

Q Question

2024년 6월 네이버 블로그 챌린지의 최대 포인트 X, 2023년 5월 챌린지에서 레벨 2 달성 시 포인트를 받은 인원 Y, 2022년 6월 챌린지 4주 참여 포인트 Z를 더한 수의 약수 개수는?

(What is the number of divisors of the sum of the following three quantities: the maximum point reward X in the June 2024 Naver Blog Challenge, the number of users Y who received points for reaching Level 2 in the May 2023 challenge, and the four-week participation reward Z in the June 2022 challenge?)

Gold: 60개 *(60)*

S Required state

Retrieve X, Y, and Z, then compute the divisor count of $X + Y + Z$.

G Gold trajectory

① Find the 2024 June challenge reward value X.
② Find the 2023 May level-2 recipient count Y.
③ Find the 2022 June four-week reward value Z.
④ Compute $X + Y + Z$.
⑤ Count the divisors.
$\rightarrow$ **60개** *(60)*

M Model trajectory

Raw search queries

```
2024년 6월 네이버 블로그 챌린지 첫 주부터 참여 3주 연속 성공
최대 포인트
(June 2024 Naver Blog Challenge maximum reward for
three consecutive successful weeks)
```

```
포토덤프 챌린지 3주 연속 성공 네이버페이 포인트 최대 2024
(Photo Dump Challenge maximum Naver Pay reward for
three consecutive successful weeks 2024)
```

```
2023년 5월 네이버 블로그 챌린지 레벨 2 달성 2천원 포인트 몇명
(May 2023 Naver Blog Challenge number of Level-2
users receiving 2,000 KRW points)
```

① Retrieves the right kind of slot values.
② Continues searching year-specific challenge evidence.
③ Miscomputes the final arithmetic or divisor count.
④ **Result: 96**

! Failure point

The failure starts after retrieval. The model fills relevant intermediate slots, but the final numerical operation produces the wrong divisor count.

Example model behavior: GPT-5.5

Figure 10: **Representative state-maintenance and intermediate-reasoning failures in K-BROWSECOMP.** Each panel shows the question, gold answer, required trajectory state, intended gold trajectory, model search queries, and the point where the trajectory first diverges. F7 illustrates a constraint-tracking failure, where the model commits to a locally plausible K-pop group candidate without enforcing the full intersection of parallel constraints. F8 illustrates an intermediate reasoning failure, where the model successfully retrieves relevant intermediate quantities but produces an incorrect final divisor count during numerical reasoning.

(a) Candidate-ledger failure

Q Question

골든디스크어워즈 디지털 음원부분에서 3곡 이상으로 수상을 한 Kpop 그룹으로, 발매한 노래의 가사 중 멤버의 이름이 들어간 노래가 있고, 멤버 중 한명이 고등학교를 배경으로 한 드라마에 출연한 적이 있으며, 멤버들의 국적이 2개 이상으로 이루어진 그룹은?

(Which K-pop group satisfies all of the following conditions: the group won awards for at least three songs in the Digital Song Division of the Golden Disc Awards, has a released song whose lyrics contain the name of one of the members, has at least one member who appeared in a drama set in a high school, and consists of members from at least two different nationalities?)

Gold: 미쓰에이 (*miss A*)

S Required state

Maintain one candidate ledger across all parallel constraints.

G Gold state

- Golden Disc digital-song winners with $\geq$ 3 awarded songs

$\cap$ lyrics mentioning a member name

$\cap$ member appeared in a high-school drama

$\cap$ multinational membership

$\rightarrow$ **미쓰에이(miss A)**

M Model trajectory

Search queries

```
골든디스크어워즈 디지털 음원 부문 수상 3곡 이상 Kpop 그룹
(Golden Disc Awards most digital song bonsang wins
Kpop group 3 or more)
```

① Search broad award-history candidates.
② Drift toward high-visibility artists such as IU / BTS.
③ Discuss lyric, drama, and nationality clues separately.
④ Never merge them into one shared candidate table.
⑤ **Result: I don't know**

! Why this fails

- Searches relevant award evidence, but never converts it into a shared candidate set.
- Treats the remaining clues as separate unresolved checks.
- Fails to intersect all constraints on the same group.

Example model behavior: K-EXAONE-236B-A23B

(b) Answer-side candidate capture

Q Question

2026년 4월 기준으로 기초 화장품과, 색조 화장품, 남성용 화장품과, 여성 케어 총 네가지 브랜드를 소유중인 한국 회사의 브랜드 중 2005년 이전에 설립된 헬스 앤 뷰티 스토어에 2018년 이전에 입점한 브랜드에서 2024–2025년에 해당 스토어에서 수상한 제품의 이름은?

(As of April 2026, among the brands owned by a Korean company that operates four categories of brands, namely basic skincare, color cosmetics, men's cosmetics, and feminine care, what is the name of the product that won an award at that health-and-beauty store in 2024–2025, from a brand that had entered a health-and-beauty store established before 2005 prior to 2018?)

Gold: 어성초 흔적 에센스 패드 (Abib Heartleaf Spot Pad Calming Touch)

S Required state

Preserve the chain: company $\rightarrow$ brand $\rightarrow$ H&B store $\rightarrow$ award product.

G Gold trajectory

① Identify the Korean company satisfying the brand-category constraint.
② Fix the qualifying brand and H&B store.
③ Resolve the store as Olive Young.
④ Search 2024–2025 award results for that fixed brand.
⑤ **Result: 어성초 흔적 에센스 패드** (Abib Heartleaf Spot Pad Calming Touch)

M Model trajectory

Raw search queries

```
2024 2025 헬스 앤 뷰티 스토어(H&S) 어워드 수상 제품 리스트
기초 색조 남성 여성케어 2005년 이전 설립 브랜드 (2024-2025
H&B store award winners skincare makeup men's care
feminine care)
```

```
헬스 앤 뷰티 스토어 H&S 화장품 입점 브랜드 2018년 이전 (H&B
store cosmetic brands before 2018)
```

```
올리브영 2024 2025 어워드 수상 제품 브랜드 2005년 설립 이전
(Olive Young 2024-2025 award winners pre-2005
brands)
```

① Search Olive Young / H&B award products early.
② Search store-entry evidence before fixing the upstream brand.
③ Anchor on the CLIO / Goodal / Peripera branch.
④ Collect locally plausible award products from that branch.
⑤ **Result: multiple CLIO / Goodal / Peripera products**

! Why this fails

- Searches answer-side award evidence before fixing the upstream brand identity.
- Commits to a locally plausible CLIO/Goodal/Peripera branch.
- Returns multiple locally supported products instead of the unique gold answer.

Example model behavior: Qwen3.6-35B-A3B

Figure 11: **Representative trajectory-level failures in K-BrowseComp.** Each panel shows the question, the required intermediate state, the gold state or trajectory, and the model's raw search queries. Panel (a) shows a candidate-ledger failure: the model retrieves partially relevant award evidence, but does not merge all constraints into a shared candidate set, failing to identify 미쓰에이(*miss A*). Panel (b) shows answer-side candidate capture: the model searches the award side of the question before fixing the upstream company and brand, then anchors on a locally plausible branch and returns multiple products instead of the gold answer, 어성초 흔적 에센스 패드 (*Abib Heartleaf Spot Pad Calming Touch*).

Shallow evidence control after a broad search

Q Question

형제가 새 왕조 개창 후 몇 년이 지나지 않아 문과 급제자 명단에 이름을 올렸으며, 동성동본의 현대 운동선수 가운데 라리가와 리그 1 무대를 모두 경험한 사람이 있고, 감찰과 관련된 관직을 역임한 적이 있으며, 사후 그의 고장에 그를 기리는 서원이 세워진 인물은?

(Which person satisfies the following conditions: his brothers appeared on the civil-service examination list shortly after the founding of a new dynasty; a modern athlete of the same surname and clan experienced both La Liga and Ligue 1; he held an inspection-related office; and a memorial academy was established in his hometown after his death?)

Gold: 이명성 *(Lee Myung Seong)*

S Required state

Maintain one historical-person candidate ledger across exam records, surname/clan evidence, the modern-athlete clue, office history, and memorial-site evidence.

G Gold trajectory

① Resolve the same-surname-and-clan clue through the modern athlete who played in both La Liga and Ligue 1.
② Use that surname/clan state to search historical figures with brothers listed in early civil-service examination records.
③ Filter candidates by inspection-related office history.
④ Verify that a memorial academy was established in the candidate's hometown.
⑤ Return the remaining candidate.
→ **이명성** *(Lee Myung Seong)*

M Model trajectory

Raw search query

```
형제 새 왕조 개창 후 문과 급제자, 동성동본 현대 운동선수 라리가와 리그 1 경험, 감찰 관련 관직, 사후 서원 설립 인물
(civil-service passer after brothers founded a new dynasty same-clan modern athlete La Liga and Ligue 1
experience censorship-related office posthumous academy figure)
```

① The model begins with one broad query that mixes all constraints into a single search.
② The search returns partially relevant historical material, including exam-list pages and visible names such as 정황, 민영환, and 김종서.
③ The model summarizes these candidates one by one, checking some local clues such as examination records, inspection-related offices, or memorial evidence.
④ The same-origin modern-athlete clue is not used as a binding constraint on the candidate set.
⑤ The trajectory never forms a shared ledger of remaining candidates, so evidence about different people is compared only locally.
⑥ **Result: I don't know**

! Failure point

The model reaches relevant Korean historical evidence, but treats retrieved snippets as candidate descriptions instead of filters over one shared candidate state. The failure occurs after retrieval: plausible names are discussed, but the full constraint set is never enforced on the same person.

Example model behavior: A.X-4.0

Figure 12: **Shallow evidence-control failure in `A.X-4.0`.** The example shows a Korean historical-entity question that requires maintaining one candidate ledger across civil-service examination records, surname/clan evidence, a modern-athlete clue, office history, and memorial-site evidence. The model issues a single broad query that concatenates most constraints, retrieves partially relevant historical candidates, and discusses locally plausible names such as 정황, 민영환, and 김종서. However, it does not enforce all constraints on the same candidate state, leaving the same-origin modern-athlete clue unbound and ending with *I don't know*. This illustrates weak evidence control after retrieval: relevant snippets are found, but they are not converted into filters over a stable candidate ledger.

Cross-source chain drift after successful initial retrieval

Q Question

2025년 7월 3째주에 카카오톡 무료 이모티콘 이벤트에서 배포한 이모티콘 중 더 작은 동물과 관련된 캐릭터 이모티콘의 제작자의 유튜브 채널 가입일은?

(Among the emoticons distributed in the KakaoTalk free-emoticon event in the third week of July 2025, what is the YouTube channel join date of the creator of the character emoticon related to the smaller animal?)

Gold: 2021년 8월 9일

S Required state

Preserve the dependency chain: event week → distributed emoticons → smaller-animal character → creator → official YouTube channel metadata.

G Gold trajectory

① Identify the KakaoTalk free-emoticon event for the third week of July 2025.
② List the animal-related character emoticons distributed in that event.
③ Determine which distributed character corresponds to the smaller animal.
④ Find the creator of that fixed character.
⑤ Open the creator's official YouTube channel metadata and read the channel join date.
→ **2021년 8월 9일**

M Model trajectory

Raw search queries

```
카카오톡 무료 이모티콘 이벤트 2025년 7월 셋째주 작은 동물 캐릭터
```

```
극철이 망그래진 헤터 오둥이 연락켓 YouTube 참니달콕라 가입일
```

```
극락 쿼커 이모티콘 제작자 YouTube
```

```
"긍" "긍" "긍" "ompp" YouTube
```

```
긍비금비 카카오 이모티콘 작가 극락 쿼커
```

① The first search enters the right event region and retrieves the Kakao announcement and related reports.
② The retrieved evidence lists candidate emoticons such as 곽철이, 망그러진 햄터, 오둥이, 극락 쿼카, and 포테토뭉.
③ The model does not freeze the event-side candidate state before moving to creator search.
④ Follow-up queries **begin to mix misspelled character names, malformed creator strings**, and unrelated YouTube terms.
⑤ The model finds a plausible creator-like entity, 긍비금비, and a YouTube channel date, but cannot verify that this creator is the maker of the target smaller-animal emoticon.
⑥ The final explanation reports this unresolved link and returns **I don't know**.

! Failure point

The initial search retrieves the relevant Kakao event evidence, but the model fails to stabilize the intermediate target emoticon before moving to creator search. After this point, the follow-up queries contain corrupted character names and no longer preserve the event–emoticon–creator–channel dependency chain.

Example model behavior: K-EXAONE-236B-A23B

Figure 13: **Cross-source chain drift in `K-EXAONE-236B-A23B`.** The example shows a KakaoTalk free-emoticon question whose answer requires preserving the dependency chain from event week to distributed emoticons, smaller-animal character, creator, and official YouTube channel metadata. The model's first query reaches the relevant Kakao event evidence and retrieves the candidate emoticon list, including 곽철이, 망그러진 햄터, 오둥이, 극락 쿼카, and 포테토뭉. After this initial retrieval, the model fails to stabilize the target emoticon before moving to creator search. Its follow-up queries contain corrupted character names and malformed creator/channel strings, and the trajectory ends without a verified creator–channel link. This illustrates weak evidence control across sources: the model reaches the right web region, but loses the intermediate entity chain needed to license the final answer.

Metadata-field extraction failure after reaching the correct repository

Q Question

극지연구소는 2015년 2월, 쇄빙연구선 아라온호를 활용한 남극 중앙해령 탐사를 통해 현무암질 해양 지각에 빙하기-간빙기 순환의 기록이 남아 있다는 사실을 세계 최초로 규명했고, 그 성과가 국제 학술지 '사이언스(Science)'에 논문으로 실렸다는 내용을 알리는 보도자료를 냈다. 이 보도자료 원문은 극지연구소가 운영하는 기관 리포지터리에 PDF 파일 한 건으로 등록되어 일반에 공개되어 있다. 해당 자료 상세 페이지에서 이 첨부 PDF 파일에 대해 표시하고 있는 파일 용량은 무엇인가?

(In February 2015, the Korea Polar Research Institute announced that research using the icebreaker Araon had found evidence of glacial-interglacial cycles in basaltic oceanic crust at the Antarctic mid-ocean ridge, with the result published in Science. The original press release is publicly registered as a PDF file in the institute repository. What file size is displayed for the attached PDF on the repository detail page?)

Gold: 698.85 kB

S Required state

Preserve the target repository item and extract the file-size field from the attached PDF metadata, not from a nearby PDF or another repository record.

G Gold trajectory

① Identify the 2015 KOPRI press release about the Araon Antarctic mid-ocean ridge study.
② Open the KOPRI repository detail page for the target press release.
③ Locate the attached PDF in the item metadata.
④ Read the displayed file-size value exactly as written on the page.
→ **698.85 kB**

M Model trajectory

Raw search queries

```
극지연구소 2015년 2월 아라온호 남극 중앙해령 사이언스 보도자료
site:repository.kopri.re.kr "6-1.pdf" 4533
(Korea Polar Research Institute February 2015 Araon Antarctic mid-ocean ridge science press release)
```

① The model retrieves the correct KOPRI topic and reaches the repository source family.
② The search results include the target repository item and several nearby PDF records with similar filenames.
③ The model does not preserve the target item when reading the file metadata.
④ It selects a nearby incorrect file-size value.
⑤ **Result:** 1.56 MB

! Failure point

The failure occurs after relevant retrieval. The model reaches the correct repository neighborhood, but does not extract the exact metadata field from the target item. This is a semi-structured parsing failure (F4).

Example model behavior: Gemini-3.1-Flash-Lite

Figure 14: **Semi-structured metadata parsing failure in the synthetic split.** The question asks for the file size displayed for an attached PDF on a KOPRI repository detail page. The model reaches the correct source neighborhood and retrieves the relevant press-release context, but it does not preserve the target repository item when reading the file metadata. It returns a nearby incorrect file-size value, 1.56 MB, instead of the gold value, 698.85 kB.

Game-level constraint tracking failure after identifying the target player

Q Question

2022 정규시즌에 평균자책점, 탈삼진, 이닝 세 부문에서 모두 리그 1위에 오른 키움 히어로즈 소속의 한 우완 투수가 있다. 이 선수가 2026 정규시즌에 등판한 경기들 가운데, 그 경기에서 상대 타자들에게 기록당한 피안타율이 가장 높았던 경기는 단 한 경기뿐이다. 그 경기의 날짜는 언제인가? 답은 기록 페이지의 표기 방식인 'MM.DD' 형식으로 적어라.

(A right-handed pitcher for the Kiwoom Heroes led the league in ERA, strikeouts, and innings pitched during the 2022 regular season. Among the games in which this player appeared during the 2026 regular season, exactly one game had the highest opponent batting average against him. What was the date of that game? Answer in the record-page format, 'MM.DD'.)

Gold: 05.08

S Required state

Maintain one game-level ledger for the target pitcher's 2026 appearances, compare opponent AVG values across games, and select the unique maximum.

G Gold trajectory

① Identify the 2022 Kiwoom pitcher who led ERA, strikeouts, and innings.
② Fix the target player as 안우진 *An Woo-jin*.
③ Retrieve the player's 2026 game-level pitching records.
④ Compare opponent AVG values for each appearance.
⑤ Select the single game with the highest opponent AVG and return its date.
→ **05.08**

M Model trajectory

Raw search queries

```
2022 KBO 평균자책점 탈삼진 이닝 1위 키움 히어로즈 투수
(KBO pitcher from Kiwoom Heroes who led the league in ERA, strikeouts, and innings pitched in 2022)
```

```
안우진 2026 시즌 등판 기록
(Ahn Woo-jin 2026 season appearance record)
```

```
안우진 2026 시즌 등판 경기 날짜 피안타율
(Ahn Woo-jin 2026 season game dates opponent batting average)
```

```
안우진 "피안타율이 가장 높았던 경기는 단 한 경기뿐" 2026
(Ahn Woo-jin "only one game had the highest opponent batting average" 2026)
```

① The model correctly identifies the target player as 안우진 *(An Woo-jin)*.
② It then searches for 2026 appearance records and game-level opponent AVG evidence.
③ The searches reach relevant baseball-record pages and news reports, including KBO player-record pages and reports about specific 2026 appearances.
④ However, the trajectory does not maintain a complete game-level ledger with date, appearance, and opponent AVG fields.
⑤ The model selects a plausible later date from the retrieved appearance evidence without verifying the unique maximum AVG condition.
⑥ **Result:** 05.14

! Failure point

The target player is correctly recovered, and the later searches enter the relevant 2026 record space. The failure occurs at the constraint-maintenance step: the model does not build a stable game-level table and therefore cannot compare the opponent AVG field across all appearances.

Example model behavior: Gemini-3.1-Flash-Lite

Figure 15: **Constraint-tracking failure in a synthetic KBO record question.** The question identifies 안우진 *(An Woo-jin)* from the 2022 league-leading pitching clues and requires comparing opponent AVG values across his 2026 game-level records. The model recovers the target player and enters the relevant baseball-record region, but does not keep a stable game-level ledger for the final comparison. It selects 05.14, while the gold date is 05.08. This example shows how the synthetic split stresses constraint maintenance and exact selection after the main entity has already been found.