DIVERSITY IS MEASURED, NOT ASSUMED: GENERATING MEMORY EVALUATIONS FROM REAL ORDER HISTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalization is crucial to the agentic commerce experience, and is in turn dependent on consumer memory. Most benchmarks for memory in this space evaluate an agent downstream of memory, confounding memory quality with agent behavior and harness. We present a pipeline to generate consumer memory benchmarks at scale and PantryBench, the resulting benchmark from running the pipeline on the public Delivery Hero Recommendation Dataset (DHRD). Each question is drafted by an LLM constrained by up to five semantic axes (memory type, lookup type, verb, noun, situation). LLM judges then evaluate the question-answer pair against evidence in the consumer's order history. Questions deemed unsupported then become abstention questions to test for hallucination. A deterministic parser check with Stanza, an open-source NLP tool, confirms that each question matches the intent of its fixed semantic axes. Our pipeline makes question diversity measurable as variety (Vendi count over embeddings), coverage (over category pairs within semantic axes), and complexity (lookup type, and number and selection of constrained semantic axes). We find that without constraints from semantic axes, we observe diversity collapse in question-drafting which replicates in both a private cohort and DHRD. On the Stockholm cohort, a Vendi count over sentence embeddings finds 7.6 distinct questions per 40 exported free draws, against 12.2 to 14.8 for constrained questions. Our variety and coverage findings replicate across three LLM families (Claude Opus 5, GPT-5.6, and Gemini 3.5 Flash) and all three cities in DHRD. We release PantryBench, its generation pipeline, and artifacts from Mem0 and Letta evaluation.

1 INTRODUCTION

A successful commerce agent adapts to user preferences. Evaluating personalization is difficult because latent user preferences cannot be observed. We cannot see whether a user was satisfied with their last order, developed a distaste for certain foods, or changed their dietary goals. However, we do observe their ordering history. Most memory benchmarks focus on chat history (§2). Further, most memory-for-personalization benchmarks test behavior downstream of memory, such as recommendations, reviews, or held-out interactions (Salemi et al., 2024; Zhang et al., 2026). Such tests mix the ability of the backend LLM and its harness with memory quality. Instead, we ask what a fixed reader can answer from the stored content to isolate memory quality from agent behavior.

We present a pipeline that generates memory evaluation tasks at scale, where tasks are grounded in user order history (§3). For each task, the pipeline samples up to five semantic axes: memory type, lookup type, verb, noun, and situation. It then prompts an LLM to draft a question under these constraints. LLM judges verify the resulting question-answer pair against the order history. If a pair is unsupported by the order history, the question becomes an abstention question that tests for hallucination (Appendix C.3).

Because the axes are fixed for each question before drafting, we can measure the diversity of the resulting corpus (§4.2). We operationalize diversity as three properties. Variety is the Vendi count over sentence embeddings (Friedman & Dieng, 2023). Coverage is the share of realized category pairs within the semantic axes. Complexity is the number and type of constrained axes and the values

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

they take. We find that unconstrained drafting, with zero fixed semantic axes, collapses and shows much lower diversity. On the Stockholm cohort, a Vendi count over sentence embeddings finds 7.6 distinct questions per 40 exported free draws, against 12.2 to 14.8 for constrained questions. This finding replicates across a private cohort and the public DHRD, three LLM authoring families, and three cities (§4.2).

We ran our pipeline on DHRD and release the resulting benchmark, PantryBench (§4.1). DHRD contains 4.4 million food-delivery orders across Stockholm, Singapore, and Taipei, spanning thousands of vendors and over a million customers (Assylbekov et al., 2023). We also release the pipeline, per-run statistics, prompts, and evaluation artifacts from Mem0 (Chhikara et al., 2025) and Letta, formerly MemGPT (Packer et al., 2023) (§4.3).

Contributions. (1) A pipeline that generates memory evaluation tasks at scale, where tasks are grounded in order history and constrained by semantic axes (§3). (2) Diversity operationalized as three properties: variety, coverage, and complexity. We observe diversity collapse when question drafting is unconstrained by semantic axes (§4.2). (3) PantryBench, the result of running our pipeline on DHRD, released alongside the pipeline code, run statistics, prompts, and artifacts from evaluating Mem0 and Letta (§4.1, §4.3).

2 Related work

Commerce benchmarks. PACEShop (Lyu et al., 2026) evaluates controlled shopping responses with injected defects rather than questions derived from observed account histories. UQABench (Liu et al., 2025) asks personalized questions over real Taobao logs, but it scores user-embedding compression, and its questions are not evidence-bound or verified. Shopping Companion (Yu et al., 2026) tests cross-session preference memory over 1.2M real products, but through task success rather than questions about the memory. MemSim (Zhang et al., 2025) constructs reliable, diverse questions for memory evaluation, but from simulated users. To our knowledge, no existing work combines real order histories with specified, evidence-bound, verified questions for memory-system evaluation with scored abstention.

Existing memory benchmarks. Some benchmarks evaluate downstream personalization on real user histories (LaMP, Salemi et al., 2024, and MemoryCD, Zhang et al., 2026) and others ask questions based on synthetic conversations (Maharana et al., 2024; Wu et al., 2025; Zhao et al., 2025a; Jiang et al., 2025a;b). We address four open gaps: separating memory assessment from agent assessment, scoring abstention as a core metric, covering commerce histories, and measuring question diversity. Scored abstention follows the negative-rejection class of RGB (Chen et al., 2024), the missing-answer class of CRAG (Yang et al., 2024), and AbstentionBench (Kirichenko et al., 2025). Of the memory benchmarks in Table 5, only LongMemEval scores it. Appendix A surveys the landscape in detail (Table 5).

Specify a question before drafting it. DataMorgana (Filice et al., 2025) generates retrieval-augmented generation (RAG) questions from declared category distributions over corpus documents. Its user categories are personas that shape phrasing, and its evidence is a sampled corpus document rather than one account's history. It has no negative class. It filters generated questions for category adherence, but it reports no adherence rate and no realized category distribution. Competency questions (Grüninger & Fox, 1995) and behavioral testing (Ribeiro et al., 2020) motivate declaring what a question tests before writing it. Our grammar applies that discipline to per-account memory evidence and adds a measurement loop. Four established instruments measure the resulting questions: mean pairwise embedding similarity, the Vendi score (Friedman & Dieng, 2023), n-gram statistics, and parse-derived labels (Appendix A.3).

Compression via memory is lossy. Letta (Packer et al., 2023) and Mem0 (Chhikara et al., 2025) extract memories at write time and serve them through retrieval. MemoryAgentBench finds that long-context models perform best on test-time learning and long-range understanding (Hu et al., 2025). Other studies find advantages for search over uncompressed history (Zheng et al., 2025) and for verbatim chunks over extracted facts (An, 2026). HaluMem finds that errors introduced during extraction and updates propagate to later answers (Chen et al., 2025a). Raw-context accuracy can also degrade as histories grow longer (Modarressi et al., 2025). We add evidence from real commerce histories at two densities, with abstention scored separately (§4.3).

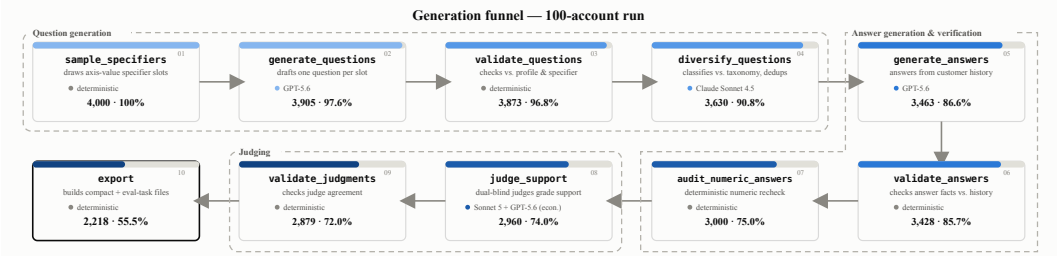

Figure 1: The generation pipeline as run on the private 100-account cohort (§4.1), laid out in the code's own module order. Each box names the deterministic check or LLM that produced its data, and the top bar shows retention relative to the 4,000 sampled slots. The numbers match Table 7.

3 Framework

Evaluation boundary. The memory system under test is a black box. Each memory approach receives the same input, the complete order-item history, and may store, index, compress, or summarize it in any way. The evaluation supplies a question and scores the answer against the exported content. Downstream tasks (recommendation, prediction) are out of scope, as are live catalogue lookups.

Semantic axes. Every question originates as a specification over five axes. *Memory function* (memory type) is episodic or semantic recall, following Tulving (1972). *Memory demand* (lookup type) is one of `retrieve`, `aggregate`, `compare`, `connect`, `temporal`, `infer`, or `assess`. *Predicate* (verb) is one of 42 behavioral claims such as `reorders`, `avoids`, or `explores`. *Target* (noun) is one of 82 subjects such as `cuisine`, `protein`, or `delivery_location`. *Ordering condition* (situation) is one of 9 conditioning contexts such as `cuisine`, `meal_period`, or `store_type`. Compatibility rules reject impossible combinations. For example, pairing the predicate `reorders` with the target `spice_level` is rejected because an account reorders dishes or vendors, not a spice level. Appendix C walks through a full specification. Each axis has a sampling *liberty*, the probability of leaving it unconstrained. This lets us attribute omission and quality effects to a specific axis. About 6% of specifications remain fully unconstrained as an unsteered control. Within an axis, draws are weighted toward categories which were less sampled for that account, so coverage is encouraged without being forced.

Diversity as three measured properties. We separate diversity into three properties, each with its own mechanism and its own reported measurement (§4.2). *Variety*: The questions differ in wording and form. The diversification stage caps repeated surface patterns and removes near-duplicates. A deterministic syntactic audit of all exported questions measures the realized distribution of question forms (Appendix C.5). *Coverage*: Over all axes, we report which enabled categories appear in at least one exported question. Coverage of category combinations across axes is smaller and is reported separately (Appendix F.4). *Complexity*: The memory demand sets most of the difficulty. It ranges from recalling a single fact to inferring a pattern the history never states. The categories of the other axes also differ in how much evidence they need. Every exported question carries its demand and its constrained semantic axes, so complexity is read off the specification.

Grounding and unanswerable questions. Questions are grounded in order history. A vendor-specific question, for example, is asked only about a vendor the account has ordered from, and a deterministic check rejects any value the history does not contain. Two independent LLM judges then confirm that the history supports each gold answer. A plausible question that cannot be supported by the history becomes an *unanswerable question* whose gold answer is "There is no evidence to answer this question". Answering such a question would be a hallucination, and the benchmark scores it as a missed abstention. Table 1 shows five exported contracts from the public PantryBench run (§4.1). They span five memory demands and all three evidence regimes, including required abstention (P5). Appendix C.2 shows a de-identified counterpart from the private export.

Question generation pipeline. Figure 1 shows the selectable, checkpointed modules. First, we sample the specifications for the questions based on the five semantic axes. Then, for each account, we use a single LLM call to generate a set of questions, with a prompt that provides, for example, relevant vendors and order items. In the next stage, we remove clear duplicates for each account, and

Table 1: Representative exported contracts from the public PantryBench Stockholm run (§4.1). The sampled specification lists memory function, memory demand, predicate, target, and ordering condition, in that order. "Free →" marks an unconstrained axis and the category the classification flywheel realized for it. Prices, and therefore the spend figures, are synthetic benchmark values (§4.1).

ID	Sampled specification	Question	Gold contract
P1	`free → semantic; aggregate; free; day_of_week; free`	Which day of the week do I place the most orders on?	Thursday. Significant evidence.
P2	`free → semantic; compare; varies_by; order_spend; store_type`	For my restaurant orders, how does what I spend vary between weekdays and weekends?	Weekend orders run about $2.20 higher per order (roughly 18% more). Significant evidence.
P3	`episodic; free → retrieve; free → orders; free → item; free`	What did I order on the very first order in my history?	Pommes frites and a Ramlösa citrus 33 cl, from Vendor 013758f5. Some evidence.
P4	`free → semantic; infer; includes_dessert; dessert_behavior; meal_period`	On my breakfast orders, what role do desserts seem to play?	A treat alongside the meal, not a substitute for it. Some evidence.
P5	`semantic; assess; owns; pet_ownership; free`	What do my past purchases suggest about whether I keep a pet at home?	Abstain. No evidence.

a separate classifier model, not the drafter, labels the axis values each question actually expresses. We use these labels to measure axis adherence to the specification and to define new taxonomy entries and compatibility rules (classification flywheel). Next, we generate answers using the full order history. Answers are requested to be succinct and separate from reasoning. To simplify audit and encourage grounding, each answer comes with up to six supporting facts as grounding evidence. We apply cost-saving measures such as prompt-cache warm-up, a lossless compact encoding of the order history that keeps every row, and batched requests. Where possible, a deterministic audit recomputes count, superlative, and comparison claims from the complete history. Last, two independent judges look at each question, answer, reasoning, supporting evidence, and order history, assess correctness, and propose changes where needed. If both judges agree, questions are exported for further review or evaluation (details in Figure 3 and Appendix B).

Memory system scoring. Candidate answers are graded separately by an LLM judge on answer correctness, and agreement with the expected evidence level (significant, some, or none). We categorize errors into different types such as `SHOULD_ABSTAIN`, `UNWARRANTED_ABSTENTION`, and `WRONG_SCOPE`.

4 EXPERIMENTS

4.1 DATA AND SETTINGS

Private cohort. The private study uses a randomly sampled cohort of 100 accounts from a production food and grocery delivery platform. Their histories hold 112,941 order-item rows in 36,516 orders, a median of 214 orders per account, and a median of two item rows per order. The history keeps all 27 audited columns of the source table. These are three identifiers, seven order-timing fields, six merchant fields, four delivery-location fields, and seven item fields for names, categories, quantities, selected options, charged prices, and ratings. The taxonomy enabled 137 category values across the five axes, and every value is groundable in this source. The histories cannot be released.

Public cohort. PantryBench applies the unchanged framework to the Delivery Hero Recommendation Dataset (DHRD), an MIT-licensed release of roughly 9.4 million food-delivery order-item rows from Singapore, Stockholm, and Taipei over about 90 days (Assylbekov et al., 2023). From each city we randomly select 100 customers with at least 20 orders. The Stockholm cohort has roughly 23× fewer order-item rows than the private cohort. DHRD lacks eight of the private cohort's 24 behavioral fields, such as item options, group-order status, item category, ratings, and location details. The missing item options and group-order context leave 125 of the 142 enabled category values[1]

[1]Five values were added to the taxonomy after the private run, which drew from 137.

groundable, and the remaining 17 are excluded from sampling before any model call. The cohort builder translates field names and values into the private data format. For example, relative day offsets become calendar timestamps from an arbitrary epoch, normalized prices become synthetic dollar values, vendor identifiers become pseudonymous vendor labels, and the order geohash serves as the delivery location (Appendix G.2). From the Stockholm questions we also reviewed a subset of 253 questions by hand. We filtered this subset for clarity, consistent grading, and discriminative power between memory systems.

Question generation pipeline settings. For each of the 100 accounts, we sample 40 specifications. The current registry has 127 compatibility rules that leave 26,058 maximally specified combinations of the roughly 434,000 nominal ones. GPT-5.6 drafts questions and gold answers on the private cohort. On PantryBench, we use Claude Opus 5 in all three cities, and on the Stockholm cohort we additionally use GPT-5.6 and Gemini 3.5 Flash. The three Stockholm runs share the cohort, the run seed, and byte-identical sampled specifications, so they differ only in the authoring model. Singapore and Taipei rerun the Stockholm profile unchanged. The diversification classifier (Claude Sonnet 4.5) and the two blind judges (Claude Sonnet 5 and a GPT-5.6 economy tier) are fixed across all runs.

Memory system settings. Two snapshot studies use the exported questions. On PantryBench, we compare six public configurations: raw history, a SQL agent, Letta core memory (Packer et al., 2023), and three Mem0 variants (Chhikara et al., 2025) (§4.3). The Mem0 variants are a default extraction store, the same store under Mem0's pattern-steering instructions, and the store read through its top-10 retrieval path. For ingesting the data into the memory systems, we use a template that maps all fields of an event into a first-person sentence, and then we add the events one by one. To isolate the stored representations, we use a single reading adapter (Claude Sonnet 4.5) that answers every question from the exported content, followed by a fixed judge of the same model that scores the answers. We could not route Letta through a GPT-family reader for technical reasons. We expect within-family grading bias to be small, because the answers are objective and the rubrics are strict. On the private cohort, seven configurations answer through a shared GPT-5.6 reader and are scored by Claude Sonnet 4.5 (Appendix F.6). These are three anonymized memory systems (Systems A–C), each alone and each combined with the raw order history, plus the raw history alone.

4.2 Question generation pipeline results

Results are displayed in Table 2 and Figure 2.

Generation funnel. Out of the 4,000 candidates per run, we keep 38.9% to 66.8% for the export. The largest losses come from quality gates, led by judge disagreement and followed by the deterministic numeric audit. Tables 7 and 18 list the typed reasons per stage. Order histories in the private cohort are much longer than in the DHRD cohort, which results in more removals, especially in the numeric audit stage, because of context rot (Hong et al., 2025; Liu et al., 2024). Gemini 3.5 Flash loses about half of its valid judgments to judge disagreement, which we attribute to model quality (Table 18). The share of exported questions that require abstention depends mostly on the authoring model. Claude Opus 5 produces the fewest in all three cities, GPT-5.6 lands close together on both cohorts, and Gemini 3.5 Flash produces the most (Table 2). The classification of questions shows very high axis adherence, with slight differences between authoring models. In every run, memory demand is the axis that most often misses its requested value. The resulting flywheel's suggestions led to five new taxonomy entries after the private run.

Coverage. Coverage depends on cohort size and sampling volume. For example, on a 10-account diagnostic run, eight target categories had zero coverage and one (`spice_level`) was discovered only through free questions. For the reported runs, every category was covered. *Pair coverage* is above 80% except for Gemini 3.5 Flash. Given that there are 2,264 to 2,342 potential pairs, a higher sampling volume could raise it substantially.

Question shape. A deterministic syntactic audit with the Stanza parser (Qi et al., 2020), which uses no LLM, confirms that the requested axes are visible in the wording of the questions (Appendix C.5). The same audit shows that each authoring model keeps its own question shape across cohorts and cities (Table 3). GPT-5.6 writes the same shape of question on private and public data. Claude Opus 5 writes longer questions with more subordinate clauses than GPT-5.6 in all three cities. The median gold answer is at least twice as long under Claude Opus 5 as under GPT-5.6 and Gemini 3.5 Flash.

Table 2: Question generation outcomes. *Yield* is the share of the 4,000 sampled specifications that end as exported questions. *Axis adherence* is the share of requested axis values that the exported questions realize. *Unanswerable* is the share of exported questions that the judges confirm to have no supporting evidence. *Judge disagreement* is the number of questions removed at export because the two judges disagree. *Pair coverage* is the share of all sampleable pairs of axis values that appear in at least one exported question. *Vendi* is the effective number of different questions in a subsample of 40 exported questions. *Vendi free* covers the questions with no constrained axis, and *Vendi constrained* covers the questions with one to five constrained axes (see also Tables 27 and 22).

Cohort	Authoring family	Yield	Axis adherence	Unanswer-able	Judge disagr.	Pair coverage	Vendi free	Vendi constrained
Private	GPT-5.6	55.5%	99.6%	13.8%	661	80.5%	7.6	12.4–15.5
Stockholm	Claude Opus 5	61.1%	99.1%	9.2%	886	81.3%	7.6	12.2–14.8
Stockholm	GPT-5.6	66.8%	99.6%	16.2%	668	83.8%	6.4	12.6–14.8
Stockholm	Gemini 3.5 Flash	38.9%	98.8%	22.2%	1,558	72.0%	9.4	13.4–16.2
Singapore	Claude Opus 5	58.6%	99.2%	9.3%	1,017	81.1%	6.3	10.3–13.8
Taipei	Claude Opus 5	63.4%	99.3%	7.7%	915	82.6%	6.8	11.1–14.2

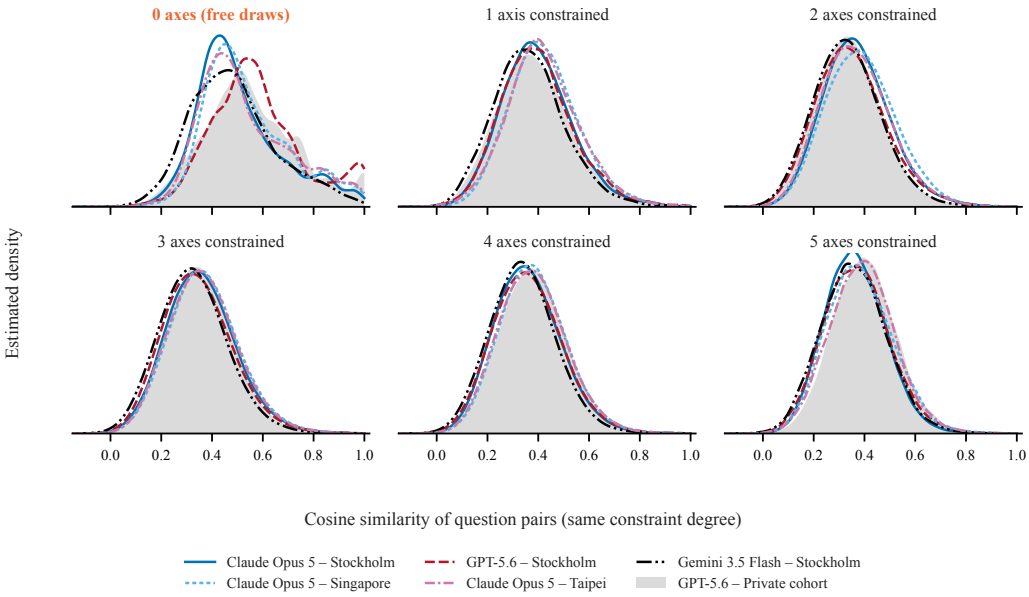

Figure 2: Unconstrained sampling collapses in embedding space, across authoring families and data sources. Each curve is the kernel density of pairwise cosine similarity (all-MiniLM-L6-v2, Wang et al., 2020; Reimers & Gurevych, 2019) between drafted questions of one run with the same number of constrained axes, one panel per constraint degree.

Closed yes/no phrasings are common only under Gemini 3.5 Flash, rare under Claude Opus 5, and absent under GPT-5.6.

Variety. Vendi scores show that free draws are by far the most self-similar population, by a factor of roughly 1.5 to 2, and that one constraint recovers most diversity. For the reported runs in Figure 2, constraining a single axis already removes most of the collapse of question variety (Table 22). At every constrained degree the curves nearly coincide, with Gemini 3.5 Flash slightly to the left. The runs differ most in the free-draw panel, and Table 25 quantifies this spread per constraint degree. GPT-5.6 sits furthest right and is the only family with a near-duplicate mode at $\sim$1.0, and the private cohort, also authored by GPT-5.6, follows it (Figure 8 in Appendix F.2 shows the private run alone). In an explorative analysis on the private cohort, we discovered that free draws also drift toward a few categories (Appendix F.2, Table 15). They concentrate on `retrieve` and `aggregate` for memory demand, the `semantic` memory function, the `orders` predicate, and the `item` target. In

Table 3: Surface statistics of the exported corpora, measured with the deterministic Stanza labeler of Appendix C.5. Q tokens: median question length in tokens. Parse depth: mean dependency-parse depth. Subord. : mean subordinate-clause count per question. Yes/no: count of closed yes/no phrasings. Temporal and Compar. : shares of questions with time reference and comparative marking. A words: median gold-answer length in whitespace-delimited words.

Cohort	Authoring family	Q tokens	Parse depth	Subord.	Yes/no	Temporal	Compar.	A words
Private	GPT-5.6	14	3.17	0.52	0	29.0%	5.1%	4
Stockholm	Claude Opus 5	17	3.65	0.75	2	46.7%	9.9%	8
Stockholm	GPT-5.6	14	3.13	0.52	0	32.3%	7.0%	4
Stockholm	Gemini 3.5 Flash	16	3.77	0.67	203	32.7%	2.8%	3
Singapore	Claude Opus 5	18	3.69	0.75	2	44.8%	11.0%	8
Taipei	Claude Opus 5	18	3.67	0.77	2	46.9%	10.9%	8

Table 4: Memory configurations scored on PantryBench and on the private cohort, one panel each. Scores are percentages. Within a panel, bold marks the largest Answer, Reviewed, and Abstain value and italics the smallest. Appendix E defines every column.

Memory approach	Answer [95% CI]	Reviewed	Abstain [95% CI]	≥4 axes	Temporal	Aggregate	Store	Context	Latency	Unique	Wins
PantryBench Stockholm, Claude Sonnet 4.5 reader (1,931 answerable, 225 unanswerable, of which 247 and 6 are hand-reviewed)											
Raw events	**63.3** [61.0, 65.7]	71.7	93.8 [90.5, 96.5]	59	53	66	6.3k	3.1k	6.8	39	24
SQL agent	60.2 [58.2, 62.3]	**79.8**	*88.0* [83.8, 91.9]	57	55	73	raw rows	5.0k	9.6	124	20
Letta	59.8 [57.0, 62.5]	68.8	95.6 [92.3, 98.2]	55	48	64	7.2k	4.0k	7.2	32	7
Mem0	57.1 [54.8, 59.2]	61.1	95.1 [92.1, 97.6]	54	45	63	6.2k	3.8k	7.0	22	1
Mem0 steered	41.8 [39.2, 44.4]	*40.5*	95.6 [92.9, 97.9]	37	29	47	6.4k	4.1k	7.0	24	0
Mem0 retrieval	*40.0* [37.4, 42.5]	41.3	**98.2** [96.4, 99.6]	42	22	40	6.2k	2.1k	6.0	24	0
Private cohort, GPT-5.6 reader (1,911 answerable, 307 unanswerable)											
Raw orders	66.5 [64.3, 68.7]	—	61.6 [55.9, 67.1]	63	63	73	—	46.6k	9.0	38	15
System A + orders	**67.8** [65.3, 70.3]	—	*52.8* [47.0, 58.6]	64	63	71	—	138.2k	12.1	39	25
System B + orders	65.9 [63.8, 68.2]	—	55.0 [48.7, 61.3]	62	64	69	—	49.6k	8.6	22	12
System C + orders	64.2 [61.7, 66.6]	—	60.3 [55.3, 65.5]	58	63	68	—	49.6k	9.5	23	4
System A	54.5 [51.8, 57.1]	—	63.8 [58.9, 68.7]	52	52	51	—	95.5k	9.1	16	0
System B	19.9 [17.6, 22.3]	—	82.7 [78.4, 87.0]	20	11	28	—	3.1k	3.8	10	0
System C	*14.6* [12.7, 16.5]	—	**88.9** [85.4, 92.2]	12	9	23	—	3.4k	3.5	10	0

contrast, constrained axes stayed well distributed after filtering. This shows that unsteered generation decreases variety by repeating the same question types and leaving certain question types untested, whereas our approach with semantic axes creates diverse questions.

4.3 Memory system scoring results

We report results on PantryBench and on its hand-reviewed subset (§4.1), shown as the Answer and Reviewed columns of Table 4.

Intuitively, raw order history performs quite well both for our private and the DHRD consumers as any sort of memory compression is lossy. On the hand-reviewed subset the SQL agent leads instead (79.8% versus 71.7% for raw history). While Letta and Mem0 stores have similar characters per account as raw history, they perform worse on the benchmark due to paraphrasing the ingestion text. This result is consistent with prior evidence of information loss during extraction (Hu et al., 2025; Zheng et al., 2025; An, 2026).

Every public configuration abstains correctly on 88.0% to 98.2% of the unanswerable questions. As expected, the systems that store the most material abstain the least (raw order history 93.8%, SQL 88.0%), and vice versa (Mem0 top-10 retrieval, 98.2%). This tradeoff is more significant in our private cohort given the long order histories, which can make the agent hallucinate (52.8–88.9% correct abstention, against 88.0% to 98.2% on PantryBench).

Memory demand orders the results more sharply than the approaches do. Every public configuration of Table 4 scores lower on time-ordered questions than on aggregate ones. Temporal accuracy spans 22–55% across configurations, against 40–73% for aggregates. The same ordering holds for six of the seven private configurations, with standalone System A the one exception. Aggregation therefore survives compression better than event sequence does, on both cohorts. Constrained questions (≥4

axes) and temporal questions score below each configuration's overall accuracy on both cohorts, apart from two near ties. Aggregate questions score above it for every configuration except standalone System A. Constrained axes and temporal demand therefore mark the hard questions, and aggregates the easier ones.

The *Unique* and *Wins* columns of Table 4 show that the questions distinguish the systems beyond their overall accuracy. On both cohorts, every configuration solves some questions that no other configuration answers. The slice wins show the strengths and weaknesses of each approach by category. A benchmark that separates approaches this way can guide the choice of a memory system for a given workload, which one accuracy number cannot.

5 Discussion

We discuss limitations and future work below.

Limitations. The semantic axes bound what the questions are able to test. For example, we do not have substitution behavior or seasonal item trends in our semantic axes. While the flywheel helps to sample all realized axes, the axes still bound the diversity of the generated corpus and it must be checked that the axes span what you want to measure.

Because of our extensive verification pipeline, resulting questions skew towards objective lookup questions as opposed to more subjective questions that admit a variety of correct answers. An example is "which cuisine do I order from most" vs. "have I been eating healthier lately". While this property makes our question-answer pairs trustworthy, arguably there are some aspects of memory useful for personalization contexts that go untested for this reason.

Many stages of our pipeline are LLM-dependent, making model intelligence crucial in ensuring quality of questions drafted as well as correct validation of the question-answer pair. In practice, we have found that different LLM families show distinct drafting styles and correct question validation is somewhat impacted by LLM quality. See also Appendix H.

Future work. While this pipeline has proven useful in the food-delivery domain, we would like to extend this framework to other domains with structured event logs (retail, browsing history, and others). We also leave more granular proxies for diversity and validity to future work. Whereas our focus was to have verified questions, we are interested in finding ways to alter the pipeline so that we can have more subjective questions. Those questions are closer to personalization, like discovery of trends and their changes.

6 Conclusion

In this paper, we introduced a framework to generate memory benchmarks from order histories and PantryBench, its public artifact release. The framework generates a diverse set of questions based on real order histories using five semantic axes. A classification flywheel labeled the axes each question realized and its suggestions extended the taxonomy. We observed that 99.6% of the requested semantic axis values are met. The semantic axes improve question variety as shown by the Vendi count over sentence embeddings. We found 7.6 distinct questions per 40 free draws, against 12.4 to 15.5 for constrained questions. Our measured coverage shows a similar effect. Every enabled category of the five axes and 80.5% of the 2,342 sampleable category pairs appear in an exported questions set. The replication on a public dataset and the ablations over cities and three generating models confirm these findings.

The questions separate memory representations. Across seven exported-content configurations, the one with the highest accuracy on answerable questions abstains correctly least often. In the private case study, no approach leads on every question type, and the breakdown by axes shows where each representation loses recall. We release PantryBench, its pipeline, and the memory-system outputs for others to build on.

AI USE STATEMENT

Generative AI tools assisted the authors with research-method design, software development, analysis workflows, literature review, and manuscript editing. Separately, generative models were used as reported components of the question-generation, answering, and judging pipeline. Sections 3–4 describe those experimental uses. The authors reviewed all AI-assisted work, verified code with automated tests and reproducible analyses, and checked claims against cited sources and generated artifacts. The authors take responsibility for the final content.

ETHICS STATEMENT

This work evaluates memory representations derived from private customer order histories. The submission reports only aggregate results and does not release customer-level data. Raw identifiers are excluded from generated questions, answers, reasoning, and judge output, and export validation rejects identifier leaks. The study does not support person-level inference from account-level behavior. Appendix H describes the remaining privacy, governance, and bias limitations, which §5 summarizes.

REPRODUCIBILITY STATEMENT

Appendices B–E document the pipeline configuration, answer contracts, and evaluation harness, and Appendix F documents the metrics and statistical analyses used in this study. Every run fingerprints the code revision, configuration, prompts, and model routes, and statistics reports are stable under reruns. Because the underlying customer histories cannot be released, external reproduction of the private-cohort results is not currently possible. Appendix H states this boundary. The DHRD replication (Appendix G) removes the private-data barrier for the generation-side results. It builds on a pinned, checksummed snapshot of a public dataset through a deterministic cohort builder, and its provenance is archived like every other run's. We release this replication as PantryBench. The release contains the generation framework code, the cohort builder with its pinned manifests, the generated question-answer corpus, and the complete run statistics and results. The release also includes the memory-system evaluation outputs and the analysis script that reproduces the tables of Appendix G.8 and the PantryBench panel of Table 4. The private panel of that table needs the private statistics and is not externally reproducible. Only cost telemetry stays internal.

REFERENCES

James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26 (11):832–843, 1983. URL https://doi.org/10.1145/182.358434.

Tao An. Fidelity before structure: Verbatim chunks beat lossy artifact extraction in long-conversation LLM memory. *arXiv preprint arXiv:2601.00821*, 2026. URL https://arxiv.org/abs/2601.00821.

Yernat Assylbekov, Raghav Bali, Luke Bovard, and Christian Klaue. Delivery hero recommendation dataset: A novel dataset for benchmarking recommendation algorithms. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, pp. 1042–1044, 2023. URL https://doi.org/10.1145/3604915.3610242.

Wicher Bergsma. A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, 42(3):323–328, 2013. URL https://doi.org/10.1016/j.jkss.2012.10.002.

Tara Bogavelli, Roshnee Sharma, and Hari Subramani. AgentArch: A comprehensive benchmark to evaluate agent architectures in enterprise. *arXiv preprint arXiv:2509.10769*, 2025. URL https://arxiv.org/abs/2509.10769.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. Large language models empowered personalized web agents. In *Proceedings of the ACM Web Conference 2025 (WWW)*, 2025. URL https://doi.org/10.1145/3696410.3714842.

Shuyang Cao and Lu Wang. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 6424–6439, 2021. URL https://aclanthology.org/2021.acl-long.502/.

Agneet Chatterjee, Rahim Entezari, Maksym Zhuravinskyi, Maksim Lapin, Reshinth Adithyan, Amit Raj, Chitta Baral, Yezhou Yang, and Varun Jampani. Stable cinemetrics: Structured taxonomy and evaluation for professional video generation, 2025. URL https://arxiv.org/abs/2509.26555.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. HaluMem: Evaluating hallucinations in memory systems of agents. *arXiv preprint arXiv:2511.03506*, 2025a. URL https://arxiv.org/abs/2511.03506.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024. URL https://doi.org/10.1609/aaai.v38i16.29728.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. *arXiv preprint arXiv:2506.02592*, 2025b. URL https://arxiv.org/abs/2506.02592.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946. URL https://press.princeton.edu/books/paperback/9780691005478/mathematical-methods-of-statistics.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*, pp. 152–164, 2024. URL https://aclanthology.org/2024.sighan-1.18/.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2368–2378, 2019. URL https://aclanthology.org/N19-1246/.

Gilles O. Einstein and Mark A. McDaniel. Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):717–726, 1990. URL https://doi.org/10.1037/0278-7393.16.4.717.

Chase M. Fensore, Kaustubh Dhole, Jason Fan, Eugene Agichtein, and Joyce C. Ho. How fine-grained should a RAG benchmark be? a hierarchical framework for synthetic question generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3715–3720. ACM, July 2026. URL https://doi.org/10.1145/3805712.3809925.

Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. Generating diverse Q&A benchmarks for RAG evaluation with DataMorgana, 2025. URL https://arxiv.org/abs/2501.12789.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

Yifan Gao, Lidong Bing, Wang Chen, Michael R. Lyu, and Irwin King. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. URL https://arxiv.org/abs/1807.03586.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. URL https://aclanthology.org/2020.findings-emnlp.117/.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. URL https://arxiv.org/abs/1803.09010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. URL https://arxiv.org/abs/1705.08500.

Michael Grüninger and Mark S. Fox. The role of competency questions in enterprise engineering. In *Benchmarking — Theory and Practice*, IFIP Advances in Information and Communication Technology, pp. 22–31. Springer, 1995. URL https://doi.org/10.1007/978-0-387-34847-6_3.

William Guey and Pierrick Bougault. Self-preference is weak or absent in verifiable instruction-following revision: A four-model test under genuine authorship. *arXiv preprint arXiv:2606.20093*, 2026. URL https://arxiv.org/abs/2606.20093.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 2025. URL https://aclanthology.org/2025.tacl-1.69/.

Zeying Hao, Hao Guo, Mengtao Xu, Yimin Hu, Yuheng Song, Zesheng Zhou, Jinsong Lan, and Xiaoyong Zhu. MMShopBench: A real-log benchmark for multimodal, multi-turn shopping agents. *arXiv preprint arXiv:2607.29002*, 2026. URL https://arxiv.org/abs/2607.29002.

Hang He, Chuhuai Yue, Chengqi Dong, Mingxue Tian, Hao Chen, Zhenfeng Liu, Jiajun Chai, Xiaohan Wang, Yufei Zhang, Qun Liao, Guojun Yin, Wei Lin, Chengcheng Wan, Haiying Sun, and Ting Su. LocalSearchBench: Benchmarking agentic search in real-world local life services. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. URL https://arxiv.org/abs/2512.07436. arXiv:2512.07436.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13: 18617, 2023. URL https://doi.org/10.1038/s41598-023-45644-9.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts LLM performance. Technical report, Chroma, 2025. URL https://research.trychroma.com/context-rot.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025. URL https://arxiv.org/abs/2507.05257.

Yuhan Hu, Zhouyang Long, Jiayi Guo, Xin Sui, Xiao Fu, Wenbo Zhao, Yanyan Zhao, and Bing Qin. OP-Bench: Benchmarking over-personalization for memory-augmented personalized conversational agents. *arXiv preprint arXiv:2601.13722*, 2026. URL https://arxiv.org/abs/2601.13722.

Yeonjun In, Wonjoong Kim, Sangwoo Park, Kanghoon Yoon, and Chanyoung Park. Personalize-then-store: Benchmarking and learning personalized memory for long-horizon agents. *arXiv preprint arXiv:2605.25535*, 2026. URL https://arxiv.org/abs/2605.25535. Introduces the PerMemBench benchmark.

Vijay Janapa Reddi. Architecting the agentic AI systems stack: What should infrastructure manage when the unit of work is a trajectory? *ACM SIGOPS Operating Systems Review*, 60(1):60–75, 2026. URL https://doi.org/10.1145/3830422.3830429.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025a. URL https://arxiv.org/abs/2504.14225.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, et al. PersonaMem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b. URL https://arxiv.org/abs/2512.06688.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2506.09038.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, 2024. URL https://aclanthology.org/2024.findings-acl.29/.

D. Richard Kuhn, Raghu N. Kacker, and Yu Lei. Combinatorial coverage measurement. Technical Report NISTIR 7878, National Institute of Standards and Technology, 2012. URL https://doi.org/10.6028/NIST.IR.7878.

Xiaochong Lan, Jie Feng, Jiahuan Lei, Xinlei Shi, and Yong Li. LocalGPT: Benchmarking and advancing large language models for local life services in Meituan. *arXiv preprint arXiv:2506.02720*, 2025. URL https://arxiv.org/abs/2506.02720. KDD 2025.

Tom Leinster and Christina A. Cobbold. Measuring diversity: The importance of species similarity. *Ecology*, 93(3):477–489, 2012. URL https://doi.org/10.1890/10-2402.1.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025. URL https://arxiv.org/abs/2502.01534.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 110–119, 2016. URL https://aclanthology.org/N16-1014/.

Langming Liu, Shilei Liu, Yujin Yuan, Yizhen Zhang, Bencheng Yan, Zhiyuan Zeng, Zihao Wang, Jiaqi Liu, Di Wang, Wenbo Su, Pengjie Wang, Jian Xu, and Bo Zheng. UQABench: Evaluating user embedding for prompting LLMs in personalized question answering. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5652–5661, 2025. URL https://doi.org/10.1145/3711896.3737385.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Shuhang Liu, Jiayuan Zhu, Lei Shu, Jialong Lin, Yiheng Chen, Haozhe Zhang, Cheng Zhang, Derong Xu, Jielin Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. PERMA: Benchmarking personalized memory agents via event-driven preference and realistic task environments. *arXiv preprint arXiv:2603.23231*, 2026. URL https://arxiv.org/abs/2603.23231.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.153/.

Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496, 2010. URL https://doi.org/10.1075/ijcl.15.4.02lu.

Weimin Lyu, Chen Luo, Guangrui Li, Yaochen Xie, Dhineshkumar Ramasubbu, Arief Koesdwiady, Wanqiu Long, Hansu Gu, Yutong Chen, Zheshen Wang, Dakuo Wang, and Yi Liu. PACEShop: Evaluating personalized, actionable, compositional, and evidence-grounded shopping assistants. *arXiv preprint arXiv:2608.26180*, 2026. URL https://arxiv.org/abs/2608.26180.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*, 2024. URL https://arxiv.org/abs/2402.17753.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025. URL https://arxiv.org/abs/2502.05167.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023. URL https://doi.org/10.6028/NIST.AI.100-1.

Thomas O. Nelson and Louis Narens. Metamemory: A theoretical framework and new findings. In *The Psychology of Learning and Motivation*, volume 26, pp. 125–173. Academic Press, 1990. URL https://doi.org/10.1016/S0079-7421(08)60053-5.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pp. 4034–4043, 2020. URL https://aclanthology.org/2020.lrec-1.497/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Feiz5HtCD0.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024. URL https://arxiv.org/abs/2404.13076. NeurIPS 2024.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the Vendi score: A family of similarity-based diversity metrics for science and machine learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. URL https://arxiv.org/abs/2310.12952.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108, 2020. URL https://aclanthology.org/2020.acl-demos.14/.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016. URL https://aclanthology.org/D16-1264/.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018. URL https://aclanthology.org/P18-2124/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. URL https://aclanthology.org/D19-1410/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4902–4912, 2020. URL https://aclanthology.org/2020.acl-main.442/.

William S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950. URL https://doi.org/10.2307/2087176.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems, 2024. URL https://arxiv.org/abs/2311.09476.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7370–7392, 2024. URL https://aclanthology.org/2024.acl-long.399/.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://aclanthology.org/2024.emnlp-main.368/.

Chantal Shaib, Venkata S. Govindarajan, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL): System Demonstrations*, 2025. URL https://aclanthology.org/2025.ijcnlp-demo.5/.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. Play favorites: A statistical method to measure self-bias in LLM-as-a-judge. *arXiv preprint arXiv:2508.06709*, 2025. URL https://arxiv.org/abs/2508.06709.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4208–4219, 2018. URL https://aclanthology.org/D18-1453/.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2309.02427.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.12150.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025. URL https://aclanthology.org/2025.findings-acl.989/.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 326–346, 2021. URL https://aclanthology.org/2021.eacl-main.25/.

Jiahao Tian and Zhenkai Wang. MARS: Multi-agent re-ranking for repeat-order food delivery recommendation. *arXiv preprint arXiv:2607.25420*, 2026. URL https://arxiv.org/abs/2607.25420.

Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson (eds.), *Organization of Memory*, pp. 381–403. Academic Press, New York, 1972. URL https://alicekim.ca/EMSM72.pdf.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Guangyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. *arXiv preprint arXiv:2604.20006*, 2026. URL https://arxiv.org/abs/2604.20006. Introduces the Memora benchmark.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. URL https://papers.nips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024. URL https://arxiv.org/abs/2410.21819.

World Wide Web Consortium. PROV-O: The PROV ontology. W3C Recommendation, 2013a. URL https://www.w3.org/TR/prov-o/.

World Wide Web Consortium. SPARQL 1.1 query language. W3C Recommendation, 2013b. URL https://www.w3.org/TR/sparql11-query/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2025. URL https://arxiv.org/abs/2410.10813.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197, 2022. URL https://aclanthology.org/2022.acl-long.356/.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of ACL*, 2024. URL https://arxiv.org/abs/2402.11436.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. CRAG – comprehensive RAG benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.04744.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2207.01206.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.12045.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. Genie: Achieving human parity in content-grounded datasets generation, 2024. URL https://arxiv.org/abs/2401.14367.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. URL https://arxiv.org/abs/1809.08887.

Zijian Yu, Kejun Xiao, Huaipeng Zhao, Tao Luo, and Xiaoyi Zeng. Shopping companion: Benchmarking and training LLM agents for long-horizon preference-grounded e-commerce tasks. *arXiv preprint arXiv:2603.14864*, 2026. URL https://arxiv.org/abs/2603.14864.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3): 1–37, 2023. URL https://doi.org/10.1145/3617680.

Weizhi Zhang, Xiaokai Wei, Wei-Chieh Huang, Zheng Hui, Chen Wang, Michelle Gong, and Philip S. Yu. MemoryCD: Benchmarking long-context user memory of LLM agents for lifelong cross-domain personalization, 2026. URL https://arxiv.org/abs/2603.25973.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. URL https://arxiv.org/abs/1809.05972.

Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. MemSim: A Bayesian simulator for evaluating memory of LLM-based personal assistants. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2409.20163.

Siyan Zhao, Mingyi Nguyen, and Aditya Grover. Do LLMs recognize your preferences? Evaluating personalized preference following in LLMs. *arXiv preprint arXiv:2502.09597*, 2025a. URL https://arxiv.org/abs/2502.09597.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025b. URL https://aclanthology.org/2025.findings-acl.927/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Yicong Zheng, Kevin L. McKee, Thomas Miconi, Zacharie Bugaud, Mick van Gelderen, and Jed McCaleb. Goal-directed search outperforms goal-agnostic memory compression in long-context memory tasks. *arXiv preprint arXiv:2511.21726*, 2025. URL https://arxiv.org/abs/2511.21726.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating structured queries from natural language using reinforcement learning, 2017. URL https://arxiv.org/abs/1709.00103.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. URL https://doi.org/10.1609/aaai.v38i17.29946.

A Extended related work

A.1 What memory benchmarks actually evaluate

Table 5 classifies prominent benchmarks by the memory construct they address, their data source, and whether they evaluate the memory itself or only behavior downstream of it. The dominant pattern scores only the final response ("indirect" in Table 5). The benchmark supplies a history, the system responds, and nothing upstream of the answer is examined. This creates several ambiguities. (1) Long context can stand in for memory: a system may score well by retaining the entire transcript in its context window. (2) The score credits the wrong part: a broken memory can produce a right answer by chance, and a good memory loses points when the model ignores it. (3) Over-retention is unpenalized: storing everything is a winning strategy despite its cost and privacy footprint. (4) Abstention is rarely scored: a system that guesses when evidence is missing is not distinguished from one that declines. Retrieval benchmarks do score it. RGB (Chen et al., 2024) tests negative rejection when the retrieved documents hold no answer. CRAG (Yang et al., 2024) gives a missing answer zero and a wrong answer a penalty, and it includes false-premise questions. AbstentionBench (Kirichenko et al., 2025) collects 20 datasets of unanswerable, underspecified, or false-premise questions and finds that reasoning fine-tuning degrades abstention. Our no-evidence class plays the same role for memory evaluation, with the absence of supporting history as the ground truth. LongMemEval (Wu et al., 2025) is the one memory benchmark in Table 5 that scores abstention, and it grounds abstention in synthetic chats. PerMemBench (In et al., 2026) is the clearest attempt at direct storage-policy evaluation. It is fully synthetic and scores retention and gating. We score answers and abstention over real histories. OP-Bench (Hu et al., 2026) penalizes inappropriate memory *use*. MemBench (Tan et al., 2025) adds capacity and efficiency, measured as read and write time. Memora (Uddin et al., 2026) grounds scoring in gold memory traces with forgetting-aware accuracy. Its gold traces are simulated. Our gold answers are computed from observed orders.

Table 5: Benchmark landscape for memory and personalization. "Evaluates" abbreviations: I = indirect (final response only), C = component-level, M = mixed, D = direct (memory bank itself).

Benchmark	Memory addressed	Data source	Main metrics	Evaluates
LaMP (Salemi et al., 2024)	profile/style from documents	real user histories	task acc./F1, ROUGE	I
LoCoMo (Maharana et al., 2024)	episodic, temporal	LLM-gen., human-edited	QA F1, ROUGE, BLEU	I
PerLTQA (Du et al., 2024)	semantic + episodic	LLM-gen., validated	class./retrieval/fusion	M
LongMemEval (Wu et al., 2025)	episodic, updates, abstention	synthetic chats	answer acc., recall	M
PrefEval (Zhao et al., 2025a)	explicit/implicit preferences	curated pairs	pref.-following acc.	I
PersonaMem (Jiang et al., 2025a)	dynamic profiles	simulated histories	MCQ accuracy	I
PersonaMem-v2 (Jiang et al., 2025b)	implicit preference memory	fully GPT-5-generated	MCQ + judge acc.	I
PersonaLens (Zhao et al., 2025b)	persona in interactive tasks	synthetic profiles	judge scores	I
MemBench (Tan et al., 2025)	factual + reflective	constructed graphs	MCQ, R@10, R/W time	M
MemoryAgentBench (Hu et al., 2025)	retrieval, TTL, forgetting	adapted + synthetic	acc., R@5, F1	M
PERMA (Liu et al., 2026)	event-driven preferences	synthetic timelines	MCQ + task perf.	I
Memora (Uddin et al., 2026)	consolidation, deletion	simulated, verified	forgetting-aware acc.	I/M
PerMemBench (In et al., 2026)	storage policy, retention	fully synthetic	retention rate, gating F1	D
OP-Bench (Hu et al., 2026)	over-personalization	synthetic, human-verified	judge + diversity scores	I

A second axis is data provenance. Real histories remain the minority. LaMP (Salemi et al., 2024) and MemoryCD (Zhang et al., 2026) demonstrate that real user data (documents and Amazon interaction histories, respectively) is a viable data source. Both, however, score downstream personalization against held-out interactions rather than asking questions about the memory. UQABench (Liu et al., 2025) comes closest on the data side. It asks personalized questions over 184K Taobao users and 31M real interactions and releases the data. Its object of study is whether a compressed user embedding preserves enough signal for the questions. The questions are generated from the logs without evidence binding or per-question verification, and there is no abstention class. PersonalWAB (Cai et al., 2025) builds personalized search, recommendation, and review tasks from real shopping behaviors, again scoring task outcomes. Shopping Companion (Yu et al., 2026) evaluates cross-session preference memory of shopping agents over 1.2M real products through task success and process rewards. MemSim (Zhang et al., 2025) is the closest prior work on question construction. It generates reliable and diverse QA pairs for memory evaluation from a Bayesian relation network over simulated users, with single-hop, multi-hop, comparative, and aggregative types. We share its goal of reliable, typed questions and differ in the evidence, which is an observed account history rather than a generated user. The synthetic majority does ask questions, but it inherits its generator. Synthetic conversations

state preferences more cleanly than real behavior does. When the generator family also judges, its artifacts sit on both sides of the measurement. PersonaMem-v2 (Jiang et al., 2025b) is the limit case: GPT-5 produces the personas, histories, questions, answers, filters, and rewards, and the intermediate memory is never scored. The category-controlled generators of Appendix A.2 show that questions can be drawn from a declared specification space, but only over synthetic data or document corpora. This framework combines the two. Real histories cannot be released publicly (Appendix H), but they provide factual evidence, natural sparsity, and true no-evidence cases. Specified generation adds machinery, but it makes coverage measurable.

A.2 Specified question generation

The framework's method belongs to a line of work that controls what generated questions test rather than generating freely.

Category-sampled generation. DataMorgana (Filice et al., 2025) is the closest prior system. Users declare question and end-user categories with target probabilities, and an LLM generates RAG questions from corpus documents under sampled combinations. Fensore et al. (2026) study how category granularity affects the resulting benchmarks. We differ in three ways. Evidence binding: DataMorgana grounds each question in a sampled corpus document, and its end-user categories are personas that steer phrasing without adding facts. We bind each specification to observed values in one account's history. Our negative class comes from grounding failures instead of adversarial writing (Rajpurkar et al., 2018). Verification: DataMorgana reports n-gram, compression-ratio, part-of-speech template, and embedding homogenization statistics of its questions, with the same sentence encoder we use. Its filtering step checks category adherence, but the paper reports no adherence rate and no realized category distribution. Our flywheel classifies every exported question against its requested axes. The free-draw drift of Table 15 shows compliance failing at the distribution level even when individual questions are fluent. Feedback: mismatches and free-draw discoveries become taxonomy suggestions, several of which shipped (Appendix F). Classifying generated content back against a sampled taxonomy also appears in text-to-video evaluation (Chatterjee et al., 2025), but without the loop into the taxonomy. PACEShop (Lyu et al., 2026) also samples structured shopping contexts, but generates responses and injects response defects. Our sampled artifact is a question specification whose grounded answerability and realized axes are subsequently checked.

Verification against structured data. Text-to-SQL benchmarks (Zhong et al., 2017; Yu et al., 2018) established questions whose answers are computable from structured records. Our deterministic numeric audit applies that executable-answer property as a verification step for LLM-drafted questions, and it rejected 428 questions that passed every model-based validator. Judge-based validation of generated QA pairs (Saad-Falcon et al., 2024; Yehudai et al., 2024) is the lineage our dual blind judging descends from. The audit exists because model-based judges share failure modes that a deterministic check does not.

A.3 Deterministic measurement instruments

The instruments in this section involve no language-model judge. They audit the questions beside the flywheel, whose match rates come from an LLM classifier (§3). This section states where our use follows precedent and where it departs.

Embedding similarity. Mean pairwise cosine similarity of sentence embeddings is a common homogenization measure, a score of how alike a set of LLM outputs is (Padmakumar & He, 2024; Kirk et al., 2024; Shaib et al., 2025; Filice et al., 2025). DataMorgana applies it to generated questions with the same all-MiniLM-L6-v2 encoder that we use (Filice et al., 2025; Wang et al., 2020; Reimers & Gurevych, 2019). We add two comparisons. Within a run, we compare drafts by constraint degree and by which axis is left free, holding the number of constrained axes fixed (Appendix F.2). Across runs, we compare authoring families on byte-identical sampled plans, so a difference is attributable to the authoring model (Appendix G.5).

Effective number of questions. The Vendi score is the exponential of the Shannon entropy of the eigenvalues of a similarity matrix (Friedman & Dieng, 2023). It descends from the similarity-sensitive diversity indices of ecology (Leinster & Cobbold, 2012). Later work generalizes it to a family of scores (Pasarkar & Dieng, 2024). Prior text applications score model outputs or retrieved passages.

To our knowledge, it has not been applied to the questions of a generated benchmark. We report it next to mean similarity and 3-gram entropy. At full specification 3-gram entropy moves against the other two (Appendix F.2). Tevet & Berant (2021) show that automatic diversity metrics estimate content diversity poorly. No single metric is therefore enough.

Lexical statistics. Unique n-gram counts go back to distinct-n in dialogue generation (Li et al., 2016), and n-gram entropy to Zhang et al. (2018). Shaib et al. (2025) recommend a low-correlation set of four scores, including compression ratios and long n-gram self-repetition. We report two lexical statistics. 3-gram entropy by constraint degree follows Zhang et al. (2018) (Appendix F.2). Unique 3-gram counts are a distinct-n style measure. Distinct-n divides the count by the number of tokens, while we compare raw counts over equal-size subsamples of drafts. The counts compare only within a cohort (Appendix G.4). We do not report a compression ratio.

Parse-derived labels. Syntactic parses have long served deterministic complexity indices (Lu, 2010). Recent work measures the syntactic diversity of LLM output with part-of-speech templates (Shaib et al., 2024) and dependency-tree kernels (Guo et al., 2025). Benchmark papers also parse their own questions. SQuAD categorized answer types with constituency parses and quantified syntactic divergence with dependency paths (Rajpurkar et al., 2016). DROP partitioned answers with a tagger and an entity recognizer (Dua et al., 2019). Sugawara et al. (2018) applied heuristics across twelve datasets. Our labels follow these audits. The departure is the test that each label is associated with the axis that was sampled (Cramér's V, Table 10). We do not report part-of-speech template repetition. Our parse-derived labels measure adherence to the sampled axis. They do not measure repetition of sentence structure.

Adherence to a requested control. Controllable question generation checks adherence with a classifier, with human judges, or with a downstream model. Examples are question type (Cao & Wang, 2021), checked by classifier and judges, and difficulty (Gao et al., 2019), checked by reading-comprehension systems. Attribute accuracy from a classifier is also the usual controllability metric in controllable generation (Zhang et al., 2023). To our knowledge, no prior work measures adherence with a parser and an association statistic. We chose a parser because it shares no failure mode with the drafter or the classifier.

Authoring-family signatures. Surface features identify the model that wrote a text (Sun et al., 2025; Herbold et al., 2023; Shaib et al., 2024). We call such a set of features a signature. Table 3 shows one for three authoring families. GPT-5.6 is the only family that appears under both data sources, and its signature is stable across them. Claude Opus 5 appears under one data source, and its signature is stable across the three cities. In the controlled family comparison, the spread of mean similarity across families is widest on free draws and shrinks under constraints (§4.2, Table 25).

A.4 Local-life-services benchmarks

Three related but distinct efforts evaluate LLMs on commerce data. **LocalSearchBench** (He et al., 2026) builds 900 multi-hop QA tasks from real user queries over 1.3M+ merchant entries across nine cities and six service categories, scoring correctness, completeness, and faithfulness. The best evaluated model reaches only 35.6% correctness with $\sim$30.7% average faithfulness, confirming that grounded multi-hop commerce search remains hard. It searches a shared merchant index for many users. We ask one account about its own history. We note a reproducibility caution from examining its released artifacts. The evaluation code depends on a merchant index and a retrieval component that the repository does not include. The file paths are internal, the index-construction code is absent, and the configuration holds placeholders. The published 900-task dataset is usable, but the full retrieval pipeline is not independently reproducible. We note this because it shaped our own limitation analysis (Appendix H), not to single out one benchmark. Grounding in private data trades public reproducibility for validity, and a paper should state that trade explicitly. **LocalGPT** (Lan et al., 2025) is a separate line of work that benchmarks and adapts LLMs for local-life services. Its benchmark has 41 task types in four categories. Its LocalInstruction method synthesizes instruction data with five cooperating agents, and a fine-tuned 7B model then approaches a 72B model. It demonstrates the value of domain specialization on the same kind of data our histories come from. **PACEShop** (Lyu et al., 2026) pairs structured synthetic personas with public product evidence and four-field shopping responses. It then injects deterministic defects with known families and locations to evaluate structured judge diagnosis. Our framework instead derives questions and answerability from observed

account histories and uses them to diagnose retained memory representations. Its negative class is absence of supporting history rather than an injected response defect. Commerce agent benchmarks score task completion instead of memory. WebShop (Yao et al., 2022) navigates 1.18M real products from text instructions, and τ-bench (Yao et al., 2025) compares the final order database against a goal state in a retail domain. MMShopBench (Hao et al., 2026) builds multimodal multi-turn shopping tasks from real annotated logs with an evidence-grounded protocol. None of them asks questions about a retained memory.

A.5 Position in the agentic systems stack

Systems-level work evaluates whole agent configurations or trajectories. AgentArch (Bogavelli et al., 2025) reads memory design off end-task success for 18 agent configurations on two enterprise tasks and six models. Janapa Reddi (2026) makes the managed trajectory the stack's unit of work and asks for metrics of useful work per verified action. When the unit of measurement is a whole trajectory, every component failure surfaces as the same lost task. This is the only-the-response-is-scored problem of Table 5, restated at the systems level. We take the complementary position and test the trajectory's components separately. These are the memory store, the agent's interaction with it, and, outside this paper's scope, the answering agent. We report the verifier coverage of our own harness (Appendices E and F).

A.6 The wider agentic-memory literature

The literature organizes agentic memory along three kinds of storage (context tokens, intermediate representations, and model parameters) and a lifecycle of construction, updating, and querying. It describes a progression from raw storage through reflection to reusable experience. Architecturally, production systems split into passive background extraction and active self-editing. MemGPT/Letta (Packer et al., 2023) exemplify the latter with explicit core/recall/archival tiers managed by agent-issued tool calls, while Mem0 (Chhikara et al., 2025) exemplifies fact-centric passive extraction. Earlier long-term dialogue work set the pattern: MSC (Xu et al., 2022) collected human multi-session chats, and MemoryBank (Zhong et al., 2024) added a forgetting-curve store evaluated on real and simulated conversations. Cognitive framings such as CoALA (Sumers et al., 2024) adapt the episodic/semantic/procedural distinctions (Tulving, 1972) to language agents, and our memory-function axis follows that lineage. Prospective memory (Einstein & McDaniel, 1990) and metamemory (Nelson & Narens, 1990) motivate question families that commerce order data cannot yet ground. We disable these question families instead of simulating them.

B Framework overview

This appendix walks the construction-and-evaluation flow of Figure 3 once, end to end. The appendices that follow detail its components. Table 6 maps the components of the figure to the section that details it. Twelve manual categories and 218 handcrafted questions are the manual origin (§1). Their failure modes motivated the semantic axes with free sampling, and manual, agent-assisted entry filling populates the axis registries (§3). The axes declare roughly 434,000 potential question specifications, multiplied by the groundings each account's history offers (§4.1). Appendix C details the grammar behind them, answerability, and how semantics become language. The task generation pipeline (Figure 1, main text) reduces sampled specifications to validated questions and answers, 2,218 exported on the 100-account run. It uses different models for different needs and a deterministic check wherever one can be computed. Table 8 maps its ten modules, and Appendices C–D detail its question and answer stages. Axis classification compares requested against realized semantics and powers the flywheel back into the taxonomy and its entries (§3). The Stanza-based syntactic audit checks adherence and variety with a deterministic instrument (Appendix C.5). Full semantic-axis coverage indicates variety that free sampling alone cannot reach (§4.2). The exported corpus feeds the memory-approach evaluation (Appendix E), whose class-separated results Appendix F.7 analyzes.

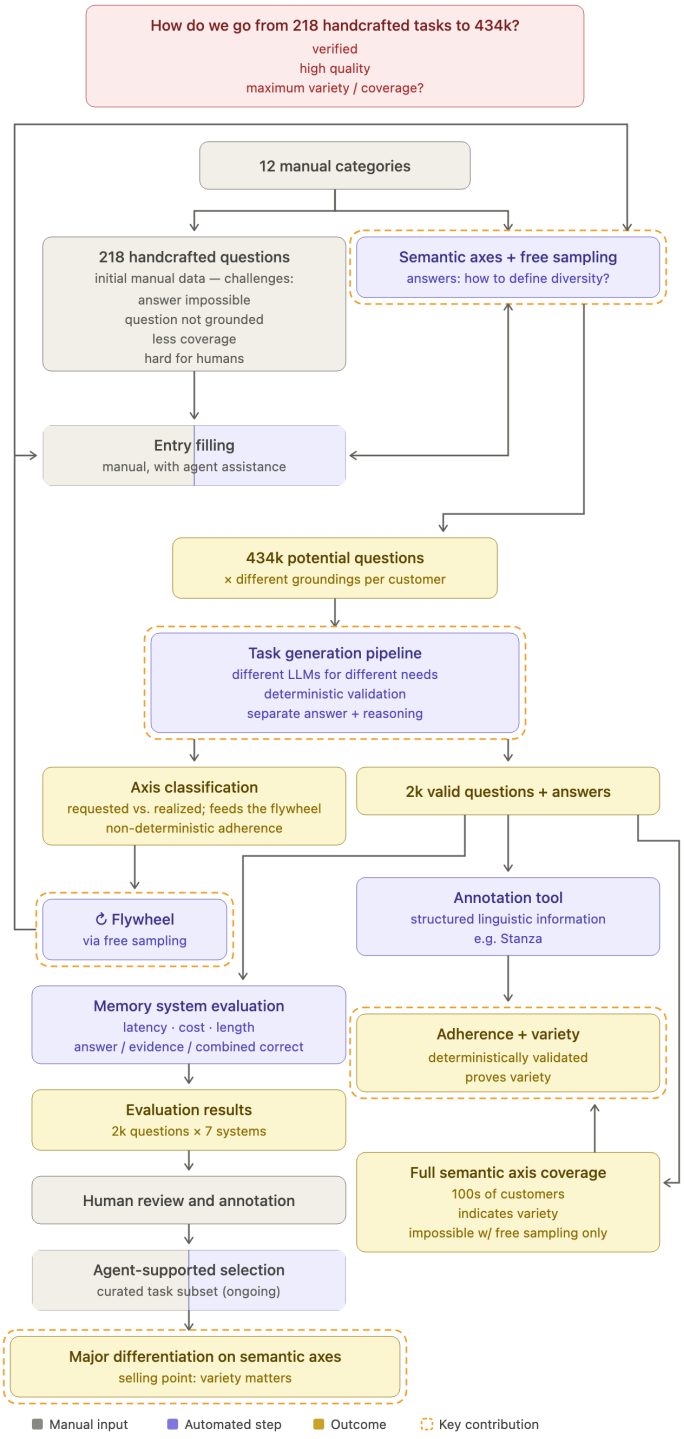

Figure 3: Benchmark-construction and evaluation flow. Handcrafted seed questions and manual categories define semantic axes and entry filling. The task generation pipeline converts their combinations into validated questions and answers, deterministic adherence and variety checks, and the seven-approach memory evaluation. Dashed halos mark the key contributions. The paper's quantitative claims use the full class-separated archive (Appendix F.7). Table 6 maps each component to the section that details it.

Table 6: The components of Figure 3: their kind in the figure's legend and where each is described.

Component	Kind	Detailed in
12 manual categories	manual input	§1
218 handcrafted questions	manual input	§1
Semantic axes + free sampling	key contribution	§3, C.1
Entry filling	manual, agent-assisted	§3, F.2
434k potential questions	outcome	§4.1, C.1
Task generation pipeline	key contribution	Table 8
Axis classification	automated step	§3, F.2
Flywheel	key contribution	§3, F.2
2k valid questions + answers	outcome	§4.2, F.1
Annotation tool (Stanza)	automated step	C.5
Adherence + variety	key contribution	C.5, §4.2
Full semantic-axis coverage	outcome	§4.2, F.3
Memory system evaluation	automated step	E
Evaluation results	outcome	F.7
Major differentiation on axes	key contribution	F.7

Table 7: Generation funnel of the 100-account run. Retention is relative to sampled slots. Loss reasons are typed. The three largest per stage are shown. A question can carry more than one typed defect, so reason counts can exceed the questions lost.

Stage	Count	Retention	Dominant loss reasons
Sampled specification slots	4,000	—	—
Drafted questions	3,905	0.976	drafting omission (95)
Valid questions	3,873	0.968	unobserved binding (32)
Retained after diversification	3,630	0.908	recommended duplicates (136), pattern cap (99)
Answered	3,463	0.866	answer omission (167)
Valid answers	3,428	0.857	unsupported fact (37)
Numeric-audit valid	3,000	0.750	winner mismatch (296), count mismatch (132)
Support judged	2,960	0.740	judgment omission (40)
Valid judgments	2,879	0.720	history alias in prose (71)
Exported	**2,218**	**0.555**	dual-judge disagreement (661)
Evaluation side (Appendix F)			
Questions in evaluation	2,218	0.555	1,911 answerable + 307 unanswerable
Judged cells (7 approaches)	15,526	—	13,377 answerable + 2,149 unanswerable cells

C Question specification and linguistic grounding

This appendix details the specification side of the construction flow (Figure 3, Appendix B). It covers the semantic axes and grammar, when a question is answerable, and how semantics become language.

C.1 Semantic axes and question grammar

Every atomic question is an instance of a shared semantic grammar with seven parts. The *subject* is fixed to one account. The *predicate* is the claim relation. The *target* is the entity or attribute. The *conditioning domain* says which observations qualify. *Time semantics* give a window, a relation, or dynamics. The *computation* is retrieval, count, ranking, comparison, or inference. The *requested result* is one entity, a ranked list, a number, a boolean, or a qualified assessment. The memory-function axis acts as a type parameter over this grammar. An episodic draw restricts predicates to event-identifying relations, while a semantic (profile) draw admits aggregate and dispositional claims. The distinctions mirror query-algebra operations (filtering, grouping, aggregation, ordering (World Wide Web Consortium, 2013b)) and interval-based temporal reasoning (Allen, 1983). The specific registries (42 predicates, 82 targets, 9 ordering conditions) are product-specific design choices. They are versioned and carry provenance metadata in the spirit of PROV (World Wide Web Consortium, 2013a) and datasheets (Gebru et al., 2021).

Two memory functions from the fuller design space, persistent instructions and prospective intentions, are disabled because retained order-history fields cannot ground them. This follows a framework rule: a question family whose evidence cannot exist in the data must be absent from the benchmark, not simulated.

Table 8: The ten pipeline modules of Figure 1: what each does, whether it is a deterministic check or a model call, and where it is detailed.

Module	Role	Check	Detailed in
`sample_specifiers`	draws one specification per slot: axis values under per-axis liberties, compatibility allowlists, and coverage-pulled weighting toward not-yet-sampled categories	deterministic	§3, C.1
`generate_questions`	drafts one natural-language question per specification, filling open slots only with values observed in the account's history, and reports every binding it used	GPT-5.6	§3, C.4
`validate_questions`	confirms the draft is well-formed and that every concrete value it uses occurs in the account's history; unobserved bindings become typed losses	deterministic	§3
`diversify_questions`	classifies the axes each question actually realizes (feeding the flywheel), drops classifier-recommended duplicates, caps repeated surface patterns run-wide, and bans closed yes/no phrasings	Claude Sonnet 4.5	§3, C.4
`generate_answers`	derives the gold answer from the complete history and splits it into a short conclusion plus separate reasoning, with an expected evidence level	GPT-5.6	D.1
`validate_answers`	checks every stated fact and citation in the gold answer against exact history rows; unsupported facts become typed losses	deterministic	D.2
`audit_numeric_answers`	recomputes every count, superlative, and comparison claim from the complete history without a model and rejects mismatches	deterministic	D.3
`judge_support`	two blind judges from different model families grade whether the history supports the gold answer at the claimed evidence level	Sonnet 5 + GPT-5.6 econ.	§3
`validate_judgments`	enforces judge-agreement gates, types disagreements, and rejects judgments that leak history aliases into prose	deterministic	§3
`export`	freezes one answer contract per question (question, gold conclusion, support, evidence level, pass criteria) and builds compact and evaluation-task files	deterministic	F.1, E

Compatibility is evaluated in a fixed rule order: the predicate constrains the target, and the target constrains the ordering condition. A rejected draw is therefore attributable to exactly one rule. A walked example: a specification samples memory function *semantic*, demand *aggregate*, predicate `reorders`, target `dish`, and ordering condition *meal period*. The predicate rule admits `dish` for `reorders`, because an account can reorder a dish. The target rule admits meal-period conditioning. Grounding then fills the open slots from the account's history. The result is a question of the form "Which dish does the account reorder most often at breakfast?" Each rule restricts the later draw to an allowlist, so an incompatible pair such as `reorders` with `spice_level` is never drawn.

C.2 Representative exported contracts

Table 9 shows five contracts selected from the frozen 100-account export used throughout this paper, in the format of the public examples of Table 1 (main text, §3). They span five memory demands and all three evidence regimes. To preserve the paper's privacy boundary, account-derived cuisines, dish types, sauces, and conclusions are replaced by type-preserving aliases. The sampled axes, question structure, answerability decision, and evidence level come from the real exported records.

E1 aggregates item-level observations into a dietary pattern, while E2 compares the frequencies of two cuisines within the restaurant condition. E3 retrieves the type of dish in one temporally selected dinner event. Its memory demand was left free by the sampler and independently realized as retrieval. E4 permits only a qualified inference from sparse sauce evidence rather than a broad preference claim. E5 has no positive conclusion: because the retained dinner evidence never records an explicit utensil decision, the contract requires abstention. Together the examples show that the same export format represents aggregate tendencies, comparisons, event recall, calibrated inference, and absence of evidence.

The examples also expose the evaluator boundary. Candidate generation receives the question and one approach's exported content, but never the gold conclusion, expected evidence level, supporting reasoning, or pass criteria. Those fields enter only after candidate generation, when the frozen contract is graded.

Table 9: Representative exported contracts from the primary private run. The sampled specification lists memory function, memory demand, predicate, target, and ordering condition, in that order. "Free → retrieve" marks an unconstrained demand that the classification flywheel mapped to retrieval. Account-derived values and conclusions are replaced by type-preserving aliases.

ID	Sampled specification	De-identified question	Gold contract
E1	`semantic; aggregate; consistently_selects; dietary_basis; free`	Which dietary pattern did my ordered items most consistently fit?	Pattern A. Significant evidence.
E2	`semantic; compare; orders; cuisine; store_type`	When ordering from restaurants, which cuisine do I order more often, Cuisine A or Cuisine B?	Cuisine A. Significant evidence.
E3	`episodic; free → retrieve; orders; dish_type; meal_period`	What dish type did I order most recently at dinner?	Dish type A. Some evidence.
E4	`semantic; infer; tolerates; sauce_type; cuisine`	When ordering Cuisine C food, what sauce type do I tolerate without repeatedly seeking it out?	Sauce A. Some evidence.
E5	`semantic; assess; omits; utensil_selection; meal_period`	How consistently did I explicitly decline utensils when ordering dinner?	Abstain. No evidence.

C.3 Evidence regimes and answerability

Answerability is a relation between the question, the account history, the permitted evidence, and the inference policy, not a property of the question alone. Each evaluation case is classified by the least interpretive derivation that produces its answer. An *observable* answer appears directly in retained evidence. A *computable* answer comes from a deterministic operation. An *inferable* answer rests on a declared inference policy that supports a qualified claim. An *unanswerable* case has no permitted derivation. The same semantic question should recur across contrasting evidence regimes: none, sparse, sufficient-and-consistent, mixed, conflicting, local-only, and stale. A model then cannot infer the expected confidence from wording alone. This is the contrast-set logic (Gardner et al., 2020) applied to evidence rather than to input perturbation. Calibration is evaluated in both directions: unsupported certainty and unnecessary hedging are both errors (Guo et al., 2017; Geifman & El-Yaniv, 2017), and exact evidence-level matching records this distinction. The three contract evidence levels are operationalized by independent-delivery count. *No* evidence means zero relevant delivery events support the conclusion. *Some* means one or two independent deliveries, or materially mixed evidence. *Significant* means at least three independent deliveries with counterevidence not overturning the conclusion. Judges count relevant deliveries in the supplied history, not cited facts, and both the answer generator and the support judges receive this same rubric.

The account-versus-person rule matters because commerce data makes it concrete. Account data may reflect a household rather than a person. A history in which every vegetarian order came from one shared-account member supports "the account frequently orders vegetarian dishes," not "the account holder is vegetarian". This is the ecological-inference caution (Robinson, 1950). Answer contracts prohibit unsupported person-level claims, and the current cohort of 100 accounts keeps this trap populated.

C.4 Linguistic realization

The generator fixes semantics before wording: a valid specification is sampled, evidence is bound, and only then is language rendered. This ordering prevents fluent-but-invalid questions and enables three validation levels borrowed from round-trip design. *Structural* agreement asks whether the realized axes match the requested slots. The flywheel measures it at 99.6% over the exported questions. *Semantic* equivalence asks whether two phrasings of one specification ask the same question, so that a paraphrase cannot silently change what is tested. *Denotational* agreement asks whether two readings of one question compute the same answer from the same history. It is testable with probe histories built so that nearby readings diverge. For example, "highest repeat-order rate" and "appears in the most orders" pick different items when a frequently bought item is rarely reordered. The first level is measured. The latter two are design targets for the pipeline. Today a prompt instruction alone enforces them: hold a single interpretation, and treat wording that supports materially different readings as unknown. Diversification enforces linguistic variety in three ways. A run-wide cap on

repeated surface patterns removed 99 questions in this run, and classifier-recommended duplicate removal dropped 136. Closed yes/no phrasings are banned in favor of open realizations. Free-axis proposals whose names only express expected absence ("unconditioned," "no condition," "all orders") are not treated as taxonomy signals. They mark an intentionally unconstrained axis, not a missing category. This small rule matters: without it, the flywheel's discovery channel fills with paraphrases of "nothing".

C.5 Syntactic audit of the realized corpus

The validation levels above rely on LLM judges. We also label all 2,218 questions of the 100-account generation run with a deterministic syntactic pipeline. This checks linguistic realization with an instrument that shares none of the judges' failure modes. It runs Stanza (Qi et al., 2020) tokenization, part-of-speech tagging, lemmatization, Universal Dependencies parsing (Nivre et al., 2020), and named-entity recognition locally over the exported question text. From each parse we derive twenty interpretable labels: interrogative type, root (main-predicate) lemma, the wh-attached focus noun, core-argument lemmas, markers for negation, comparison, superlatives, and temporal or frequency reference, entity types, and structural-complexity measures. The labels first describe the corpus plainly. Question length, parse depth, subordinate clauses, temporal and comparative marking, and gold-answer length are reported in Table 3 (§4.2), where the private run stands alongside the public runs. No closed yes/no phrasing survives to export. We quantify each label's association with the five sampled semantic axes by Cramér's V (Cramér, 1946) with the bias correction of Bergsma (2013). The statistic is computed over the questions on which each axis was sampled (Table 10). The correction matters because the lemma-valued labels enter the contingency tables at full cardinality, where the uncorrected estimator is inflated by sparsity.

Table 10: Association between parser-derived question labels and the sampled semantic axes (Cramér's V, where 0 means independent), for the eleven labels with the highest mean association over the 100-account generation run, using the bias-corrected estimator of Bergsma (2013). V is computed over questions whose label value is resolved. "Unres. " reports the share of questions excluded as unresolved (e.g., no direct object, or a wh-word that attaches to no noun), so sparse labels are association among the questions that realize the construction.

Derived label	Mean	Unres.	Demand	Function	Ordering	Predicate	Target
Root lemma	0.46	0%	0.51	0.62	0.15	0.69	0.35
Object lemma	0.43	29%	0.24	0.40	0.43	0.49	0.60
Focus noun	0.41	45%	0.24	0.35	0.26	0.47	0.71
Oblique lemma	0.37	25%	0.33	0.51	0.36	0.38	0.25
Temporal marking	0.35	0%	0.65	0.27	0.15	0.41	0.29
Subject lemma	0.31	1%	0.28	0.46	0.10	0.33	0.40
Interrogative type	0.31	0%	0.21	0.30	0.16	0.45	0.45
How-subtype	0.30	82%	0.31	0.34	0.00	0.53	0.32
Frequency marking	0.27	0%	0.16	0.46	0.00	0.45	0.27
Negation	0.27	0%	0.09	0.06	0.04	0.78	0.37
Superlative	0.26	0%	0.29	0.31	0.07	0.39	0.23

The audit yields three findings. First, the sampled semantics control surface form. The requested predicate is realized as the question's main verb ($V = 0.69$). Negated phrasings occur almost only under predicates that assert something is left out, such as `omits`, `removes`, and `avoids` ($V = 0.78$). Time expressions concentrate in questions whose memory demand is `temporal`, the demand category for questions about when or in what order events happened ($V = 0.65$). Named-entity types track the ordering condition ($V = 0.57$): cuisine conditions surface as demonyms and location conditions as geopolitical entities. The target concept appears verbatim in 90.5% of target-bearing questions, predominantly in grammatical focus positions (object, compound, or subject). Per-axis requested-versus-realized status is carried over from the flywheel's LLM-judged classification (§3). It is *matched* on 98.1% to 100% of requested values, with at most two undecided judgments per axis. The parser-based associations above thus confirm that LLM-judged structural agreement with a deterministic, LLM-free measurement. Second, the axes are coupled as sampled. The predicate strongly determines the memory function ($V = 0.67$) and correlates with the target ($V = 0.52$). The episodic memory function is never co-sampled with a memory demand: the function is *semantic* on all 342 questions that carry both axes. The five axes therefore span fewer

effective degrees of freedom than their count suggests. Third, the corpus consists only of open wh-questions (*what* 33%, *which* 28%, *when* 19%, *how* 18%, *where* 2%), with zero polar realizations. This confirms that the yes/no ban of the diversification stage holds. Negation (3%), comparatives (5%), and conditionals (1%) remain rare, marking concrete headroom for exclusion-, comparison-, and hypothetical-framed question families.

D Answer generation and verification

This appendix covers the answer side of the pipeline. It explains how gold answers are derived from the history, how they are validated, and the deterministic audit that rechecks their numeric claims.

D.1 Gold-answer contracts and labels

Answer-contract separation. Mixing conclusions with reasoning in gold answers made verdicts inconsistent. One identical answer text passed under four memory approaches and failed under the fifth with the same contract hash. It also rewarded vague answers: a right conclusion with no numbers passed, while a right conclusion with approximate numbers failed. Splitting gold answers into a short conclusion plus separate reasoning dropped multi-sentence answers from 15.7% to 1.0%. The numeric working moved into the reasoning field (digit rate 12.3% $\rightarrow$ 36.2%) instead of being lost.

Label-blind evidence classification. The drafter's early no-evidence guess is not trusted as a final label. In one affected run, 104 of 190 compact records kept the draft `no_evidence` kind, while 87 of those had a concrete supported answer. The final question kind is now derived after support judging from the answer shape and judge verdicts. A live canary (one account, 372 deliveries, 911 items, 24 normalized fields) validates the classification path end-to-end.

D.2 Answer validation

Every fact and citation in a gold answer is checked deterministically against exact history rows. An answer whose stated support does not match the rows it cites is rejected as an unsupported fact. This is the dominant loss reason of this stage (37 answers in the 100-account run, Table 7).

D.3 Deterministic numeric audit

An LLM-free audit stage recomputes every count, superlative, and comparison claim from the complete history. In this run it rejected 428 questions (296 winner mismatches, 132 count mismatches) that had passed every model-based validator. This is the strongest single argument for pairing judges with deterministic checks. Judge families and biases are a known risk (Zheng et al., 2023; Liu et al., 2023; Koo et al., 2024; Panickssery et al., 2024). PACEShop (Lyu et al., 2026) evaluates pointwise judges through controlled persona, cross-component, grounding, and localization defects and a structured diagnostic output contract. That protocol diagnoses a supplied shopping response. Our blind support judges instead decide whether a generated question and authored answer are supported by the complete account history, and disagreements block export. Exact checks, blind dual judging, and model adjudication of disputed records are the mitigations used here. A stratified human audit of axis and answerability labels remains open.

E Evaluation harness

The evaluation side is a standalone runner, separate from generation. Its contract has four stages. **(1) Task freezing.** The generation pipeline exports one frozen contract per task: question, gold conclusion, gold support, expected evidence level, explicit answer- and support-pass criteria, and an abstention policy. Contracts are frozen before any memory approach is graded, so every approach faces byte-identical tasks. **(2) Candidate generation.** An adapter stage answers each question from one exported content file per approach. Tasks are projected down to task ID, pseudonymous account ID, and question text *before* any prompt is built. Gold-contract fields therefore can never leak into an adapter prompt. The account ID selects the content but is never rendered. Environment-specific exporters produce the content files out of band and fail closed. Source loading rejects any nested

raw identifier field before a model call. Legacy or cross-approach content wrappers are rejected, not relabeled. Empty exported content aborts candidate generation as a source-availability failure instead of being scored as abstention. A shared answer model (GPT-5.6) produces, for every task, a candidate answer, a candidate evidence level, and separate free-text support. **(3) Judging.** A grading judge (Claude Sonnet 4.5) scores answer correctness against the pass criteria *blind* to the candidate's support. It scores the evidence level as exact-match and separately checks support consistency. Typed failure codes record the reason for an incorrect answer or evidence level. **(4) Reporting.** Per-request latency, token, and cost ledgers are persisted alongside verdicts. Manifests label every artifact a component diagnostic rather than a system score (persisted as `system_score_valid: false`). If a memory approach cannot receive a question at all, an automatic failure is recorded rather than a skip.

Reported metrics. For answerable questions the harness reports *answer accuracy*. For unanswerable questions it reports the *correct-abstention rate*. Evidence-level accuracy, full-pass rate, typed failure codes, paired win/loss counts against the raw-history baseline, oracle rate, and universal failures are partitioned by the same two classes. Rates and paired differences carry 95% percentile intervals from 10,000 account-clustered bootstrap resamples. Paired comparisons use an exact account-cluster sign-flip test. Holm adjustment controls familywise error across the six archive-wide comparisons per class and, separately, across all displayed axis-slice comparisons per class. Balanced accuracy (the unweighted mean of answerable-question accuracy and unanswerable-question correct abstention) is the only cross-class summary. Per-request latency, token, and cost telemetry remain class-agnostic operational measures. Support consistency is tracked as a separate non-scoring diagnostic.

Columns of Table 4. "Answer" is accuracy on the answerable questions of that panel, and "Abstain" is correct abstention on its unanswerable questions. "Reviewed" is accuracy on the 247 answerable questions of the hand-reviewed PantryBench subset (§4.1), which exists for that panel only and is too small for a useful interval. Scores are percentages, and brackets are account-clustered 95% bootstrap intervals. "$\geq$4 axes" is answerable accuracy on the questions whose specification constrained four or more semantic axes (616 private, 611 public). "Temporal" and "Aggregate" are answerable accuracy on the temporal and aggregate memory-demand questions (229 and 293 private, 293 and 391 public). "Store" is the median memory text per account in characters, which the private snapshots do not export. "Context" is the median adapter prompt tokens per question, and "Latency" is the median adapter answering time in seconds. Both describe that panel's shared reader, not production serving. "Unique" counts answerable questions that only this configuration answered correctly within its panel. "Wins" counts the displayed axis-value slices on which it has the highest point accuracy (49 private slices in Appendix F.7, 48 public slices in Appendix G.8). Wins use no significance test, and a tie counts for each tied configuration.

Answerability as a two-class evaluation. We call questions with significant or some supporting evidence *answerable* and questions with no supporting evidence *unanswerable*. These correspond to positive and negative classes only in the sense of answerability. The label says whether the retained history contains sufficient evidence, not whether a candidate response is correct. On the answerable (positive) class, a wrong answer and an unwarranted abstention both count against answer accuracy. The harness's typed failure codes keep them distinguishable. On the unanswerable (negative) class, a correct abstention is true-negative behavior and an invented answer is a false positive. Balanced accuracy is appropriate because the negative class is the designed minority (13.8% of exported questions). Pooled accuracy changes with that class ratio and can make answering and abstaining cancel each other. We therefore use class-wise rates as the primary results and balanced accuracy only as a secondary summary.

F Full results on the private cohort

F.1 Generation statistics

This subsection reports three measurements that Table 7 aggregates away.

Numeric-audit outcomes. Of the 3,428 answer-valid questions, the deterministic audit verified 200 numeric claims exactly and rejected 428 (12.5%) for mismatches. It could not ground 845 (claims with no computable counterpart) and skipped 1,955 that carried no numeric claim.

Unanswerable questions through the funnel. Table 11 traces the drafted no-evidence label through the pipeline. That label is a hint, not a verdict (Appendix D.1). Support judging assigns the final evidence level. Almost two thirds of the draft-unanswerable exported questions turned out to have supporting evidence, leaving the benchmark's 13.8% negative class. The drafted fraction varies strongly per account (5%–90%, median 28%), tracking real differences in history density rather than a generation artifact.

Table 11: The drafted no-evidence label through the funnel of the 100-account run. Support judging assigns the final label, so a draft-unanswerable question can export as answerable.

Stage	Questions
Drafted as unanswerable (35.9% of drafts)	1,401
Surviving to answer generation	1,285
Exported still carrying the drafted label	849
of which judged answerable after all	542
of which confirmed unanswerable after judging	307

Yield by specification tightness. Table 12 shows survival from sampled slot to exported question by the number of constrained axes. Moderately constrained specifications (2–4 axes) survive best at 57–59%, while the fully unconstrained control survives worst (32%). This agrees with the drift findings of Table 15 (Appendix F.2).

Table 12: Survival from sampled specification slot to exported question, by the number of constrained axes.

Constrained axes	Sampled	Exported	Survival
0 (unsteered control)	217	70	32%
1	164	73	45%
2	780	445	57%
3	1,511	887	59%
4	1,129	654	58%
5	199	89	45%

Stage retention per axis and category. Figures 4–7 refine Table 12 along the other dimension: retention of sampled slots after each pipeline stage, for the slots that constrained each axis and each category. Attrition is close to uniform across axes through the middle stages. The axes separate mainly at the export gate, where dual-judge disagreement costs demand-constrained slots most (0.53 retained) and function-constrained slots least (0.60). The category view attributes losses to specific gates. The `aggregate` and `compare` demands fall mainly at the deterministic numeric audit, and `beverage_category` conditioning thins already at drafting.

F.2 Classification flywheel detail

During the diversification stage, a classifier reads each drafted question and labels which value of every semantic axis the question actually expresses. The flywheel then compares these realized labels against the values the specification requested. This classification succeeded for every exported question, and zero were unclassifiable. Defective classifier responses were rare and typed: 6 already-declared-missing-field and 10 duplicate-axis-classification across all 3,873 classified drafts. Memory function shows zero confusions: every constrained draw realized the requested function (episodic 830/830 and semantic 552/552). The 16 memory-demand mismatches are sparse and spread over ten adjacent requested-realized pairs (Table 13). A semantic mismatch is not itself a removal criterion. A mismatched question can still be removed for an independent duplicate, non-atomicity, or repeated-pattern reason. For retained questions, requested and realized labels persist as audit metadata and feed the suggestion engine. Ambiguity concentrates in the axes with the largest registries: 2 target, 1 predicate, and 1 ordering-condition case, each with a written rationale. The syntactic audit (Appendix C.5) confirms these classifications with an LLM-free instrument. Free-draw realization pressure by category is the clearest diversity signal: Table 15 contrasts sampled with realized volume for the dominant categories. The suggestion engine emitted 14 compatibility-rule proposals (all memory-demand adjacencies) and taxonomy-addition candidates ranked by

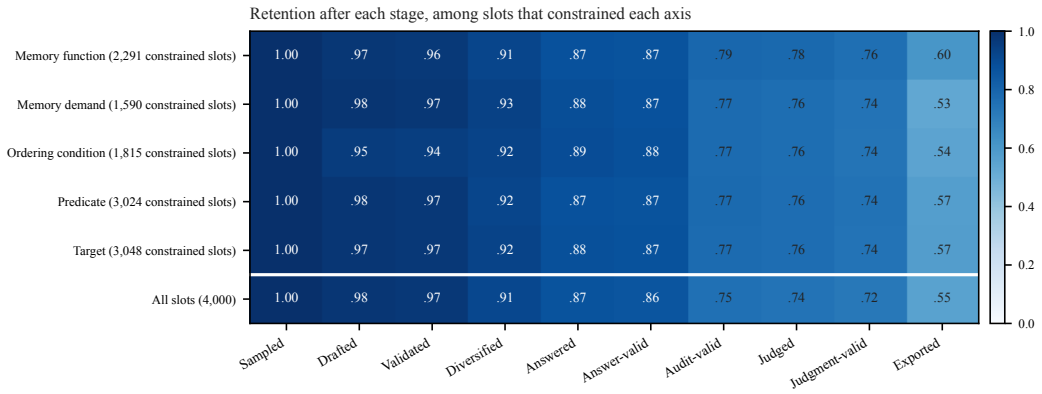

Figure 4: Retention of sampled specification slots after each pipeline stage, among the slots that constrained each axis. Each row label carries its denominator. The all-slots row is the funnel of Table 7. An axis is constrained on only part of the 4,000 slots, because its sampling liberty leaves it free elsewhere. Axis rows therefore need not track the all-slots row.

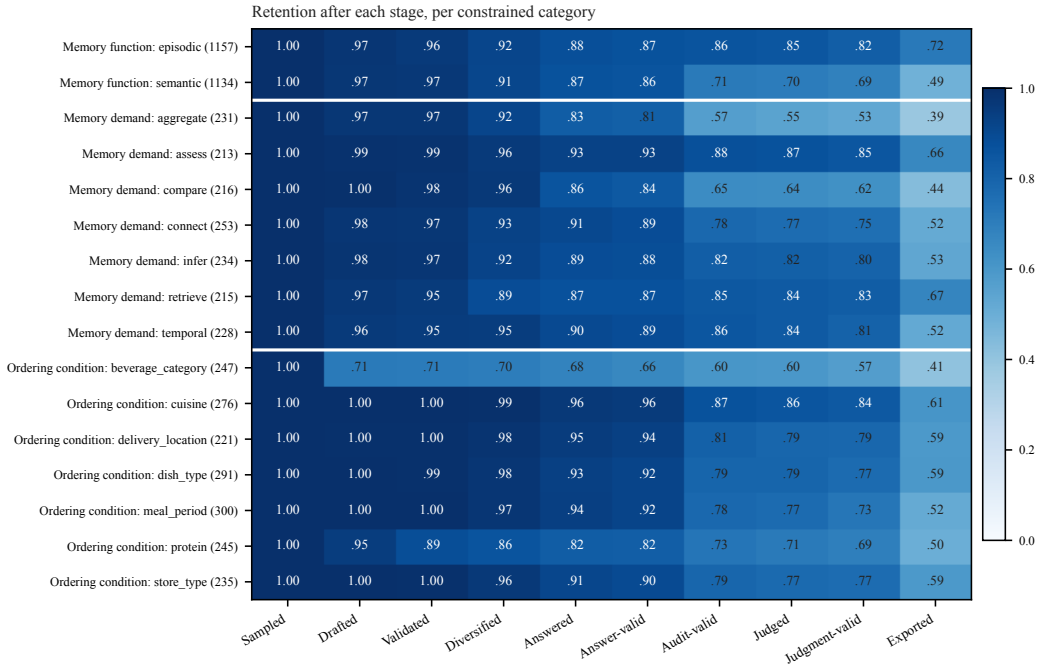

Figure 5: Stage retention per constrained category of the memory function, memory demand, and ordering-condition axes. The `aggregate` and `compare` demands lose most of their slots at the deterministic numeric audit. `beverage_category` conditioning loses 29% at drafting, before any gate runs.

evidence (Table 14). Several of these subsequently became real taxonomy entries (`day_of_week`, `day_type`, `selected_option`), closing the loop the flywheel was designed for. Proposals such as `all_orders` instead confirmed the expected-absence suppression rule.

Figure 8 confirms the same collapse at the surface level of the drafted questions themselves, independent of the reviewer's taxonomy. The free-draw control is the most self-similar population by a wide margin, and moderate constraint (two to three axes) yields the most diverse drafts. The same ordering persists after diversification removes near-duplicates from the exported questions (Table 27). Free

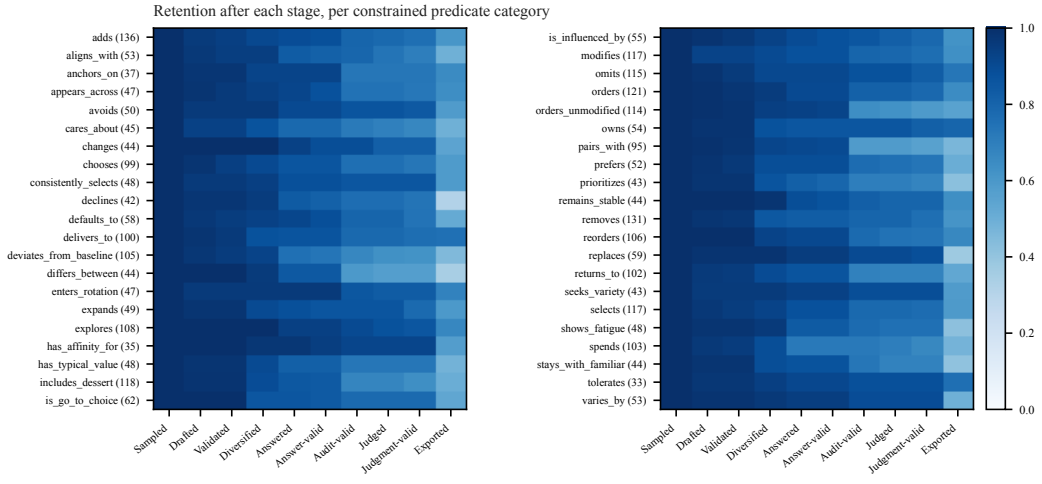

Figure 6: Stage retention per constrained predicate category, alphabetical in two column halves. Each label carries its sampled-slot count.

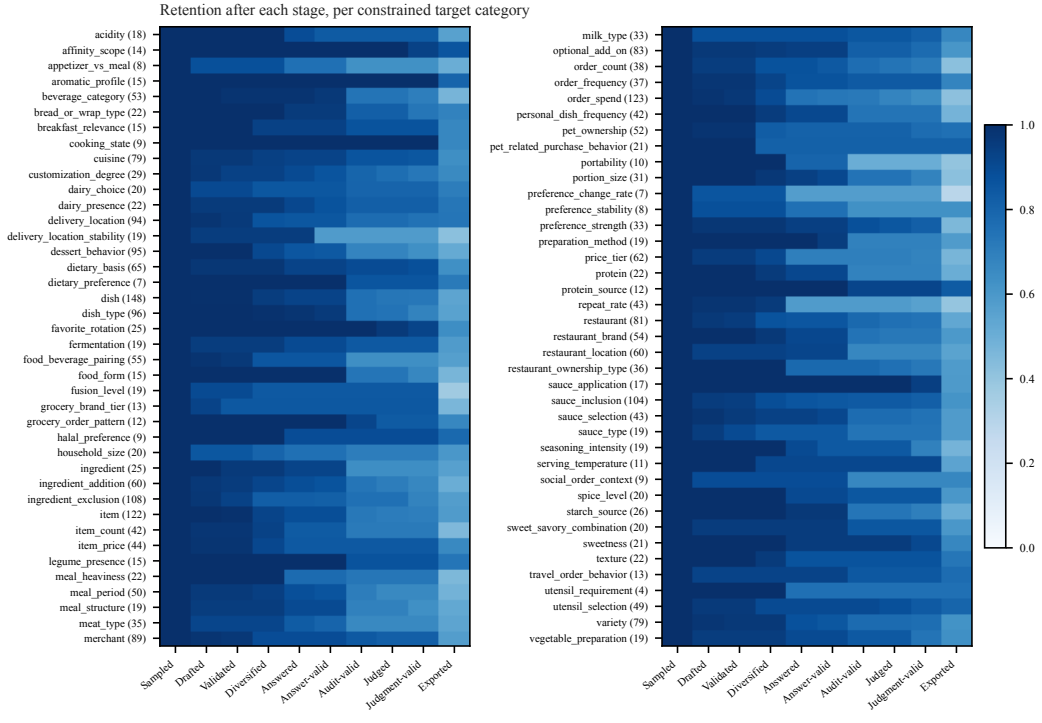

Figure 7: Stage retention per constrained target category, alphabetical in two column halves. Each label carries its sampled-slot count.

draws remain the most self-similar population. Two constrained axes give the lowest mean similarity, which rises again at full specification.

Figure 9 decomposes the collapse by axis. The axes contribute very unequally, in a way their surface realization does not predict. The syntactic audit (Appendix C.5) finds the predicate and the memory demand almost equally strongly realized in the text (main-verb association $V{=}0.69$ versus temporal-expression association $V{=}0.65$). Yet their marginal effect on corpus diversity differs sharply. Leaving memory demand free is the costliest freedom of any axis. The cost grows with specification depth:

Table 13: The 16 memory-demand mismatches of the 100-account run: requested category versus the category the classifier judged realized.

Requested	Realized	Count
connect	assess	3
connect	compare	2
connect	aggregate	1
connect	temporal	1
aggregate	compare	2
aggregate	temporal	1
retrieve	aggregate	2
assess	temporal	2
compare	assess	1
infer	aggregate	1

Table 14: Taxonomy-addition candidates emitted by the suggestion engine on the 100-account run, ranked by the number of supporting observations.

Candidate	Evidence
specific_item	30
has_characteristic	22
all_orders	15
specific_date	15
has_attribute	14
exhibits	11
day_of_week	9

$+0.021$, $+0.039$, and $+0.064$ mean pairwise similarity at two, three, and four constrained axes. The predicate's penalty stays small and roughly flat ($+0.018$ to $+0.028$). Given the subject but not the computation, drafts converge on the same few demand patterns. Given everything except the predicate, they still vary. Part of the memory-demand gap is compositional. Demand-free cells are predominantly episodic, and episodic drafts are more self-similar (0.466 versus 0.407 for semantic drafts within the degree-4 demand-free group). The gap survives this split, and the compositional pathway is itself a consequence of leaving the axis free. The ordering condition runs the other way at every degree: constraining it *raises* mean pairwise similarity by ~ 0.06. Its values are surface-visible qualifiers such as a cuisine or a time window, so specifying one templates the phrasing instead of diversifying the intent. For a fixed number of constrained axes, constraining the memory demand helps diversity most, constraining the predicate changes little, and constraining the ordering condition reduces diversity.

Figure 10 shows the full distributions behind those means at the depth where every single-freedom cell is populated, bracketed by the two extremes. It adds two observations a mean cannot carry. First, the free-draw collapse is an effect of *joint* freedom, not the sum of per-axis freedoms. Even the least diverse single-freedom group (memory demand free, 0.429) is a unimodal distribution with no mass near 1.0. The all-free control (0.562) carries the heavy high-similarity shoulder and the near-duplicate mode. The costliest single freedom adds $+0.03$ over full specification. All five together add $+0.17$ and a qualitatively different tail. Second, full specification is not the diversity optimum. The fully specified population (0.397) sits visibly to the right of the ordering-condition-free group (0.336), because pinning the last surface-visible qualifier templates the phrasing. The mild rise from three to five constrained axes in Figure 8 is the same effect. Diversity therefore does not increase monotonically with the number of constraints. It depends on which axes stay free, and one well-chosen free axis gives more diversity than none.

Entropy states the collapse in a unit that is easier to picture: a count of questions rather than an average similarity. The Vendi score (Friedman & Dieng, 2023) is such an entropy measure. For a sample of drafts it answers "after merging near-duplicates, how many different questions are left?" Forty near-identical drafts count as roughly one, and 40 unrelated drafts count as roughly forty. By this count, 40 free draws contain only 5.7 different questions, while 40 constrained drafts contain 12.5–15.2 at every constraint degree. The exported corpus keeps the gap: 7.6 against 12.4–15.5

Table 15: Free-draw drift on the 100-account run: categories that absorb most of the volume when axes are left unconstrained. Constrained draws hold near parity (§4.2). Free draws collapse toward retrieval, aggregation, and semantic recall.

Axis	Category	Sampled	Realized	From free draws
memory demand	retrieve	215	1,197	995
memory demand	aggregate	231	905	678
memory demand	temporal	228	452	218
memory function	semantic	1,134	2,681	1,586
memory function	episodic	1,157	1,192	82
predicate	orders	121	418	301
target	item	122	340	217

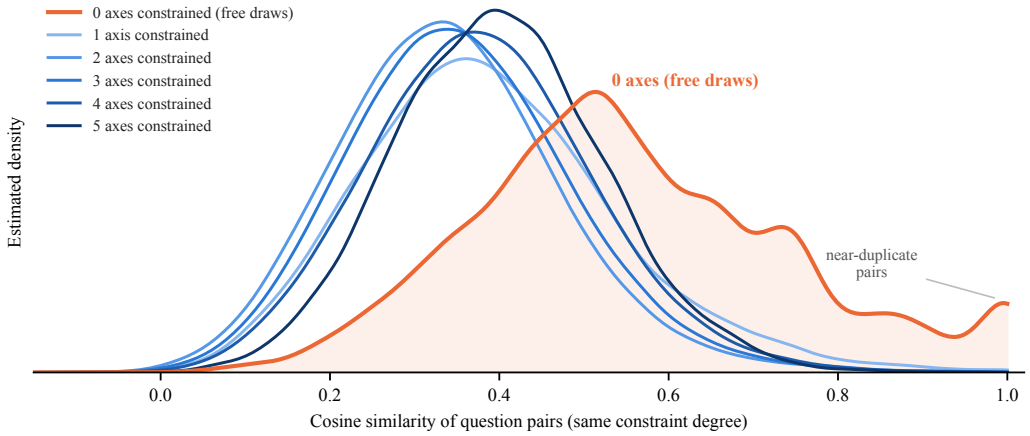

Figure 8: Unconstrained sampling collapses in embedding space on the private cohort. Kernel density of pairwise cosine similarity (all-MiniLM-L6-v2, Wang et al., 2020; Reimers & Gurevych, 2019) between drafted questions with the same number of constrained axes, over all 3,905 drafts of the 100-account run. Similarity is measured on drafts, before diversification removes near-duplicates. Table 2 reports the exported corpus. Free draws (0 axes) are the least diverse population. Their distribution sits far right of every constrained regime, with a heavy high-similarity tail and a mode at $\sim$1.0 from near-duplicate drafts. Constraining even a single axis removes most of the collapse. Similarity rises mildly again at full specification as cells become templated (per-degree means in Table 22, category-level drift in Table 15).

(Table 27). Export-stage APS and Vendi use equal-size subsamples of 40 and 40 exported questions, with 25 resamples each. Unconstrained drafting spends most of its budget re-asking the same few questions, and constraining even one axis more than doubles the yield. The gap is much starker in this unit than in the similarity averages (0.56 versus 0.34–0.40). The count is not just the average in disguise (Spearman -0.92 across the sixteen groups measured). The count also explains the mild right rise in Figure 8. As more axes are constrained, drafts use a wider vocabulary (3-gram entropy keeps rising). Yet they ask fewer different questions (15.2 down to 13.0 after two constrained axes). Fully specified drafts *sound* more varied but *ask* fewer distinct things.

F.3 Cohort-size effect on coverage

A 10-account diagnostic run exported 225 of 400 slots (56.3%), of which 195 became evaluator-ready tasks after the evaluation-side gates. It left eight target categories with zero explicit and zero realized coverage: `bread_or_wrap_type`, `cooking_state`, `dairy_presence`, `delivery_location_stability`, `legume_presence`, `portability`, `social_order_context`, and `starch_source`. One category (`spice_level`) was reached only through free questions. The 100-account run covers ev-

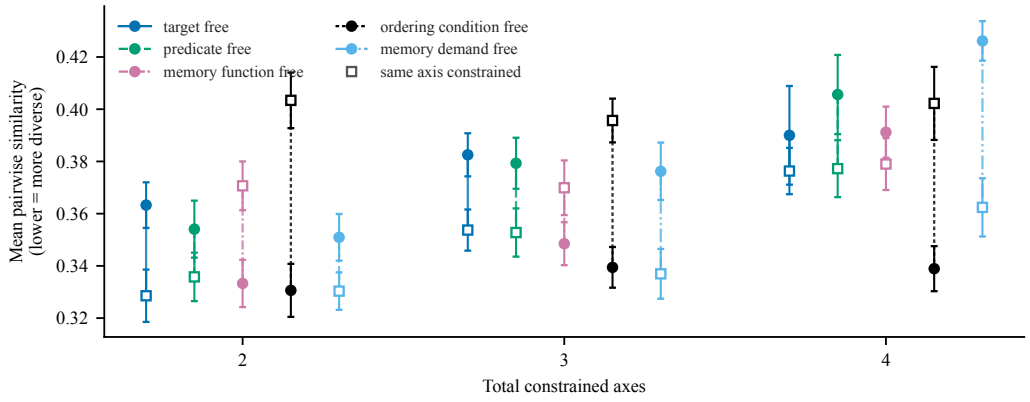

Figure 9: Marginal diversity effect of each semantic axis at fixed specification depth: mean pairwise cosine similarity (all-MiniLM-L6-v2, equal-size bootstrap resamples, ± 1 sd, lower = more diverse) among degree-k drafts that leave the axis free (filled dot) versus those that constrain it (open square), for $k \in \{2, 3, 4\}$. Holding the number of constrained axes fixed separates *which* axis is free from *how many* axes are free. This is necessary because axis freedoms travel together (an episodic memory function, for example, always leaves memory demand free). Degree 1 is omitted: two axes lack a $\geq$30-question constrained cell there.

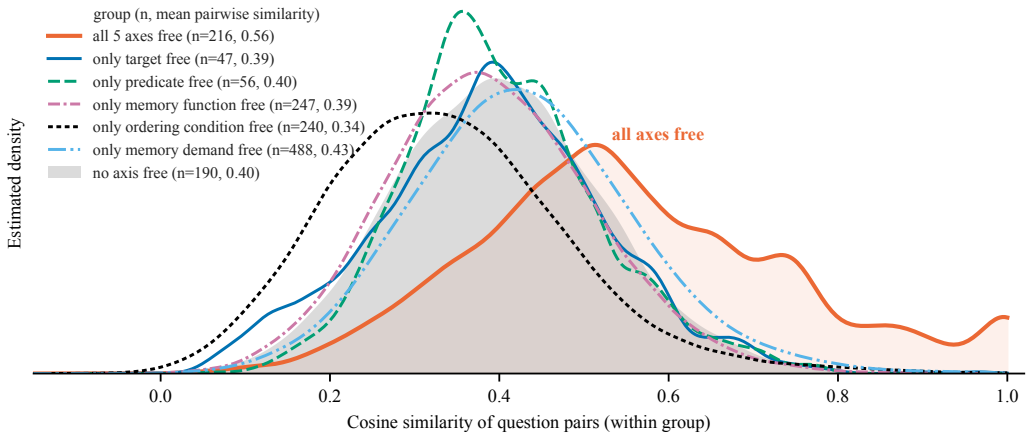

Figure 10: Distributional view of Figure 9 at four constrained axes: kernel densities of within-group pairwise cosine similarity for drafts with exactly one axis free, one curve per free axis, bracketed by the two extreme populations: all five axes free (the free-draw control, orange, filled) and no axis free (fully specified, grey shaded area). The legend reports group size and mean pairwise similarity (equal-size bootstrap resamples, lower = more diverse).

ery enabled category with zero empty cells. A coverage claim therefore needs its denominator. The same generator, configuration, and taxonomy produce different coverage at different cohort sizes, and only a framework that enumerates its cells can report this (Kuhn et al., 2012; Ribeiro et al., 2020).

F.4 Combination coverage

Per-axis coverage says nothing about combinations, so we report those separately, against what the compatibility rules make sampleable rather than against the nominal product space. Section 4.1 reports roughly 434,000 possible combinations for the current registry, which has 82 targets and 9 ordering conditions. The evaluated run used an earlier registry with 79 targets and 7 ordering conditions.

This registry defines 325,164 possible combinations. The flywheel identified gaps in the evaluated registry. Based on its results, we added `day_of_week`, `day_type`, and `selected_option` as targets. We also added `day_type` and `item` as ordering conditions. The 124 compatibility rules reduce the maximally specified sampleable space to 20,008 combinations. A rule can leave an axis free, as the episodic memory function does for memory demand, so such a specification is maximal with four constrained axes and counts once. The 2,218 exported questions realize 706 of them (3.5%). Full-combination coverage is not claimed and is not reachable at this corpus size. The corpus size is adjustable. The run samples 40 specifications per account. Pair and combination coverage rise with a higher per-account sampling rate or a larger cohort, at proportional generation cost. Pairwise combinations are a different matter. The exported corpus realizes 80.5% of the sampleable axis-value pairs, and target × ordering condition (the two largest registries) is the weakest pair (Table 16). Realized labels come from the flywheel's classification, so free draws extend reach beyond the sampleable space. Exported questions additionally realize 85 maximal combinations that the rules never request jointly, mostly a target the rules do not allow for the realized predicate. Every number in this paragraph is recomputed from the run's archived configuration and export by `analysis/axis_pair_coverage.py`, versioned with the paper source.

Table 16: Pairwise combination coverage of the exported corpus: for each axis pair, the number of value pairs the compatibility rules allow to be sampled jointly, and how many of them at least one exported question realizes.

Axis pair	Sampleable pairs	Realized	Coverage
memory function × memory demand	7	7	100%
memory function × predicate	58	58	100%
memory function × target	158	153	96.8%
memory function × condition	14	14	100%
memory demand × predicate	263	256	97.3%
memory demand × target	553	406	73.4%
memory demand × condition	49	49	100%
predicate × target	522	403	77.2%
predicate × condition	294	252	85.7%
target × condition	424	288	67.9%
total	2,342	1,886	80.5%

F.5 Memory-approach evaluation

Seven memory approaches were evaluated: raw orders, System A + orders, System B + orders, System C + orders, and standalone Systems A, B, and C. Every reported number is a component diagnostic over *exported snapshots* of the memory systems. Production retrieval, ranking, tools, and the production answer agent are not exercised, and the artifacts carry an explicit no-system-score contract. The approaches' exported contents also differ greatly in size: standalone Systems B and C export compact profile representations, while orders-backed approaches carry the complete order history. That size difference is part of what is measured (whether the representation retains enough information), not a confound to remove. Token-budget effects on the shared answer model are not separately controlled, however.

Evaluation archive. The evaluator analysis uses all 2,218 exported questions, i.e., 15,526 candidate-answer cells across the seven approaches. Excluding the 295 questions whose evidence label lacked dual-judge consensus raises each approach's answerable-question accuracy by at most 0.025 and changes no ordering. All 295 are answerable, so the 307-question negative class carries full consensus. The archive's manifest pins input hashes and retains an explicit no-end-to-end-score contract: the cells are component diagnostics of exported content, not production-system scores.

F.6 Memory representations on the private cohort: retention versus calibration

Table 4 shows that all memory systems with raw order history are better than the plain ones. Paired account-level tests analyze significance (Appendix F.7). Separation by answerable and abstention

question reveals a trade-off that one aggregate score would hide. The ranking of systems for answerable questions is the inverse from the unanswerable question. Systems B and C abstain correctly most often, because they compress more. The configuration with the highest answerable accuracy, System A + orders history, has the lowest correct-abstention rate. Paired account-level tests confirm the asymmetry. The answerable advantage of System A + orders history over raw history is not distinguishable from zero (p_{Holm}=0.55). Its abstention deficit is significant (p_{Holm}=0.0043, Appendix F.7). Figures 14 and 15 (Appendix F.7) break down answerable accuracy by memory demand, target, and predicate. In Figure 11, we can see a wide range of how many systems agree on a single request, which can be seen as an indicator for complexity. The added columns show how much the approaches differ beyond the aggregate score. System A + orders history reads three times the raw context (138.2k versus 46.6k median prompt tokens) for an answerable gain that is not distinguishable from zero. Systems B and C compress the history to at most 3.4k tokens and answer in 3.8 seconds or less, against 8.6 seconds or more for every orders-backed approach. They keep only 0.09–0.11 of accuracy on time-ordered questions against 0.23–0.28 on aggregate questions, so compression removes event sequence before it removes summaries. Accuracy also falls with specification complexity for every approach, from 0.83 to 0.63 for the raw history. Every approach, including the two weakest, answers some questions that no other approach answers correctly and no single approach leads on every question type.

F.7 How the memory approaches differ: archive-wide differentiation analysis

This subsection asks how the approaches differ and what the benchmark's own structure contributes to telling them apart. We analyze the full archive. After the location/price-corrected candidates were re-judged, all 2,218 questions carry fresh judgments for all seven approaches, so no cells are excluded. Semantic-axis labels join from the generation export for 2,199 of 2,218 questions. Every gold answer in the archive passed the full pipeline: answer validation, the deterministic numeric audit, and dual support judging. Humans have spot-checked exported contracts during ongoing curation, but a systematic blinded audit of gold answers is future work, so absolute rates remain component diagnostics. Class-conditional differences are the object of interest.

Table 17: Class-separated archive-wide answer metrics (all seven approaches fully judged). "Answer" is accuracy on the 1,911 answerable questions. "Abstain" is correct abstention on the 307 unanswerable questions. Brackets are account-clustered 95% bootstrap intervals. "Bal." is their unweighted mean. Paired columns are wins/losses against raw orders within each class.

Memory approach	Answer [95% CI]	Abstain [95% CI]	Bal.	Ans. W/L	Unans. W/L
Raw orders	0.665 [0.643, 0.687]	0.616 [0.559, 0.671]	0.640	—	—
System A + orders	0.678 [0.653, 0.703]	0.528 [0.470, 0.586]	0.603	204/178	11/38
System B + orders	0.659 [0.638, 0.682]	0.550 [0.487, 0.613]	0.605	147/157	15/35
System C + orders	0.642 [0.617, 0.666]	0.603 [0.553, 0.655]	0.622	135/179	15/19
System A	0.545 [0.518, 0.571]	0.638 [0.589, 0.687]	0.592	162/390	33/26
System B	0.199 [0.176, 0.223]	0.827 [0.784, 0.870]	0.513	65/955	82/17
System C	0.146 [0.127, 0.165]	0.889 [0.854, 0.922]	0.518	46/1,037	90/6

Information sufficiency separates the tiers, and abstention calibration separates the leaders. On answerable questions the four orders-backed approaches cluster at the top, standalone System A sits in between, and standalone Systems B and C trail far behind (Table 17). Unanswerable questions invert the latter result: Systems B and C abstain correctly more often than every orders-backed approach. System A + orders shows the cancellation directly. It wins more answerable comparisons against raw orders than it loses, but loses most unanswerable comparisons (the paired win/loss columns of Table 17). The paired answerable difference is 0.014 (95% CI $[-0.009, 0.037]$, exact account-cluster sign-flip p_{Holm}=0.55). Its unanswerable difference is -0.088 ($[-0.135, -0.041]$, p_{Holm}=0.0043). After Holm correction within each six-comparison class, none of the three orders-backed approaches differs from raw history on answerable questions. All three standalone systems are worse. On unanswerable questions, Systems B and C are better than raw history, Systems A + orders and B + orders are worse, and standalone System A and System C + orders are not distinguishable from raw history. Balanced accuracy separates the orders-backed approaches only narrowly (Table 17). A seven-approach oracle is a hypothetical router that answers each question with whichever of the seven approaches handles it correctly. It is an upper bound for any combination of them. This oracle answers 84.0% of answerable questions and abstains correctly on 93.8% of unanswerable ones. The

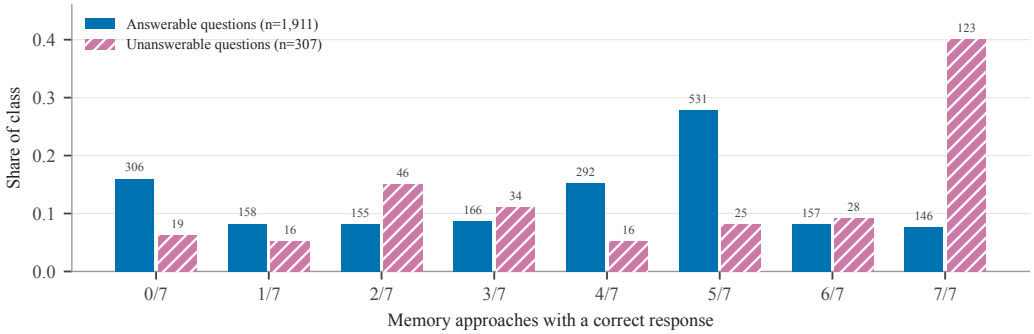

Figure 11: Questions by the number of memory approaches with a correct response, per question class. Every value from 0 of 7 to 7 of 7 occurs in both classes. 306 answerable questions defeat every approach, and 146 are answered by all seven. The count reflects both question difficulty and each approach's information sufficiency, since standalone Systems B and C fail most answerable questions by construction. The histogram therefore shows the benchmark's difficulty spread, not a pure item-difficulty scale.

remaining 306 answerable and 19 unanswerable questions are universal failures that defeat all seven approaches. Roughly one in six answerable questions is answered correctly by some approach but not by the leader. Failures therefore overlap little. Part of the gap reflects how the standalone systems are built, not complementarity among peers. The 306 answerable universal failures show the benchmark is not saturated even under best-case routing. They are also the natural priority slice for a human audit of gold answers. A question that some approach answers correctly corroborates its gold answer. A universal failure does not separate a hard question from a defective gold answer.

Difficulty spread (Figure 11). For each question we count how many of the seven approaches produce a correct response. Every count from 0 of 7 to 7 of 7 is populated in both classes. The benchmark therefore contains questions at every level of difficulty and does not cluster at trivially easy or uniformly impossible ones. This is the per-question evidence behind the complexity component of §3. On answerable questions the distribution peaks at 5 of 7 correct (27.8% of the class), the four orders-backed approaches plus standalone System A. On unanswerable questions it peaks at 7 of 7 (40.1%), where every approach abstains correctly.

Evidence regimes are the benchmark's sharpest discriminator (Figure 12). On significant-evidence questions the orders-backed approaches cluster at 0.65–0.70 while standalone Systems B and C reach 0.20–0.27. On *some-evidence* questions (sparse support, the regime a persistent memory should cover), Systems B and C collapse to 0.03–0.04, the single largest separation in the benchmark. Their standalone representations retain dominant patterns but lose the marginal observations that sparse questions require. The unanswerable questions then invert the ranking. Systems B and C abstain best because their context contains nothing to over-read. The richest context (System A + orders) abstains correctly least often (Table 17). Neither number is good on its own. The benchmark exposes the trade-off instead of rewarding one side of it.

Failure codes turn the trade-off into a calibration axis (Figure 13). Unwarranted abstentions on answerable questions rise monotonically as context shrinks, from 1.6% for System A + orders to 62% for standalone System C, over half of that class. In the opposite direction, missed abstentions on unanswerable questions fall from 47% to 11% along the same ordering. Within the answerable class, orders-backed approaches otherwise share similar failure profiles (led by `WRONG_EVIDENCE_LEVEL` at 607–677 and `CONTRADICTORY_EXTRA_CLAIM` at 294–373). Unanswerable-question failures are reported separately as missed abstentions rather than folded into those counts.

Semantic axes localize what compression destroys (Figure 14). On answerable questions, standalone Systems B and C have episodic accuracy of 0.07–0.11 and retrieval accuracy of 0.07–0.13, versus 0.29–0.39 on inference demands and 0.20–0.26 on semantic questions. The separate unanswerable panel shows why the class split matters. Systems B and C abstain correctly on nearly all

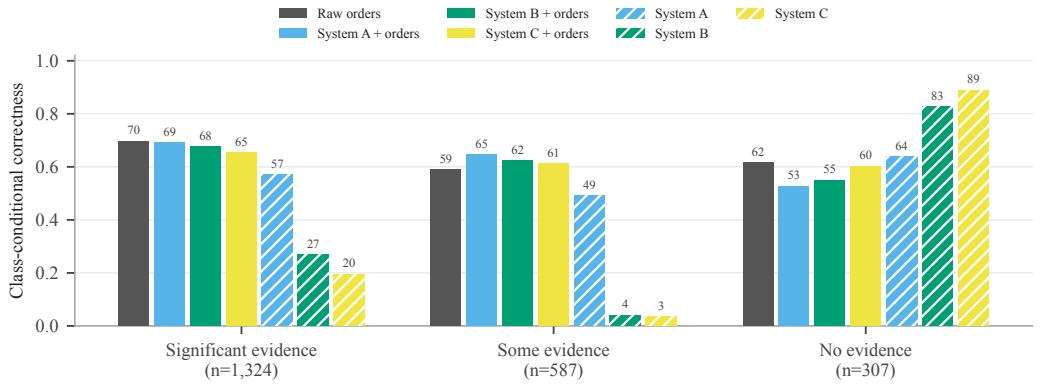

Figure 12: Class-conditional correctness by expected evidence regime (fully judged archive, n=2,218). The significant- and some-evidence bars are answerable-question accuracy. The no-evidence bar is the unanswerable-question correct-abstention rate, with a separate denominator printed for every regime. The evidence axis is what distinguishes the approaches. Standalone Systems B and C nearly zero out on some-evidence questions yet lead on unanswerable questions. They lead not because they judge evidence well but because they lack the information to be tempted. Richer context inverts the pattern: System A + orders is strongest on answerable questions and weakest at abstaining.

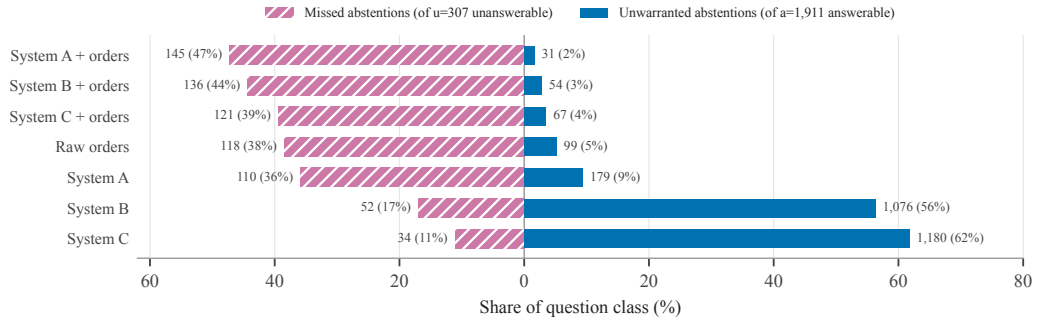

Figure 13: The two abstention failure modes per memory approach, as shares of their question class. Unwarranted abstentions (right) are answerable questions the approach declined to answer. Missed abstentions (left, `SHOULD_ABSTAIN`) are unanswerable questions it answered anyway. Approaches are ordered by context richness, and the two sides move in opposite directions along it: richer context answers more but abstains less.

episodic and retrieval questions, so a pooled score would credit their abstention as if it were recall. The answerable panel shows that event-level detail is gone while dispositional summaries partially survive, without credit from a different task type.

Fine-grained targets, predicates, and conditions (Figures 15 and 16). The finer registries show the same pattern in more detail and add slices the coarse axes hide. We applied Holm correction to all comparisons shown in the slice figures, separately for answerable and unanswerable questions. A slice is one semantic category, such as cuisine or delivery location. After correction, no orders-backed approach outperformed raw history in any slice. Among answerable questions, 103 differences remained significant, and every one was worse than raw history. Among unanswerable questions, six differences remained significant, all from the higher correct-abstention rates of standalone Systems B and C. Four patterns stand out.

- *(a) Aggregation survives compression. Entities do not.* On the `cuisine` target, System A + orders reaches 0.70 versus 0.51 for raw orders. Even standalone System A reaches 0.67, so cuisine-level tendencies survive in its exported representation. Entity-level targets collapse

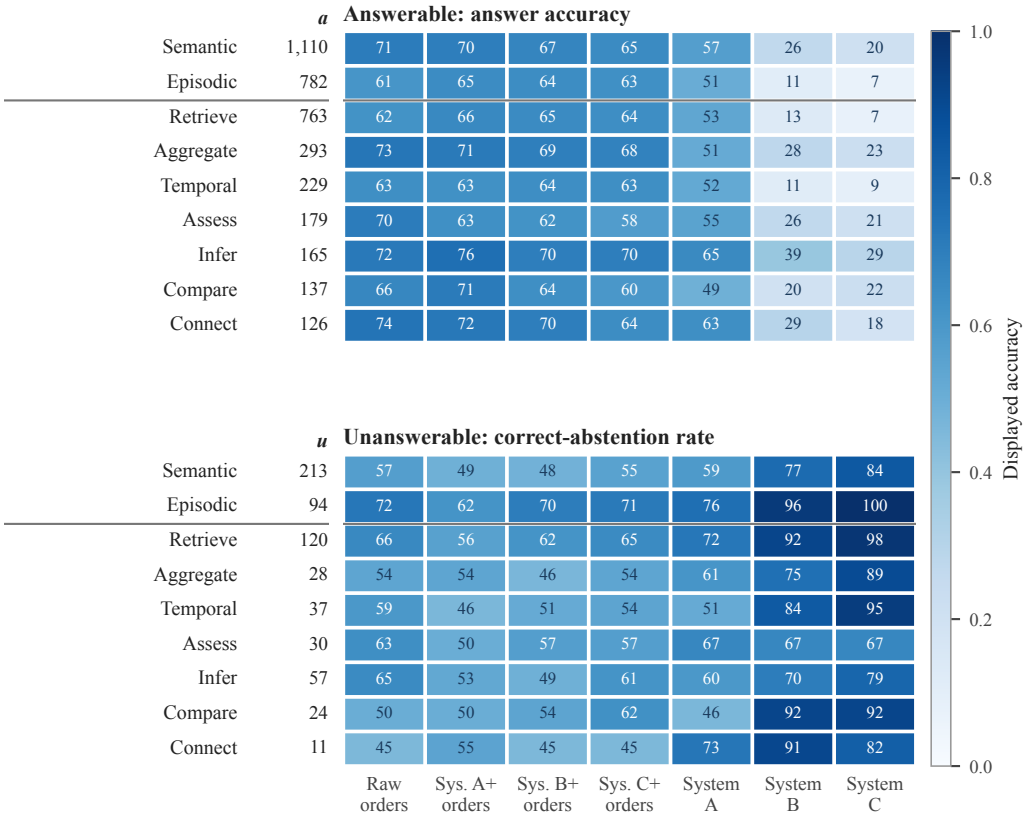

Figure 14: Class-separated performance by realized memory function and memory demand. The top panel reports answer accuracy on the 1,892 axis-joined answerable questions. The bottom panel reports correct-abstention rate on all 307 axis-joined unanswerable questions. The top panel's a column gives the number of answerable questions, and the bottom panel's u column gives the number of unanswerable questions. Nineteen answerable questions lack a joined semantic-axis record.

without raw history: standalone Systems B and C reach only 0.10–0.14 on `merchant`, `variety`, and `item`.

- *(b) Absence claims are uniformly hard.* `orders_unmodified` asserts that the account does *not* customize. It is the weakest predicate for every approach: 0.37–0.47 orders-backed, 0.24 for System A, and 0.03–0.05 for Systems B and C. Verifying a universal negative requires scanning the full history, and no representation currently supports that well. The related `dish_type` target slices sit at 0.38–0.48 even for orders-backed approaches, marking the questions where item-name-to-category normalization is the bottleneck.

- *(c) Location conditioning shows the largest descriptive lift from augmentation, but it does not survive correction.* On the 122 answerable delivery-location-conditioned questions, System A + orders leads at 0.73 versus 0.60 for raw orders. The paired difference is 0.13 (account-clustered 95% CI $[0.04, 0.22]$), but its Holm-adjusted p-value is 1.0 in the pre-specified family of 294 answerable axis-slice contrasts. It is therefore descriptive, not confirmatory. Delivery location reaches a question through three axes, and the union of those 200 answerable questions separates the configurations more than any other slice in the archive: System A + orders reaches 0.79 while standalone Systems B and C reach 0.03–0.04. Address structure is explicit in System A's exported context but must be reconstructed from raw rows. This slice also shows the evaluation's correction loop. Every displayed value uses the re-judged, location-corrected candidates, so a candidate-export defect is not attributed to the memory representation.

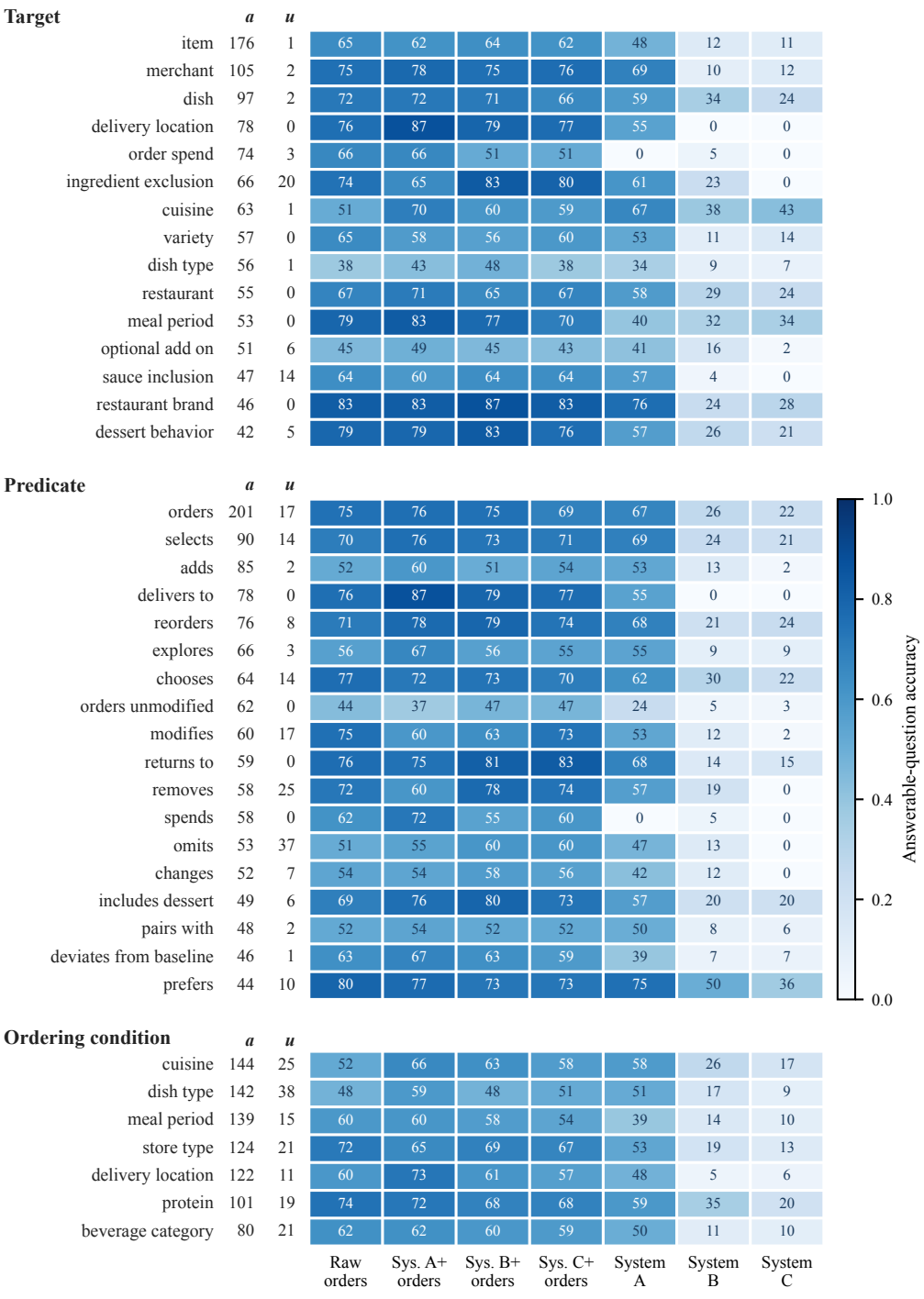

Figure 15: Answerable-question accuracy by realized target (top, $a \geq 40$), predicate (middle, $a \geq 40$), and ordering condition (bottom, $a \geq 15$) for each memory approach. The adjacent a and u columns give the numbers of answerable and unanswerable questions. Inclusion thresholds use a only. Fine-grained slices localize each representation's strengths. The standalone systems recover cuisine-level aggregates but lose entity-level targets. Absence claims (`orders_unmodified`) are hard for every approach. Delivery-location conditioning shows the largest displayed descriptive system-context lift over the raw-history baseline, but it does not survive the slice-wide multiplicity correction.

- *(d) Unanswerable questions concentrate unevenly.* Some slices are dominated by their unanswerable class. `owns` contains only 9 answerable questions but 36 unanswerable questions, and pet ownership contains 8 answerable questions but 35 unanswerable questions. Neither reaches the answerable-question display threshold in Figure 15. Their abstention behavior is shown instead in Figure 16.

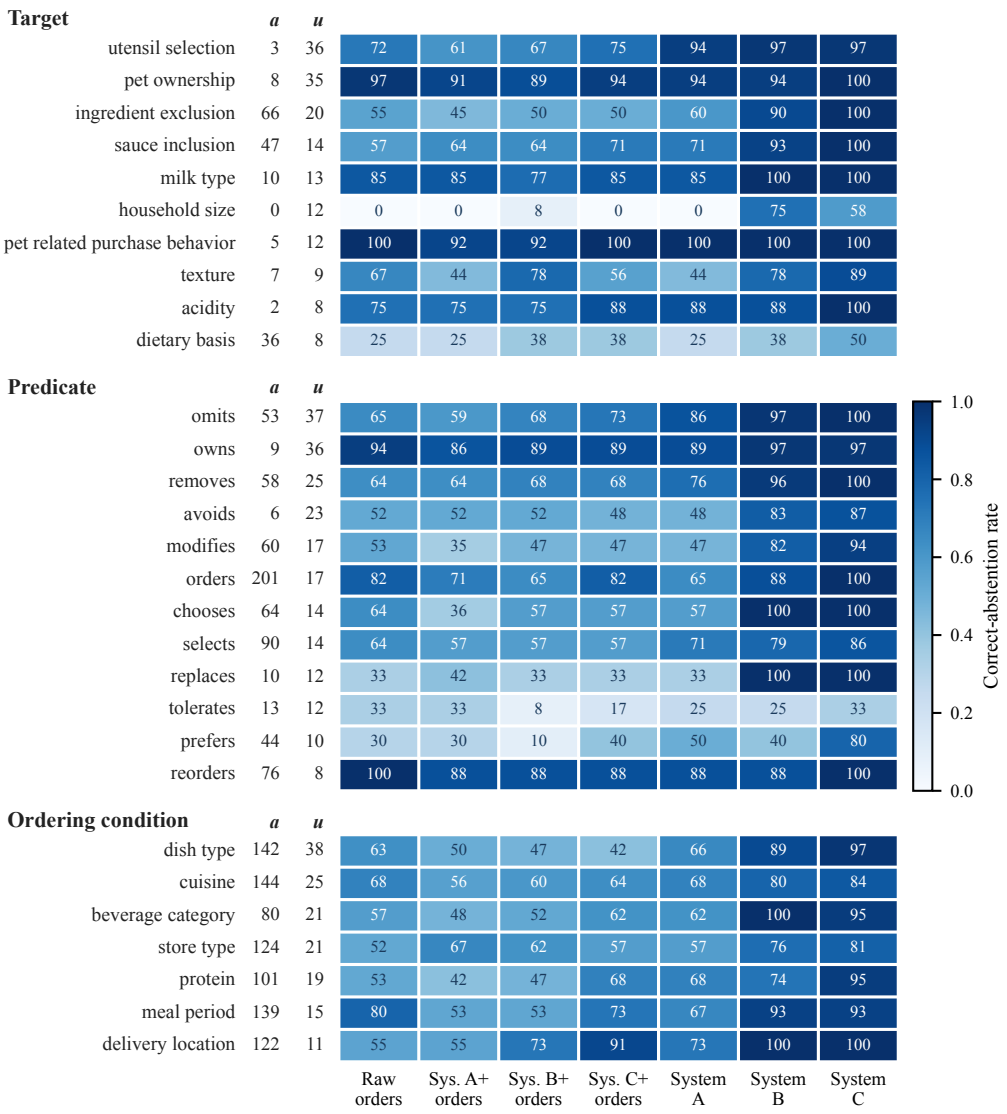

Figure 16: Correct-abstention rate on unanswerable questions by realized target (top, $u \geq 8$), predicate (middle, $u \geq 8$), and ordering condition (bottom, $u \geq 10$) for each memory approach. The adjacent a and u columns give the numbers of answerable and unanswerable questions. Inclusion thresholds use u only. This is the unanswerable-question counterpart to Figure 15, not a pooled continuation of it.

In this sense the benchmark distinguishes memory systems. It gives no single ranking. It attributes accuracy to the kind of recall each representation still supports, the evidence regime it can still calibrate against, and the abstention behavior its context induces.

G Replication on the Delivery Hero Recommendation Dataset

This appendix is the detailed reference for PantryBench (§4.1): the dataset and its provenance, the adapter, the run protocol, and the alignment with the private cohort. PantryBench builds on the Delivery Hero Recommendation Dataset (DHRD) from RecSys '23 (Assylbekov et al., 2023) and uses no private data at any step. The reproducibility statement lists the released artifacts. Compared with the private run, the replication changes the platform, the catalog languages, the history density, and the authoring model family at the same time. Four results are aligned across the two runs: the funnel yield, complete coverage of the supported taxonomy, the match rate between requested and realized axes, and the diversity collapse of free draws. The private memory-approach comparison of Appendix F.7 is out of scope here. Its public counterpart is the study of §4.3, whose ablations and per-axis breakdowns close this appendix (Appendix G.8). The replication runs in all three cities (Appendix G.3). One comparison then varies a single factor: the authoring family, on the fixed Stockholm cohort (Appendix G.5).

G.1 Dataset and provenance

DHRD contains roughly 9.4 million real food-delivery order-item rows from three cities (Singapore, Stockholm, Taipei) over about 90 days. The upstream release anonymizes all identifiers with salted hashes and carries the MIT license. Each city ships three tables. The orders table holds pseudonymous customer, order, vendor, and product identifiers, the weekday, the time of day, the day offset from the start of the window, and a five-character location geohash. The product catalog holds the display name in the local language mix and a normalized unit price. The vendors table holds chain membership and the primary cuisine. The upstream distribution is an unversioned file share, so the exact bytes are pinned. The three archives are vendored with SHA-256 manifests together with the upstream documentation. Every cohort and run below records the archive checksums it consumed. MARS (Tian & Wang, 2026) is the only other LLM work on this dataset that we found. It re-ranks repeat-order recommendations on the Stockholm and Singapore splits and does not evaluate memory.

G.2 Adapter: from DHRD to the canonical history schema

A deterministic cohort builder converts DHRD into the pipeline's canonical input format, so planning, generation, validation, and judging run unchanged. Four mapping decisions carry the semantics. (1) *The timeline is synthetic.* DHRD publishes only relative day offsets. Dates therefore derive from an arbitrary epoch, chosen so that every derived calendar weekday agrees with the source weekday column (verified per row). Clock times, weekdays, and day spacing are real. The dates are not. (2) *Merchants are pseudonymous.* DHRD has no vendor names, so stores and chains appear as "Vendor n" and "Chain n" labels. (3) *Quantity is never inferred.* DHRD does not define what a repeated order–product row means, so each row counts as one product occurrence. (4) *Prices are synthetic benchmark values.* A fixed rule maps the normalized catalog price to a synthetic price between \$1 and \$25. Price and spend questions therefore ground, but they test retention of deterministic benchmark values, not historical charges. The five-character order geohash is copied unchanged into both the delivery area and the address surrogate. The upstream release publishes these values openly, so they may appear in questions and answers. Sampling is source-aware: before any LLM call, the profile resolves which semantic values the 19 retained fields can ground. Of the 142 enabled values, 125 are supported. The 17 unsupported values all depend on modifier selections or group-order context. They are five predicates such as `adds` and `removes`, and twelve targets such as `milk_type` and `utensil_selection`. They are excluded from sampling, and the run report lists them separately from covered and uncovered categories. A value the source can never ground is therefore not counted as a coverage gap or a generation loss. Customer-level absence remains a valid no-evidence case, because the source could observe the behavior for another customer.

Cohorts are deterministic and pinned. The builder reads only the upstream train split (days 0–75), which keeps the upstream test days clean for a later temporal-update study. A customer is eligible

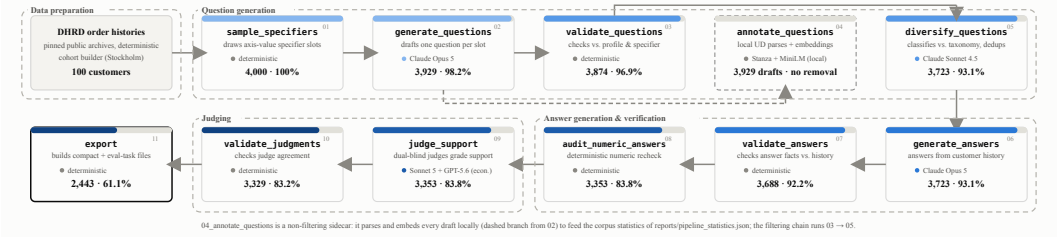

Figure 17: The generation pipeline with per-stage retention on the DHRD Stockholm run (compare Figure 1, the private-cohort run). This run includes the non-filtering annotation stage (dashed). It parses and embeds every draft locally and removes nothing, so the filtering chain routes from question validation directly to diversification. The counts match Table 18.

with 20 or more distinct train orders. Selection orders eligible customers by a seeded hash, so rebuilds are byte-identical and a smaller same-seed cohort is a strict prefix of a larger one. Each cohort manifest records source and output hashes. It also records a representativeness check. The check compares the cohort against its whole eligible pool with a two-sample Kolmogorov–Smirnov test on each of seven behavioral metrics: distinct orders, item rows, distinct vendors, distinct products, active span in days, mean order hour, and weekend share. On the cohorts for all three cities, no metric differs significantly (minimum $p = 0.13$).

G.3 Replication runs

The replication runs once per city with one configuration. Stockholm anchors the report: 100 customers, 2,642 orders, and 4,823 order-item rows, drawn from the city's 1,082 eligible customers. This is roughly 23× sparser than the private cohort's 112,941 rows, so the run also probes the thin-history end of the design space. Sampling uses the production all-axes profile with 40 specifications per customer, 4,000 slots in total, restricted to the 125 source-supported values. The diversification classifier (Claude Sonnet 4.5) and the judge pair (Claude Sonnet 5 primary, GPT-5.6 economy secondary) are unchanged. The authoring family is different on purpose: Claude Opus 5 drafts questions and gold answers, where the private run used GPT-5.6. Agreement between the runs is therefore evidence of robustness to all four changes at once. A disagreement would not identify which change caused it.

Table 18: Generation funnel of the DHRD Stockholm run, in the format of Table 7. Retention is relative to sampled slots. A question can carry more than one typed defect, so reason counts can exceed the questions lost.

Stage	Count	Retention	Dominant loss reasons
Sampled specification slots	4,000	—	—
Drafted questions	3,929	0.982	drafting omission (71)
Valid questions	3,874	0.969	wrong-field condition (49), compound question (6)
Retained after diversification	3,723	0.931	leading questions (96), pattern cap (26)
Answered	3,723	0.931	—
Valid answers	3,688	0.922	unsupported fact (31), schema wording (10)
Numeric-audit valid	3,353	0.838	winner mismatch (171), count mismatch (164)
Support judged	3,353	0.838	—
Valid judgments	3,329	0.832	evidence-kind mismatch (22)
Exported	**2,443**	**0.611**	dual-judge disagreement (886)

Figure 17 and Table 18 show the funnel. The shape replicates: intended quality gates dominate the losses, led by dual-judge disagreement at export and the deterministic numeric audit. The end-to-end yield is close to the private run's (Table 27). Because source-impossible requests never enter sampling, question validation loses only 55 drafts, and no draft fails for citing a field the source lacks. Confirmed unanswerable questions are 225 of the exported questions, a smaller share than on the private cohort (Table 27). With schema gaps excluded up front, the remaining unanswerable questions reflect customer-level absence only. The diversification stage also removed 96 leading

questions, through a filter added after the private run. Coverage is complete over the source-supported taxonomy. Every one of the 125 supported values appears in at least one exported question, no explicitly sampled category lost all its drafts, and no category was reached only through free draws. The flywheel's fidelity also transfers: Table 19 compares the per-axis match rates over exported questions with the private run's, computed identically (§4.2). Memory demand is again the axis with the most mismatches.

Table 19: Constrained-axis match rates over exported questions, per axis: the private run versus the DHRD Stockholm run.

Axis	Private	Stockholm
memory function	100.0%	99.9%
ordering condition	99.9%	99.8%
predicate	99.7%	98.8%
target	99.5%	99.1%
memory demand	98.1%	97.9%
overall	99.6%	99.1%

Pairwise combination coverage also replicates, close to the private share, with target $\times$ ordering condition again the weakest pair (Table 28).

The Singapore and Taipei cohorts rerun the replication profile without any other change. Claude Opus 5 authors all three runs, and the cohort builder, sampling profile, run seed, classifier, and judges stay fixed. The cities are fixed replication environments, not samples from a population, so the runs stand side by side and are never pooled. Customer labels are city-prefixed, so the three runs are independent draws rather than paired ones.

Table 20 shows the three funnels, and the replication findings hold in every city. Yield stays in a narrow band across cities (Table 27). Coverage stays complete over the 125 source-supported values. Match rates, unanswerable shares, and pairwise coverage stay in narrow bands (Tables 27 and 28). Target $\times$ ordering condition stays the weakest pair in every city. The loss profile also transfers, with dual-judge disagreement and the deterministic numeric audit dominating everywhere.

Table 20: Generation funnels of the three city runs with the authoring family fixed at Claude Opus 5, in the format of Table 18. Retention is relative to the 4,000 sampled slots.

	Stockholm		Singapore		Taipei	
Stage	Count	Ret.	Count	Ret.	Count	Ret.
Sampled specification slots	4,000	—	4,000	—	4,000	—
Drafted questions	3,929	0.982	3,868	0.967	3,925	0.981
Valid questions	3,874	0.969	3,832	0.958	3,888	0.972
Retained after diversification	3,723	0.931	3,685	0.921	3,792	0.948
Answered	3,723	0.931	3,685	0.921	3,792	0.948
Valid answers	3,688	0.922	3,669	0.917	3,773	0.943
Numeric-audit valid	3,353	0.838	3,372	0.843	3,469	0.867
Support judged	3,353	0.838	3,372	0.843	3,469	0.867
Valid judgments	3,329	0.832	3,361	0.840	3,449	0.862
Exported	**2,443**	**0.611**	**2,344**	**0.586**	**2,534**	**0.634**

Table 21 grounds the replication in the catalogs themselves. The same machinery, running unchanged, asks about semla pastries and café breakfasts in Stockholm, kopi and nasi lemak in Singapore, and stir-fried noodles and milk tea in Taipei. City-specific cuisine enters only through the histories. No prompt, profile, or taxonomy value changes between the runs.

Table 1 (main text, §3) draws five further exported contracts from this run. They span five memory demands and all three evidence regimes, including a confirmed no-evidence abstention.

Table 21: Exported question–answer pairs from the three city runs: one composition question and one trend question per city. The vocabulary shifts with the catalog while the question machinery stays fixed.

City	Question	Gold answer
Stockholm	When I order at breakfast time, what kind of dish do I usually end up with?	A cheese breakfast roll from the café, usually alongside orange juice and a semla.
Stockholm	Looking at how my orders have shifted, which kind of dish has taken the place of another one I used to order?	A croissant now fills the pastry slot that the Semla used to occupy.
Singapore	What food-and-drink combination did I put together within a single order?	Mee Siam and Nasi Lemak with Kopi Tarik and Teh O
Singapore	For my breakfast orders, how has the range of different dishes I try widened over time?	Breakfast started as noodle and bee hoon plates and has widened into porridge, chicken rice, Malay value meals and a multi-item dim sum spread.
Taipei	When I order Taiwanese food, which food-and-drink combination do I usually put together in the same order?	Grilled-pork stir-fried noodles with a milk-based drink (latte or milk tea).
Taipei	When it comes to stir-fried noodle dishes, how has my appetite for them changed over the months I've been ordering?	Increased — from one stir-fried noodle order in February to two in March, with none in January.

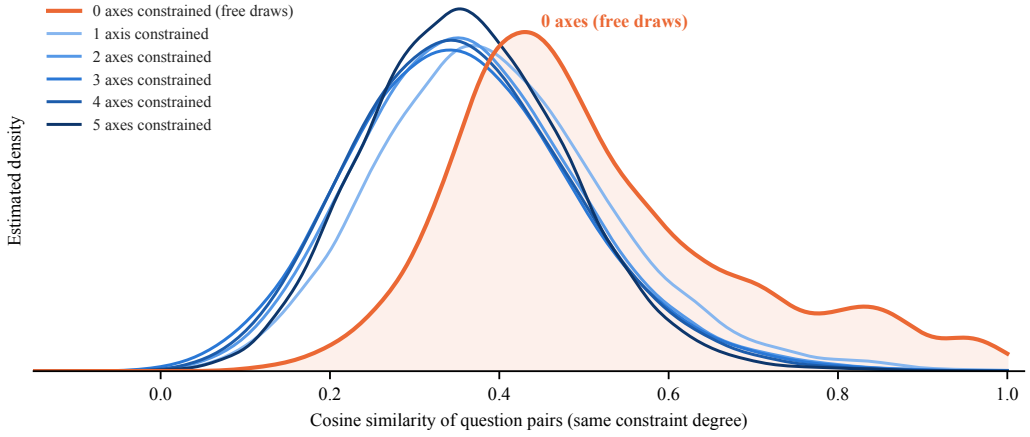

Figure 18: The measurement of Figure 8 repeated on the DHRD Stockholm run: kernel density of pairwise cosine similarity (all-MiniLM-L6-v2) between drafted questions with the same number of constrained axes, over all 3,929 drafts. Free draws are again the least diverse population by a wide margin, with a heavy high-similarity tail. The private run's near-duplicate mode at $\sim$1.0 does not appear.

G.4 The diversity collapse replicates

Figure 18 repeats the measurement of Figure 8 on the 3,929 Stockholm drafts, and Table 22 aligns the private cohort with the three city runs. The main findings of Appendix F.2 recur. Free draws are by far the most self-similar population, and constraining a single axis removes most of the collapse (Table 22). The constrained degrees then form a shallow plateau. The private run's mild rise at full specification does not clearly recur: the five-axis mean sits within one bootstrap standard deviation of the constrained minimum. Two differences between the cohorts are informative. The DHRD free-draw distribution has a heavy high-similarity tail but no near-duplicate mode at $\sim$1.0, and the collapse is milder than the private run's (Table 22). This fits a different authoring family that paraphrases its repeated questions instead of copying them. Appendix G.5 tests this reading. The same histories authored by GPT-5.6, the private run's family, bring the mode back (Figure 2). Unique 3-gram counts are higher at every degree (629 versus 419 for free draws, 1,255–1,851 versus 880–1,342 for constrained drafts). The multilingual item names and the alphanumeric vendor labels widen the vocabulary, so this lexical measure compares only within a cohort, not across data sources.

Free-draw drift selects the same dominant categories as Table 15. With axes left unconstrained, realized volume concentrates on the same five categories, at volumes close to the private run's (Table 23).

Table 22: Mean pairwise cosine similarity of drafted questions by constraint degree, before diversification (all-MiniLM-L6-v2 embeddings, equal-size bootstrap subsamples, mean $\pm$ 1 sd, lower = more diverse). The table compares the private cohort with the three PantryBench city runs. The exported corpus is summarized in Table 27. Every run samples the same specification profile. Degree 0 is the unsteered free-draw control. The private and Stockholm runs differ in data source and authoring family. The three city runs share Claude Opus 5 authoring.

Degree	Private		Stockholm		Singapore		Taipei	
	n	Similarity	n	Similarity	n	Similarity	n	Similarity
0	216	0.561 ± 0.009	235	0.508 ± 0.005	220	0.538 ± 0.005	230	0.536 ± 0.007
1	162	0.387 ± 0.003	162	0.397 ± 0.004	173	0.425 ± 0.004	154	0.420 ± 0.002
2	772	0.341 ± 0.012	810	0.369 ± 0.009	741	0.387 ± 0.012	788	0.366 ± 0.009
3	1,487	0.354 ± 0.009	1,478	0.360 ± 0.008	1,487	0.371 ± 0.013	1,519	0.362 ± 0.010
4	1,078	0.381 ± 0.009	1,059	0.359 ± 0.008	1,050	0.381 ± 0.011	1,049	0.376 ± 0.010
5	190	0.400 ± 0.005	185	0.365 ± 0.004	197	0.382 ± 0.004	185	0.391 ± 0.003

Table 23: Free-draw drift on the DHRD Stockholm run, for five dominant categories of Table 15. The private-run counterparts stand there.

Axis	Category	Sampled	Realized	From free draws
memory demand	retrieve	219	1,083	877
memory demand	aggregate	244	867	628
memory function	semantic	1,122	2,753	1,668
predicate	orders	157	479	331
target	item	108	302	192

The collapse of unsteered drafting is therefore not a quirk of one platform's histories or one model family's habits. It reproduces on public data, on the same categories. This supports the paper's core position: treat diversity as a property to construct and measure, not to assume.

Singapore and Taipei repeat the pattern (Table 22). Free draws are the least diverse population in every city, and constraining a single axis removes most of the collapse in every city. The Vendi count, which measures diversity as a number of distinct questions instead of a mean similarity, shows the same collapse in every city (Table 27). The constrained degrees again form a shallow plateau in each city.

G.5 Authoring-family comparison

The Stockholm cohort also supports a controlled comparison of authoring families. All three runs share the cohort, the sampling profile, the run seed, and the classifier and judge models, and their sampled plans are byte-identical. Only the authoring family changes: Claude Opus 5, GPT-5.6, and Gemini 3.5 Flash. Claude Opus 5 is the replication run above, and GPT-5.6 is the private run's family. Low reasoning effort applies to Gemini 3.5 Flash. Default reasoning settings apply to Claude Opus 5 and GPT-5.6. The same 4,000 slots therefore enter every run, and the runs pair slot by slot. Table 24 shows the three funnels.

The funnels differ strongly between families. End-to-end yield is family-specific and spans 38.9% to 66.8%. The gates that separate the families differ as well. Gemini 3.5 Flash loses the most at question validation, with 99 unobserved bindings and 84 wrong-field conditions. Claude Opus 5 loses the most at answer validation, with 31 unsupported facts. GPT-5.6 drafts cleanly but repetitively: diversification removes 58 duplicate recommendations and 81 questions at the run-wide pattern cap, both the highest of the three. GPT-5.6 also drafts the most unanswerable questions. It drafts 1,451 no-evidence questions, 37.5% of its drafts versus 12.5% to 13.4% for the other two. Its confirmed unanswerable share of the export (16.2%) is second only to Gemini 3.5 Flash (22.2%, Table 27). The Gemini 3.5 Flash share is high for a different reason. Between answering and export it keeps 90% of its abstentions but 39% of its grounded answers. The numeric audit loses between 237 (Gemini 3.5 Flash) and 335 (Claude Opus 5) questions per family. Its checks run deterministically against the shared histories, so its losses vary little with the author. Dual-judge disagreement at export is not flat: the fixed judge pair discards far more Gemini 3.5 Flash answers than Claude Opus 5 and GPT-5.6 answers (Table 27). Gemini 3.5 Flash loses 50% of its valid judgments there.

Table 24: Generation funnels of the three authoring-family runs on the Stockholm cohort, in the format of Table 18. All three runs consume byte-identical sampled plans. Retention is relative to the 4,000 sampled slots.

Stage	Claude Opus 5 Count	Ret.	GPT-5.6 Count	Ret.	Gemini 3.5 Flash Count	Ret.
Sampled specification slots	4,000	—	4,000	—	4,000	—
Drafted questions	3,929	0.982	3,873	0.968	3,894	0.974
Valid questions	3,874	0.969	3,865	0.966	3,702	0.926
Retained after diversification	3,723	0.931	3,616	0.904	3,513	0.878
Answered	3,723	0.931	3,616	0.904	3,513	0.878
Valid answers	3,688	0.922	3,614	0.904	3,483	0.871
Numeric-audit valid	3,353	0.838	3,360	0.840	3,246	0.812
Support judged	3,353	0.838	3,360	0.840	3,246	0.812
Valid judgments	3,329	0.832	3,339	0.835	3,113	0.778
Exported	**2,443**	**0.611**	**2,671**	**0.668**	**1,555**	**0.389**

What the authoring family does not change is the benchmark's structure. All three families pass the replication's structural checks. Coverage is complete in every run: each of the 125 source-supported values appears in at least one exported question. Constrained axes match their request 98.8% to 99.6% of the time, with memory demand the weakest axis in every run. Pairwise combination coverage tracks export volume and spans 72.0% to 83.8% of the 2,264 jointly sampleable axis-value pairs (Table 28). Target $\times$ ordering condition is the weakest pair in every run. Table 27 aligns the three families with the private run. This split is the comparison's main finding. The authoring family decides how many questions survive and where they fail, while the specification machinery keeps coverage and fidelity fixed.

Table 25: Mean pairwise cosine similarity of drafted questions by constraint degree, per authoring family, computed as in Table 22 (all-MiniLM-L6-v2 embeddings, equal-size bootstrap subsamples, mean $\pm$ 1 sd, lower = more diverse). Degree 0 is the unsteered free-draw control.

Degree	Claude Opus 5 n	Similarity	GPT-5.6 n	Similarity	Gemini 3.5 Flash n	Similarity
0	235	0.508 ± 0.005	235	0.578 ± 0.010	201	0.482 ± 0.006
1	162	0.397 ± 0.004	160	0.397 ± 0.004	159	0.365 ± 0.002
2	810	0.369 ± 0.009	804	0.349 ± 0.010	804	0.336 ± 0.007
3	1,478	0.360 ± 0.008	1,460	0.344 ± 0.009	1,480	0.327 ± 0.009
4	1,059	0.359 ± 0.008	1,033	0.359 ± 0.011	1,064	0.340 ± 0.009
5	185	0.365 ± 0.004	181	0.369 ± 0.005	186	0.360 ± 0.004

Table 25 repeats the diversity measurement per family. The free-draw collapse appears in all three runs, and the families disagree most on the free draws. Mean free-draw similarity ranges from 0.482 (Gemini 3.5 Flash) to 0.578 (GPT-5.6), a spread of 0.096. At every constrained degree the spread across the three families is at most 0.033. Gemini 3.5 Flash has the lowest similarity at every degree. Constraints therefore equalize the families, and only unconstrained drafting exposes each model's own repetition habit. GPT-5.6 pairs the strongest collapse with the smallest vocabulary. It has 327 unique 3-grams on free draws, versus 629 and 691 for Claude Opus 5 and Gemini 3.5 Flash. The Vendi count agrees: in every family, 40 exported free draws contain far fewer different questions than 40 constrained questions at any degree (Table 27). Constraining a single axis removes most of the collapse in every family. The collapse of unsteered drafting, and the constraint mechanism that repairs it, are therefore not specific to one frontier model.

Figure 2 (main text, §4.2) draws the full curves behind Table 25, one panel per constraint degree. It adds the Claude Opus 5 replications on Singapore and Taipei from Table 22. The free-draw panel is the panel in which the families separate most. GPT-5.6 sits furthest right and is the only family with a near-duplicate mode at $\sim$1.0. The Claude Opus 5 free-draw curve has a heavy high-similarity tail but no such mode. Every panel also shows the private cohort, which GPT-5.6 authored on a different data source. That shaded curve follows the GPT-5.6 curve, including the near-duplicate mode, and not the Claude Opus 5 curve on the same histories. The near-duplicate mode is therefore a habit of the authoring family and not of the data source. The Singapore and Taipei curves make the same point from the other side: the same family on three cities gives three nearly identical curves, none with the mode. In the five constrained panels the three curves nearly coincide in position and width,

with Gemini 3.5 Flash slightly to the left. The private cohort's shaded area lies within the same band in each constrained panel. Its small shift to the right at four and five axes is the mild rise at full specification of Figure 8. This is the picture behind the small constrained-degree spread in Table 25.

Table 26: One sampled slot, three authoring families. The slot requests memory demand `connect`, predicate `is_go_to_choice`, target `restaurant`, and ordering condition `cuisine`, with the memory function left free. Every run drafted this slot from the same customer history, and each question survived to export with all constrained axes matched. Vendors appear under their pseudonymous DHRD labels, and DHRD publishes its cuisine vocabulary openly.

Family	Question	Gold answer
Claude Opus 5	When I'm in the mood for Asian food, which restaurant is my reliable go-to?	Vendor 1a9c0c9a
GPT-5.6	When ordering Asian cuisine, which restaurant has been my go-to choice?	Vendor 1a9c0c9a
Gemini 3.5 Flash	Which specific restaurant location is my go-to choice when I want sallad?	Vendor 649b7d23

Table 26 shows the three questions drafted for one shared slot. All three realize the same specification, but each family frames it in its own way. Claude Opus 5 and GPT-5.6 bind the free condition value the same way and name the same vendor. Gemini 3.5 Flash binds it differently and names a different vendor. The specification pins what a question must probe. The authoring family still owns the framing and the concrete binding, and downstream consumers of the benchmark inherit both.

G.6 Cross-run summary

Table 27 collects the headline numbers of all six runs: the private run, the three authoring families on the Stockholm cohort, and the two further cities. Table 28 does the same for pairwise combination coverage. Three regularities span every run. Match rates stay at or above 98.8%. Pairwise coverage tracks export volume, and target $\times$ ordering condition is the weakest pair everywhere, on both data sources. Free draws are always the least diverse population. Yield and the confirmed unanswerable share are the volatile quantities, each spanning a wide range. The authoring family drives the widest spread in both.

Table 27: The six generation runs side by side. Yield is exported questions over the 4,000 sampled slots. Adherence is the share of requested axis values that the exported questions realize. Unanswerable is the confirmed no-evidence share of the export. Disagreement is the dual-judge disagreement count at export. Pair coverage is the realized share of the jointly sampleable axis-value pairs of each data source (2,342 private, 2,264 DHRD, Table 28). APS is the mean pairwise cosine similarity of the exported questions (equal-size subsamples of 40, 25 resamples, mean $\pm$ 1 sd). Vendi is the effective number of different questions in a subsample of 40 exported questions (Appendix F.2). Free draws are the unconstrained population. Constrained is the min–max over constraint degrees 1–5. Tables 22 and 25 give the drafting-stage per-degree values, before diversification.

Cohort	Authoring family	Exported	Yield	Axis adh.	Unansw.	Disagr.	Pair cov.	APS Free draws	APS Constrained	Vendi Free draws	Vendi Constrained
Private	GPT-5.6	2,218	55.5%	99.6%	13.8%	661	80.5%	0.488 ± 0.014	0.329–0.390	7.6 ± 0.6	12.4–15.5
Stockholm	Claude Opus 5	2,443	61.1%	99.1%	9.2%	886	81.3%	0.499 ± 0.015	0.361–0.399	7.6 ± 0.5	12.2–14.8
Stockholm	GPT-5.6	2,671	66.8%	99.6%	16.2%	668	83.8%	0.535 ± 0.022	0.344–0.393	6.4 ± 0.5	12.6–14.8
Stockholm	Gemini 3.5 Flash	1,555	38.9%	98.8%	22.2%	1,558	72.0%	0.454 ± 0.018	0.336–0.350	9.4 ± 0.6	13.4–16.2
Singapore	Claude Opus 5	2,344	58.6%	99.2%	9.3%	1,017	81.1%	0.551 ± 0.019	0.371–0.439	6.3 ± 0.4	10.3–13.8
Taipei	Claude Opus 5	2,534	63.4%	99.3%	7.7%	915	82.6%	0.525 ± 0.018	0.362–0.425	6.8 ± 0.4	11.1–14.2

Table 3 reports deterministic surface measurements for every run, with the instrument of Appendix C.5. They show an authoring-family signature that is stable across data sources. GPT-5.6 writes the same shape of question on private and public data, with question length, parse depth, subordination, and answer length nearly unchanged. Claude Opus 5 writes longer and structurally deeper questions than GPT-5.6 in all three cities and marks time most often. Gemini 3.5 Flash writes both the deepest questions of all and the shortest gold answers. Closed yes/no phrasings are absent from both GPT-5.6 exports and rare for Claude Opus 5. They are common under Gemini 3.5 Flash.

Table 28: Pairwise combination coverage of all six runs. n is the sampleable pair count of the data source. A value pair is sampleable when the compatibility rules allow both values jointly and, on DHRD, the source supports both values. The remaining columns give the realized share in percent. The three unmarked DHRD model columns are the Stockholm runs, with long family names shortened. SG and TW are Singapore and Taipei with Claude Opus 5.

	Private		DHRD					
Axis pair	n	GPT-5.6	n	Opus 5	GPT-5.6	Gemini 3.5	SG	TW
memory function × memory demand	7	100	7	100	100	100	100	100
memory function × predicate	58	100	48	100	100	100	100	100
memory function × target	158	96.8	140	96.4	95.7	91.4	98.6	98.6
memory function × condition	14	100	18	100	100	100	100	100
memory demand × predicate	263	97.3	228	97.4	97.8	85.1	97.4	97.8
memory demand × target	553	73.4	490	77.1	81.4	69.6	77.1	77.6
memory demand × condition	49	100	63	100	100	100	100	100
predicate × target	522	77.2	458	80.1	84.7	69.9	81.0	81.2
predicate × condition	294	85.7	333	85.3	85.0	71.2	82.6	84.1
target × condition	424	67.9	479	66.4	69.9	57.2	66.0	71.2
total	2,342	80.5	2,264	81.3	83.8	72.0	81.1	82.6

G.7 Scope and limits of the replication

DHRD supports transfer and reproducibility claims, not variance claims. The three cities are fixed replication environments, not a sample that could estimate between-cohort variance. The replication therefore reports the cities side by side and never pools them (Appendix G.3). The authoring-family comparison is paired and controlled, but its yield differences mix gold-answer quality with judge affinity. The judge pair is fixed across families, yet it shares model families with two of the three authors, which leaves one of them (Gemini 3.5 Flash) outside both judges' families. The same judges score every family, so the scores are comparable, but the judges may still prefer answers that resemble their own families' output. The disagreement records still argue against a single-judge bias: both judges find more to correct in Gemini 3.5 Flash than in their own families (Table 29).

Table 29: Share of each authoring family's dual-judge disagreement records in which the judge corrects the gold answer, per judge. Both judges correct Gemini 3.5 Flash more often than Claude Opus 5 and GPT-5.6.

Judge	Claude Opus 5	GPT-5.6	Gemini 3.5 Flash
GPT-5.6 economy	44%	50%	74%
Claude Sonnet 5	22%	18%	46%

The residual risk is a preference both frontier judges share, not a bias of one judge. The kin split also cannot separate such a shared preference from differences in author quality. It also ranks families only under this pipeline's contracts, gates, and configured reasoning efforts, not in general. All questions are grounded in the observed train window, days 0–75. Modifier selections, ratings, and delivery outcomes are absent from the source, so no gold answer can rest on them. Item quantity is retained but constant: each source row is one product occurrence, so quantities carry no signal. Dates and prices are synthetic benchmark values. The public-system study of §4.3 and Appendix G.8 gives the representation-level diagnostics of Appendix F.7 their public counterpart. The store exports, judged results, and per-call token ledgers are released alongside the corpus.

G.8 Public memory systems: ablations and per-axis breakdowns

This subsection reports the full results of the public-system study (§4.3). Table 30 shows the third-person rendering ablation of each approach. Only Letta changes materially. The Letta difference comes from write reliability, not from the rendering itself. The third-person ingestion left 9 of 100 memory stores below the 500-character empty-store threshold, against 1 in the first-person run. Figure 19 reports class-separated performance by memory function and memory demand. Figures 20 and 21 report the fine-grained axes. The three figures are the public counterparts of Figures 14, 15, and 16, rendered with the same code, scale, and inclusion thresholds. The slices show where each representation loses accuracy. The SQL agent is weakest on delivery-location questions. The steered

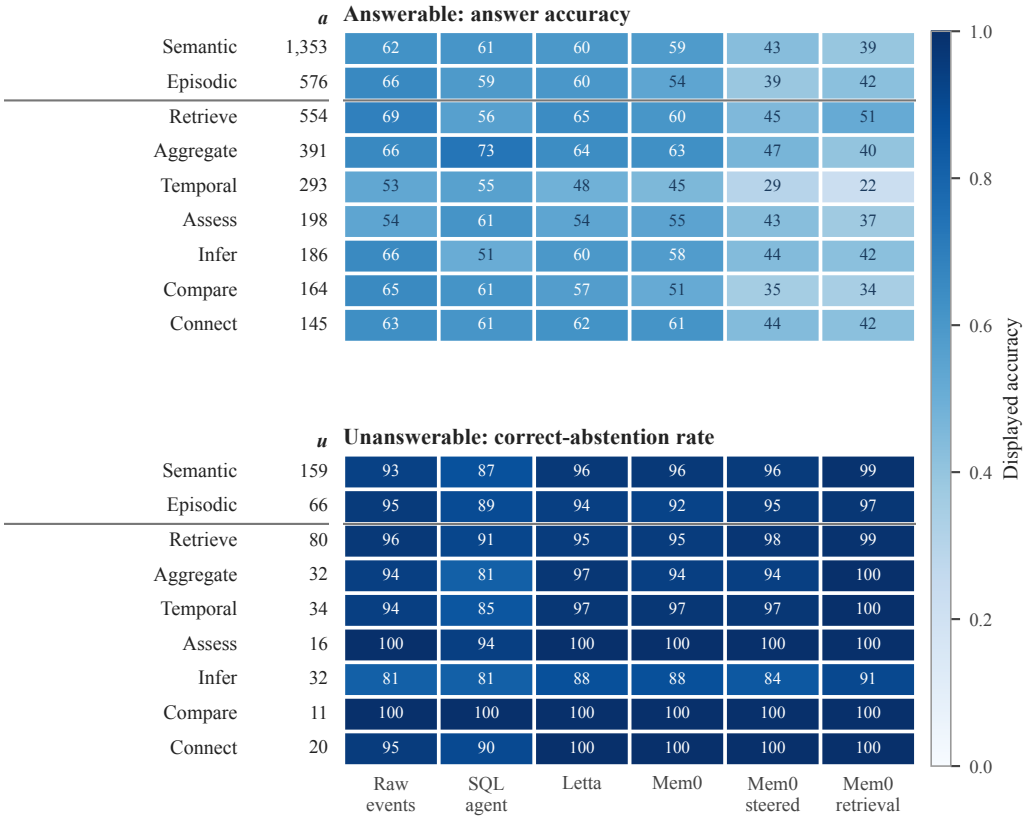

Figure 19: Class-separated performance of the six public-system configurations by realized memory function and memory demand on the PantryBench Stockholm cohort. The top panel reports answer accuracy on answerable questions. The bottom panel reports the correct-abstention rate on unanswerable questions. The a and u columns give the per-category question counts. This is the public counterpart of Figure 14.

and retrieval configurations lose most on change-over-time predicates. The SQL agent leads on spend aggregates, which it computes with a query instead of reading them from context.

Table 30: Rendering ablation of the public-system study on the PantryBench Stockholm cohort. Each memory approach re-ingests the same histories rendered in the third person instead of the first person, to test whether the first-person rendering favors any vendor. The first-person columns repeat the PantryBench panel of Table 4, whose metric definitions apply.

Memory approach	Answer [95% CI] first person	third person	Abstain [95% CI] first person	third person
Raw events	63.3 [61.0, 65.7]	62.6 [60.0, 65.2]	93.8 [90.5, 96.5]	92.9 [89.7, 95.8]
Letta	59.8 [57.0, 62.5]	55.7 [51.4, 59.7]	95.6 [92.3, 98.2]	94.2 [91.0, 97.0]
Mem0	57.1 [54.8, 59.2]	56.4 [54.1, 58.7]	95.1 [92.1, 97.6]	94.7 [91.7, 97.2]
Mem0 steered	41.8 [39.2, 44.4]	39.7 [37.1, 42.4]	95.6 [92.9, 97.9]	94.7 [91.5, 97.3]
Mem0 retrieval	40.0 [37.4, 42.5]	40.8 [38.4, 43.2]	98.2 [96.4, 99.6]	96.9 [94.5, 98.9]

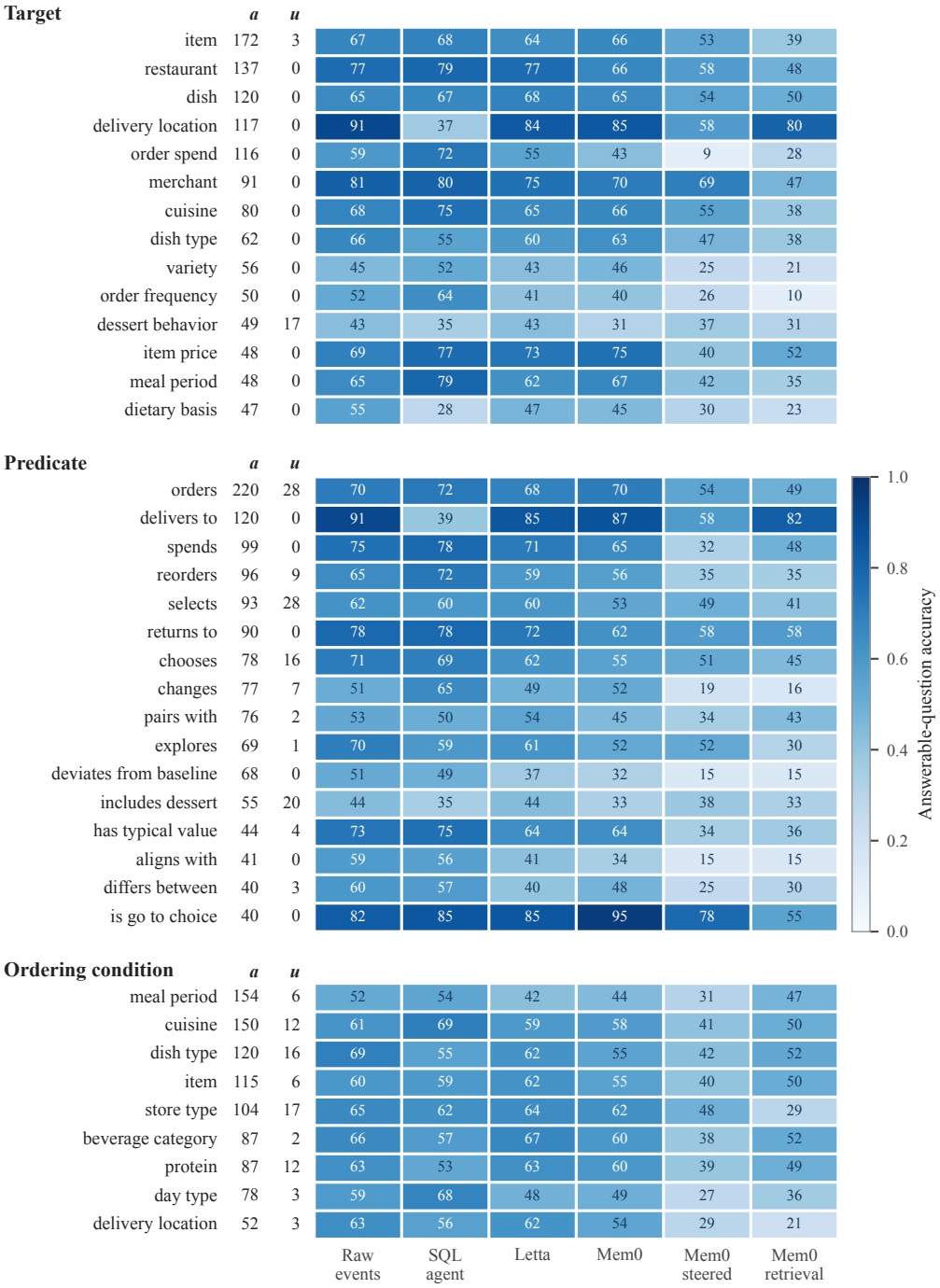

Figure 20: Answerable-question accuracy of the six public-system configurations by realized target (top, $a \geq 40$), predicate (middle, $a \geq 40$), and ordering condition (bottom, $a \geq 15$). The adjacent a and u columns give the numbers of answerable and unanswerable questions. Inclusion thresholds use a only. This is the public counterpart of Figure 15.

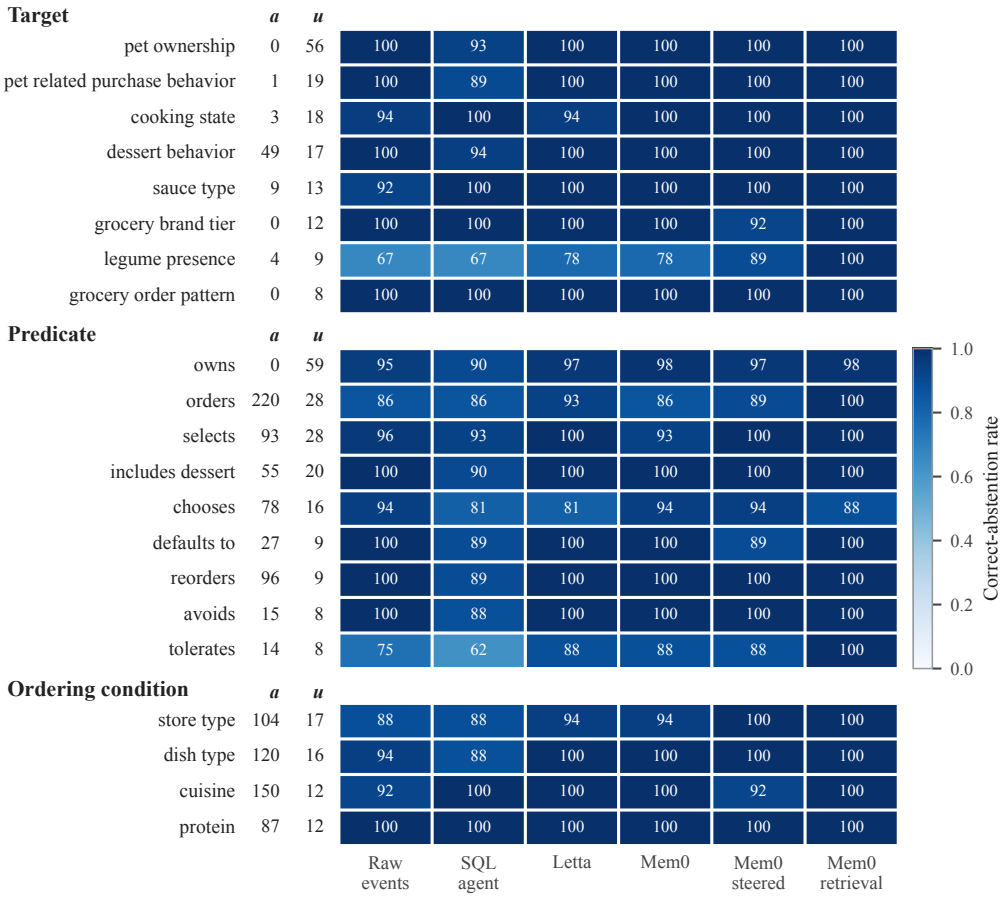

Figure 21: Correct-abstention rate of the six public-system configurations on unanswerable questions by realized target (top, $u \geq 8$), predicate (middle, $u \geq 8$), and ordering condition (bottom, $u \geq 10$). The adjacent a and u columns give the numbers of answerable and unanswerable questions. Inclusion thresholds use u only. This is the public counterpart of Figure 16.

H LIMITATIONS AND ETHICS

Private data and reproducibility. The benchmark is generated from real customer order histories and cannot be publicly released. We describe the framework, aggregate statistics, and methodology in this paper. The exact empirical results depend on private customer histories and are not currently externally reproducible. Which parts of the data, tasks, and registries can be shared is pending legal and privacy review. Until that review concludes, the paper reports aggregate results and methodology only and reveals no customer-level details. We accept the same criticism we direct at LocalSearchBench's missing retrieval components (Appendix A): grounded-but-private benchmarks trade external reproducibility for validity. The mitigations here are determinism (byte-identical reports, hashed provenance, resumable runs) and full internal archival of every run and evaluation. Any published number is therefore reproducible by the data owner. The DHRD replication (Appendix G) removes the private-data barrier for the generation-side findings. Its released form, PantryBench, passes a license and privacy review before publication. The memory-approach comparison itself still rests on the private cohort.

Snapshot scope. All memory-approach numbers are component diagnostics over exported content snapshots evaluated with one shared answer model. They measure what a fixed reader can answer from each representation, a question of information sufficiency rather than end-to-end product quality. Production retrieval, ranking, tool use, and answer agents are not exercised, and each artifact carries an explicit no-system-score contract. Readers should not cite Table 17 as a ranking of production systems.

Judge dependence. Support judging, answer grading, and axis classification are LLM-based and inherit known judge biases (Zheng et al., 2023; Koo et al., 2024; Panickssery et al., 2024). The mitigations are dual blind judges from different model families and a deterministic numeric audit. Appendix F quantifies the residual risk. Two same-family couplings remain unmitigated. The model family that drafts the questions and authors the gold answers (GPT-5.6) also produces every candidate answer. The secondary support judge is an economy tier of the same family. Either coupling may favor the family's conventions and interpretations. Prior work finds this bias small in settings like ours. Once answer quality is controlled, self-preference is small for strong judges (Spiliopoulou et al., 2025), shrinks with judge capability (Chen et al., 2025b), and is weak or absent on verifiably checkable tasks (Guey & Bougault, 2026). The proposed familiarity mechanism (Wataoka et al., 2024) predicts that the bias concentrates in open-ended stylistic grading, not in evidence-anchored verdicts. Severe amplification is documented mainly where judge output feeds training (Li et al., 2025; Xu et al., 2024). Here no judge output feeds a training loop, and the primary judge is cross-family. Verdicts are anchored to evidence-grounded contracts and a deterministic numeric audit, not to free-form preference. A planned rerun nevertheless varies exactly the authoring stages and the sampling seed, holding everything else fixed, to measure the residual effect directly.

Domain and language. The taxonomy spans food and grocery commerce, realized in English. The grammar, evidence regimes, and flywheel are domain-agnostic by construction. The DHRD replication (Appendix G) shows the generation machinery transfers to a second food-delivery platform with multilingual catalogs. Transfer beyond food commerce remains untested until the registries cover other domains.

Cohort. The cohort is 100 accounts drawn at random from those covered by the evaluated memory-system exports, with no further selection control. History density varies widely across accounts (drafted-unanswerable fractions of 5–90%). Geographic or demographic representativeness is neither controlled nor claimed. The confidence intervals account for differences among accounts in this cohort. Because we evaluated only one cohort, we cannot measure how results vary across cohorts. Controlling and varying cohort composition becomes relevant when absolute rates, rather than the framework, are the claim.

Privacy. Accounts appear only under digest pseudonyms. Prompts carry aliased row identifiers. Raw identifiers never appear in questions, answers, reasoning, or judge prose, and export validators reject leaks. In this run, 71 questions were rejected for a history alias appearing in prose. Person-level inference is scored as an error unless explicitly evidenced, which aligns the benchmark's incentives with data minimization. Governance follows documented-evaluation practice (National Institute of Standards and Technology, 2023; Gebru et al., 2021).