CypherTurn: A Multi-Turn Benchmark for Conversational Text-to-Cypher Evaluation and the Autonomy Divergence

Yuzhe Zhang[1,3] **Weijie Zhu**[2] **Haolin Yang**[1] **Ziyun Zhang**[1,3] **Xianwei Xue**[2]
Mengke Chen[2] **Qiutong Pan**[2] **Huaqian Cai**[1,3*]

[1]Peking University [2]Baidu Inc.
[3]National Key Lab of Data Space Technology and System

Abstract

Graph databases are increasingly queried through natural language, yet every existing benchmark evaluates isolated single-turn queries rather than the multi-turn sessions through which analysts actually work. We introduce **CypherTurn**, the first benchmark for conversational Text-to-Cypher evaluation, comprising 721 sessions and 5,927 turns across 7 knowledge graphs and 13 conversational phenomena. We evaluate 15 models under a guided oracle protocol and a fully autonomous agentic protocol, yielding four findings. First, the best model reaches only 64.7% execution accuracy, and session-level correctness remains below 5%. Second, despite strong overall rank correlation, frontier models exhibit a consequential reordering of the top of the leaderboard under autonomous operation, a phenomenon we term the **Autonomy Divergence**, which reveals error-management as a partially independent capability from raw generation skill. Third, scaling action budgets from $\times 3$ to $\times 10$ fails to close the autonomy gap, as the strongest frontier models self-limit to approximately two actions per turn regardless of available budget. Fourth, single-turn Cypher fine-tuning degrades multi-turn instruction following, while architecture-appropriate specialization outperforms several frontier models. These results establish CypherTurn as an open challenge for conversational graph database reasoning. Code and data are available at https://github.com/BarryQ/CypherTurn.

1 Introduction

Knowledge graphs have become foundational infrastructure in enterprise analytics, scientific discovery, and digital humanities. Deploying natural language interfaces over graph databases requires models to translate free-form user questions into Cypher queries, the declarative query language powering Neo4j and related property graph systems (Francis et al., 2018). In practice, however, users rarely retrieve the information they need from a single self-contained question. Real analytical work unfolds as a sequence of related queries, each anchored to what was returned before: a product analyst exploring a company knowledge graph might first ask which departments had the highest headcount last year, then which projects those teams worked on, then which clients each project served, and finally what revenue each client contributed. As illustrated in Figure 1, a model asked *"which students are enrolled in those courses?"* must resolve *"those courses"* against the prior turn's result set, and a follow-up *"what is their average grade?"* must further resolve *"their"* against the students just retrieved.

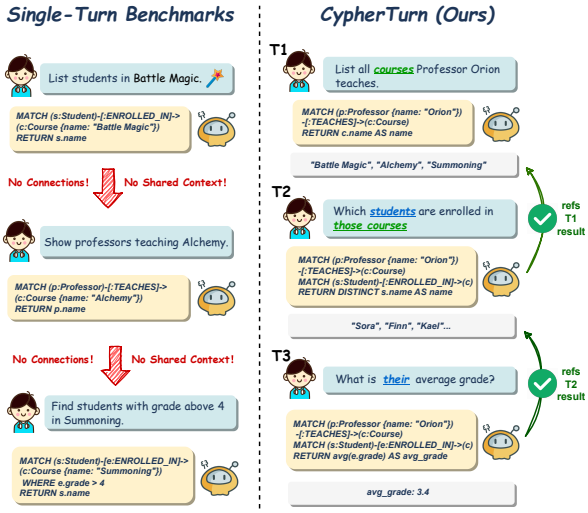


Figure 1: Existing single-turn benchmarks evaluate each query in isolation (left); CypherTurn evaluates chain-dependent turns in sequence, where each query must be grounded in its predecessor's result (right).

Existing Text-to-Cypher benchmarks do not capture this conversational structure: every published dataset in this space evaluates models on fully independent single-turn queries, and CypherBench

*Corresponding author: caihuaqian@pku.edu.cn

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 27173–27207
October 24-29, 2026 ©2026 Association for Computational Linguistics

(Feng et al., 2025), the largest and most widely used, is single-turn by design. A parallel line of work on knowledge-graph reasoning has explored LLM agents that traverse KGs interactively—e.g., Think-on-Graph (Sun et al., 2024) and Reasoning on Graphs (Luo et al., 2024)—but these systems still assume self-contained input questions and provide no protocol for multi-turn dialogue. Recent work such as PankRAG (Li et al., 2026) further models dependencies among sub-questions for graph retrieval, but remains single-query rather than conversational. The field therefore has no way to measure the capability that real deployment actually demands: coherent reasoning across a session in which each query depends on what came before. Strong single-turn accuracy reveals little about whether a model can resolve cross-turn references, propagate intermediate result sets, and stay grounded as the conversation unfolds.

The multi-turn setting also raises a second, more subtle challenge. Once a model is responsible for managing its own query history, an error at any turn corrupts the context available to every turn that follows, and small mistakes can cascade into session-wide failure. No existing benchmark measures this error propagation systematically, nor provides the infrastructure to separate pure query-generation ability from robustness to accumulated context errors.

To address both gaps, we introduce **CypherTurn**, the first multi-turn benchmark for conversational Text-to-Cypher evaluation with rigorous human verification, where each turn's Cypher must reference the result of its predecessor. We make three contributions:

- We introduce the **first multi-turn Text-to-Cypher benchmark** comprising 721 sessions, 5,927 turns, seven purpose-built knowledge graphs, a taxonomy of 13 graph-native conversational phenomena, and 9 chain modes that characterise how each query depends on its predecessor.
- We design two complementary evaluation protocols: a **Guided Protocol** that isolates query generation ability via oracle context, and an **Agentic Protocol** that requires models to manage their own multi-turn history, mirroring real deployment conditions.
- We evaluate **15 frontier and specialized models** under both protocols and identify the **Autonomy Divergence**: despite strong overall rank correlation, frontier models exhibit a consequential reordering of the top of the leaderboard under autonomous operation, revealing error-management as a partially independent capability from raw Cypher generation skill. We further find that single-turn Cypher fine-tuning actively degrades multi-turn instruction following, while architecture-appropriate domain specialization can outperform several frontier models.

2 The CypherTurn Benchmark

CypherTurn is, to our knowledge, the first Text-to-Cypher benchmark simultaneously offering multi-turn evaluation, contextual anaphora annotation, and a fully autonomous agentic evaluation protocol (Table 1).

Benchmark	MT	Cypher	Ctx	Agent	T/S
CypherBench	✗	✓	✗	✗	1.0
ZOGRASCOPE	✗	✓	✗	✗	1.0
SM3-Text-to-Query	✗	✓	✗	✗	1.0
Text2GQL-Bench	✗	✓	✗	✗	1.0
MTGQL†	✓	✗	✗	✗	6.5
CypherTurn	✓	✓	✓	✓	**8.2**

Table 1: Comparison of CypherTurn against existing Text-to-Cypher and graph query benchmarks. **MT**: multi-turn sessions; **Cypher**: Neo4j Cypher target; **Ctx**: contextual anaphora evaluation; **Agent**: agentic protocol; **T/S**: average turns per session. †MTGQL targets nGQL, not Cypher.

We construct seven unique knowledge graphs, each embodying a distinct graph topology to surface challenges specific to property graph traversal, and annotate each session with a taxonomy of 13 conversational phenomena capturing the full range of chain-dependent reasoning challenges. Every session is assigned a naturalistic persona to diversify utterance register, and all gold queries undergo a two-stage human review before inclusion.

2.1 Knowledge Graphs

CypherTurn comprises seven purpose-built knowledge graphs, each designed around a topological pattern that poses a distinct reasoning challenge. All graphs use synthetic worlds, ensuring every correct answer must be derived from the graph itself. Three graphs target structural complexity: **Ancient Empire** features a directed-cycle topology with dense numeric edge properties; **Ocean Kingdom** uses a diamond-convergence structure that creates path ambiguity; and **Magic Academy** introduces a bipartite cycle requiring precise entity resolution under structural ambiguity. Two graphs

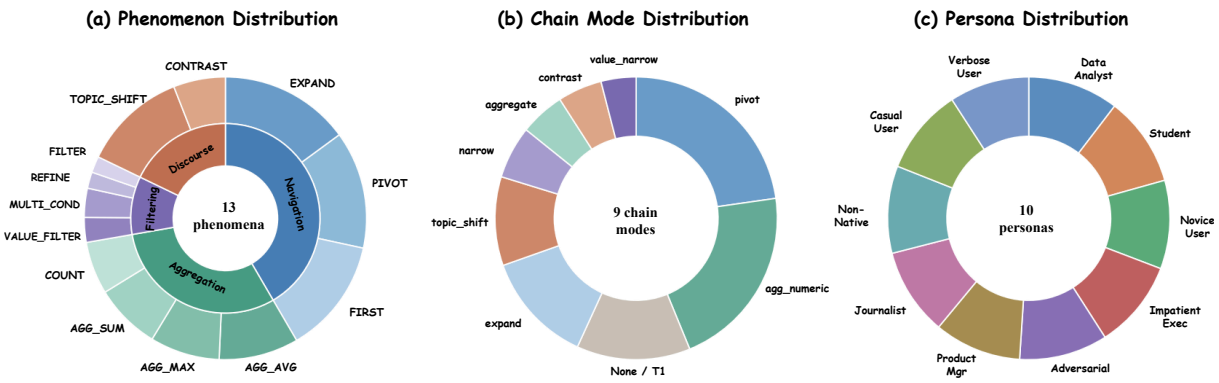


Figure 2: Benchmark composition. **(a)** Phenomenon distribution across 5,206 phenomenon-bearing turns (721 first turns carry no phenomenon; note: the FIRST phenomenon refers to rank-then-select navigation, not to first turns): inner ring shows four categories (Navigation, Aggregation, Filtering, Discourse); outer ring shows individual phenomena. **(b)** Chain mode distribution across all 5,927 turns. **(c)** Persona distribution across 721 sessions; ten types are sampled near-uniformly.

stress recursive reasoning: **Arcane Archive** embeds a citation loop enabling recursive multi-hop queries; and **Celestial Court** models a bureaucratic hierarchy with officials supervising other officials at arbitrary depth. The remaining two cover complementary patterns: **Stellar Colony** pairs a tree hierarchy with lateral alliance edges between peer factions; and **Merchant Harbor** uses a central invoice node bridging traders, buyers, and storehouses with rich numeric edge properties. Full schema details appear in Appendix D.1.

2.2 Conversational Phenomena

We design 13 distinct conversational phenomena to capture the full range of multi-turn graph querying challenges, organised into four categories that reflect different types of chain-dependent reasoning (shown in Figure 2 a). **Navigation** phenomena are the most graph-native: they require the model to traverse edges outward from the prior result set, sometimes after ranking entities by a property to identify a specific target before traversal. **Aggregation** phenomena ask the model to compute numerical summaries—means, maxima, sums, or counts—over a prior result set, testing whether entity references are correctly carried forward into aggregate queries. **Filtering** phenomena narrow prior results by adding attribute constraints, ranging from single-condition filters to compound multi-criteria conditions on numeric or categorical properties. **Discourse** phenomena test context management at the session level: one type queries the logical complement of a prior result, while the other marks a fresh independent query and is the only non-chain-dependent phenomenon in the taxonomy. Full definitions and Cypher examples appear in Appendix D.2.

2.3 Chain Modes

Annotation Example: Phenomenon vs. Chain Mode

Turn 3 (*ancient_empire*, persona: *impatient_exec*)
User: "Give me the highest population province's provinces now."
Phenomenon: PIVOT (navigation) — rank by property, then traverse
Chain mode: *pivot* — rank-then-traverse from prior result set

Turn 4 (*ancient_empire*, persona: *novice_user*)
User: "um, can I just see the ones with an occupation?"
Phenomenon: REFINE (filtering) — narrow prior result by attribute
Chain mode: *narrow* — categorical attribute filter on prior set

Both FILTER and REFINE map to chain mode *narrow*; the phenomenon distinguishes whether the filter is new or iteratively tightens a prior constraint.

Each chain-dependent turn is further characterised by a **chain mode** that specifies how the current query relates to the prior result (shown in Figure 2 b). Phenomena and chain modes form two complementary annotation layers: phenomena are fine-grained linguistic labels capturing query intent (13 types), while chain modes are coarser construction-time labels describing how a turn structurally depends on its predecessor (9 types). Multiple phenomena may map to the same chain mode. CYPHERTURN defines 9 chain modes spanning two non-chain-dependent types (*none* at T1 and *topic_shift*) and seven chain-dependent types: *expand* (one-hop traversal from the prior result), *pivot* (rank-then-traverse), *agg_numeric* (numeric aggregation over the prior set), *aggregate*

(cardinality count), *contrast* (logical complement), *value_narrow* (numeric threshold filter), and *narrow* (categorical attribute filter). Full chain mode definitions and Cypher examples appear in Appendix D.3.

2.4 Session Personas

Each session is assigned a **persona** drawn uniformly from 10 types: journalist, student, product manager, novice user, data analyst, casual user, impatient executive, adversarial user, non-native speaker, and verbose user. Personas serve as a register-diversification device that diversifies utterance style rather than as an independently validated experimental factor. Personas govern the utterance register throughout the session, and as shown in Figure 2(c) the ten types are sampled near-uniformly across the 721 sessions. The full dataset spans 5,927 turns with a mean of 8.2 turns per session. Full persona definitions and representative utterance examples appear in Appendix D.4.

2.5 Human Annotation and Quality Control

Utterances are generated programmatically from phenomenon-specific templates conditioned on persona register and prior-turn context, then optionally rewritten by DeepSeek-V3.2 for stylistic naturalness. The user simulator in the agentic protocol is also powered by DeepSeek-V3.2 with persona-specific system prompts.

All candidate sessions undergo a two-stage human review before inclusion in CypherTurn. Two Cypher-proficient annotators independently verify each turn against three criteria: (1) the gold Cypher correctly captures the utterance intent, (2) cross-turn anaphora dependencies are unambiguous, and (3) the utterance is natural and well-formed for the assigned persona. Turns marked *revise* are corrected and re-executed against a live Neo4j instance. Inter-annotator agreement on the binary accept/reject decision yields $\kappa = 0.83$ across all 5,927 turns, indicating strong agreement. Approximately 4.1% of turns required revision and 2.3% of candidate sessions were rejected outright. Every retained gold query is independently verified to execute correctly on the corresponding graph.

3 Evaluation Framework

3.1 Two Evaluation Protocols

We evaluate models under two protocols designed to measure complementary aspects of multi-turn Cypher generation (Figure 3).

Guided Protocol. At each turn, the model receives the graph schema, the complete gold-annotated history of all prior turns comprising the prior gold natural-language utterances and their gold execution results, and the current user utterance. A budget of **3 actions per turn** is enforced to encourage direct generation. This protocol provides an oracle upper bound on performance, decoupling Cypher generation skill from the ability to manage accumulated, potentially corrupted context.

Agentic Protocol. The model operates as an autonomous agent over an entire session. No oracle context is provided: the model must discover the graph schema, maintain its own prediction history, and manage a **shared action budget** of $B = T \times m$ for a session of T turns, where $m \in \{3, 5, 10\}$. Under Agentic $\times 3$ this averages three actions per turn, but the model may freely allocate more or fewer to any individual turn. When the model invokes `ASK_USER`, a persona-aware user simulator responds appropriately. This protocol mirrors the information structure of deployment, where the model must self-correct errors, plan across a long horizon, and handle context that degrades as mistakes compound.

Both protocols use the same **5-action space**: `EXECUTE_CYPHER` (run Cypher against Neo4j and receive result rows), `ASK_USER` (pose a clarifying question), `INSPECT_SCHEMA` (retrieve node labels and relation types), `SEARCH_VALUES` (fuzzy-search entity property values by keyword), and `SUBMIT_ANSWER` (finalize the predicted query for the current turn). Full protocol specifications, including prompt templates and action-space definitions, appear in Appendix D.5.

3.2 Metrics

We report metrics at both the turn level and the session level.

Execution Accuracy (EX) is our primary metric. For a predicted query $\hat{q}_t$ and gold query q_t executed over graph $\mathcal{G}$:

$$\mathrm{EX}_t = \mathbf{1}[\mathrm{exec}(\hat{q}_t, \mathcal{G}) = \mathrm{exec}(q_t, \mathcal{G})] \quad (1)$$

where equality requires identical rows *and* column names. EX is strict but unambiguous.

Provenance Subgraph Jaccard Similarity (PSJS) measures structural correctness at the graph

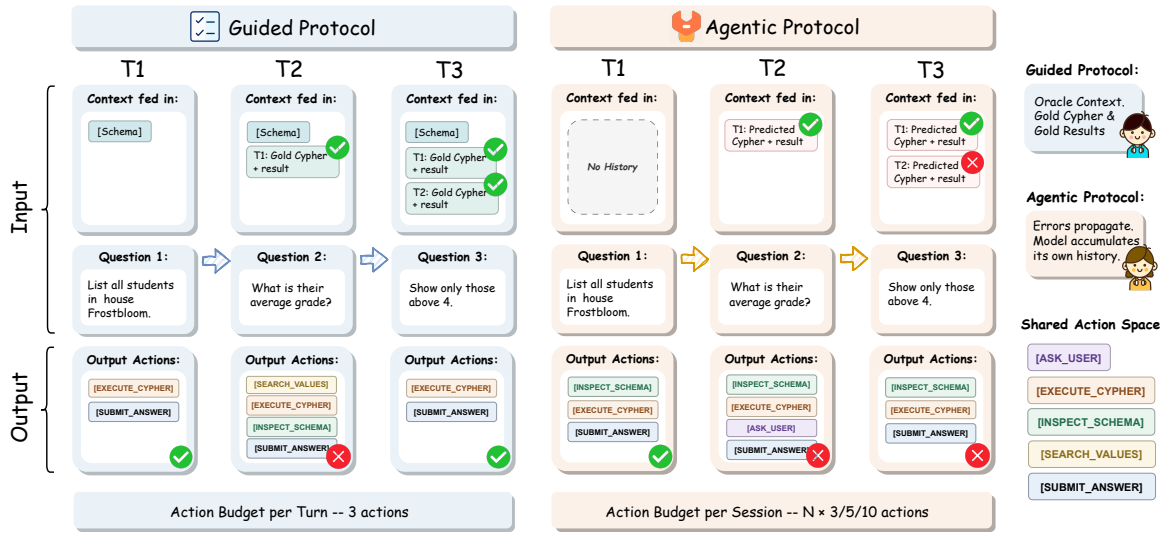


Figure 3: The two evaluation protocols in CYPHERTURN. The **Guided Protocol** (left) supplies gold-annotated prior turns as oracle context, isolating raw Cypher generation ability. The **Agentic Protocol** (right) requires models to autonomously accumulate their own history under a configurable shared session budget ($B = T \times m$); early-turn errors propagate forward through chain dependencies, compounding across the session.

traversal level:

$$\text{PSJS}_t = \frac{|V(\hat{q}_t) \cap V(q_t)|}{|V(\hat{q}_t) \cup V(q_t)|} \quad (2)$$

where $V(q_t)$ is the set of node element-IDs bound during pattern matching in q_t (for aggregation queries returning scalar values, $V(q_t)$ comprises the nodes matched by the `MATCH` clause). PSJS serves as a diagnostic for cases where a model retrieves the correct subgraph but fails at output formatting.

Chain Error Rate (CER) quantifies error propagation through session history. For all turns whose immediately preceding turn failed:

$$\text{CER} = P(\text{EX}_t = 0 \mid \text{EX}_{t-1} = 0) \quad (3)$$

A high CER indicates that failures cascade: once a turn fails, the next turn is likely to fail as well. CER is computed over all turns whose immediately preceding turn scored $\text{EX} = 0$, so the first turn of each session and turns that follow a successful turn do not contribute.

Session Exact Match (SEM) is a session-level binary metric: a session receives $\text{SEM} = 1$ only if every turn achieves $\text{EX} = 1$, and 0 otherwise. Full per-model numeric results under all three agentic budgets appear in Appendix E.

4 Experiments

4.1 Experimental Setup

We evaluate fifteen models—10 frontier API models, 2 small open-source baselines, and 3 fine-tuned or specialized architectures (full descriptions in Appendix C). All models are accessed via their respective APIs or deployed locally with identical vLLM configurations. CypherRI-7B is excluded from aggregate statistics due to fill-in-the-middle token contamination; its results appear in tables marked with †.

4.2 Main Results

Table 2 presents the full turn-level results under both protocols, and Figure 4 shows agentic EX under all three action budgets for all evaluated models. Full numeric tables appear in Appendix E.5.

Under the guided protocol, Claude Opus 4.7 leads at 0.647 EX. MiniMax-M2.7 is an outlier at 0.139, a score that likely reflects format-compliance issues. Across the remaining frontier models the spread is roughly 19 points. Session-level correctness is rare: Claude leads with an SEM of 0.046, the highest of any model, representing 33 of 721 sessions. The specialized STRuCT-LLM-Novo reaches 0.526 and surpasses six frontier models.

Every model declines under the agentic protocol except MiniMax-M2.7, whose gain of 0.018 is a floor effect that follows from its near-chance guided EX. Among frontier models, the guided-to-agentic drop ranges from 0.140 for Gemini-3.1-Flash-Lite to 0.314 for Qwen3-235B. Crucially, the rankings reorder under autonomy. Gemini-3.1-Flash-Lite rises from fourth place under the guided protocol to first under the agentic protocol at 0.432, while

Model	Guided Protocol					Agentic ×3					
	EX↑	PSJS↑	CER↓	SEM↑	Tok/S	EX↑	PSJS↑	CER↓	SEM↑	Tok/S	ΔEX↓
Frontier Models											
Claude Opus 4.7	**0.647**	**0.846**	0.434	**0.046**	617	0.402	0.681	0.712	**0.019**	639	−0.245
GPT-5.5	0.625	0.828	**0.424**	0.025	504	0.401	0.673	0.698	0.006	554	−0.224
Kimi-K2.5	0.602	0.819	0.469	0.028	753	0.335	0.612	0.749	0.006	965	−0.267
Gemini-3.1-Flash-Lite	0.572	0.744	0.508	0.017	**500**	**0.432**	**0.698**	**0.695**	0.008	**427**	**−0.140**
Qwen3-235B	0.513	0.737	0.466	0.006	896	0.199	0.548	0.833	0.001	927	−0.314
GLM-5	0.506	0.816	0.496	0.003	926	0.274	0.623	0.773	0.006	749	−0.232
DeepSeek-V3.2	0.506	0.729	0.474	0.004	894	0.204	0.551	0.835	0.000	965	−0.302
ERNIE-5.0	0.468	0.676	0.575	0.003	976	0.295	0.584	0.830	0.003	867	−0.173
MiMo-V2.5-Pro	0.460	0.752	0.531	0.003	969	0.250	0.591	0.781	0.001	871	−0.210
MiniMax-M2.7	0.139	0.379	0.856	0.000	1,644	0.156	0.403	0.862	0.000	1,805	*+0.018*
Small Open-Source Models											
Llama-3.1-8B	0.291	0.456	0.764	0.000	1,079	0.137	0.371	0.962	0.000	924	−0.154
Gemma-2-9B	0.281	0.475	0.742	0.000	1,082	0.136	0.378	0.948	0.000	844	−0.145
Fine-tuned / Specialized Models											
STRuCT-LLM-Novo	0.526	0.763	0.506	0.015	1,902	0.333	0.680	0.766	0.003	1,346	−0.193
text-to-cypher-gemma	0.261	0.498	0.759	0.000	1,054	0.143	0.369	0.934	0.000	792	−0.118
CypherRI-7B†	0.013	0.283	0.989	0.000	29,380	0.001	0.068	0.999	0.000	23,495	−0.012

Table 2: Turn-level results under the Guided and Agentic ×3 protocols. **Tok/S**: mean clean output tokens per session (excl. <think>). **Bold**: best per column (excluding †). † CypherRI-7B's near-zero EX and inflated token count both reflect a format compliance failure caused by fill-in-the-middle token contamination in its RL training; excluded from aggregate statistics.

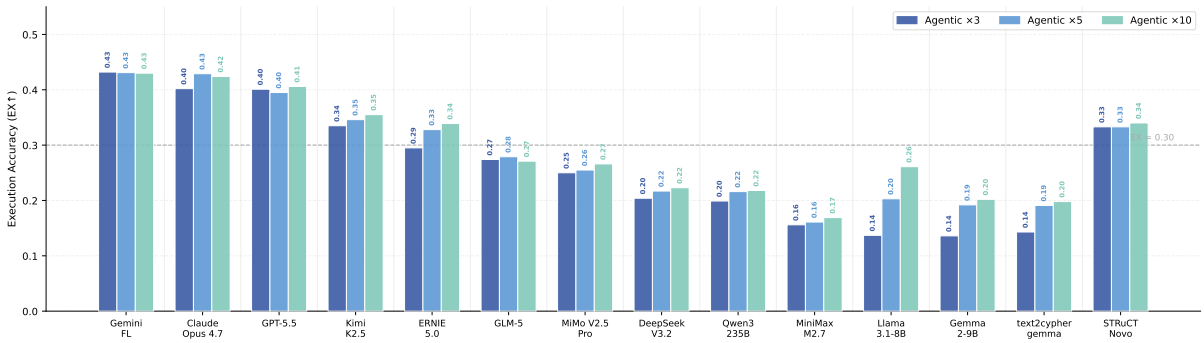


Figure 4: Agentic EX under ×3 / ×5 / ×10 action budgets for 14 evaluated models. Full numeric tables in Appendix E.5.

Qwen3-235B falls from fifth to near the bottom at 0.199. Claude Opus 4.7 and GPT-5.5 are separated by only 0.001 EX under autonomy, a difference within run-to-run variance, and the two effectively tie for second behind Gemini. All gaps and the reordering between Gemini and Claude are significant at the 95% level under a 10,000-resample bootstrap, and we analyze this reordering, which we term the Autonomy Divergence, in §5.2. Expanding the budget from ×3 to ×10 yields limited frontier gains but benefits small models substantially, as Llama-3.1-8B nearly doubles to 0.261 in Figure 4.

5 Analysis

5.1 The Difficulty Landscape

Figure 5 presents guided EX across three dimensions. Phenomenon type is the strongest predictor of turn difficulty: REFINE averages 0.244 across 14 models and CONTRAST averages 0.273, while AGG_MAX averages 0.650 and AGG_AVG 0.602. Persona and graph topology are secondary factors, with within-model EX spread across personas typically ten to sixteen points for frontier models. Per-graph difficulty tracks structural entity-membership ambiguity: Stellar Colony and Merchant Harbor are hardest at 0.426 and 0.434, while Celestial Court reaches 0.485 (Table 10).

Figure 5: Guided EX across three dimensions for 14 models (CypherRI-7B excluded; MiniMax-M2.7 additionally excluded from heatmaps). **(a)** Per-phenomenon mean EX with min–max range; dashed line shows the unweighted macro mean across 13 phenomena. **(b)** Per-persona heatmap (13 models $\times$ 10 personas). **(c)** Per-graph heatmap (13 models $\times$ 7 graphs).

A distinct source of difficulty also emerges on the output side: the gap between PSJS and EX. Claude Opus 4.7 leads PSJS at 0.846 yet achieves only 0.647 EX, and its SEM of 0.046 represents just 33 fully correct sessions out of 721. Output formatting is independent of traversal correctness and remains unsolved.

5.2 The Autonomy Divergence: Why Rankings Shift

Guided and agentic rankings are strongly correlated overall, with a Spearman ρ of 0.86 across all 14 models. This global figure is propped up by the weakest models, which sit stably at the bottom under both protocols. Restricted to the ten frontier models, ρ drops to 0.73, and to 0.63 (95% CI [0.58, 0.70]) after excluding MiniMax-M2.7. Within the frontier tier, 11 of the 45 pairwise comparisons reverse significantly under a 10,000-resample bootstrap, including Gemini overtaking Claude for the top position. We therefore frame the Autonomy Divergence as a differential-robustness effect rather than a systemic reversal: autonomy disproportionately degrades a few frontier models, enough to change the top of the leaderboard, while the overall order holds. The Autonomy Divergence isolates a graph-specific mechanism in which each turn's entity set seeds the next, so that an early error changes what later queries compute over rather than merely adding independent noise. We further test three protocol variants in Appendix F, and all preserve the overall ranking.

CER quantifies the cascade. Even under the guided protocol, where gold context insulates each turn from error propagation, frontier models show a median conditional CER of 0.47, excluding MiniMax-M2.7 (computed from the model's own per-turn EX, not the injected oracle context), indicating that chain-dependent turns cluster their failures intrinsically. Under the agentic protocol this figure rises to 0.77. For DeepSeek-V3.2 and Qwen3-235B it exceeds 0.83, which means that once a turn fails, the next one fails more than 83% of the time.

Resilience tracks a schema-heavy action strategy, which we interpret as a behavioral marker of resilient models rather than the cause of resilience. The resilient models recalibrate graph structure on every turn, and Claude and Gemini-3.1-Flash-Lite devote 85.8% and 83.5% of their actions to `INSPECT_SCHEMA`. The collapsing models instead spend 62–73% of their actions on `EXECUTE_CYPHER` retries without recovering the schema, which leaves them exposed to compounding corruption. As Table 3 shows, `INSPECT_SCHEMA` allocation correlates with agentic EX at a Spearman ρ of 0.93 ($p<0.001$), stable above 0.88 under leave-one-out removal. Protocol variants that make the schema persistent or disclose the session horizon do not eliminate the gap, as confirmed by the ablations in Appendix F.

Figure 6 makes the divergence visible turn by turn. Under the guided protocol (panel a), all six frontier models remain stable across T_1–T_{10}, confirming that oracle context quarantines per-turn errors. The agentic trajectories then split into two trends. Claude and Gemini-3.1-Flash-Lite plateau without collapsing (panel b); GLM-5 dips at T_3 but recovers to roughly 0.44 rather than continuing to decline. ERNIE-5.0, DeepSeek-V3.2, and Qwen3-

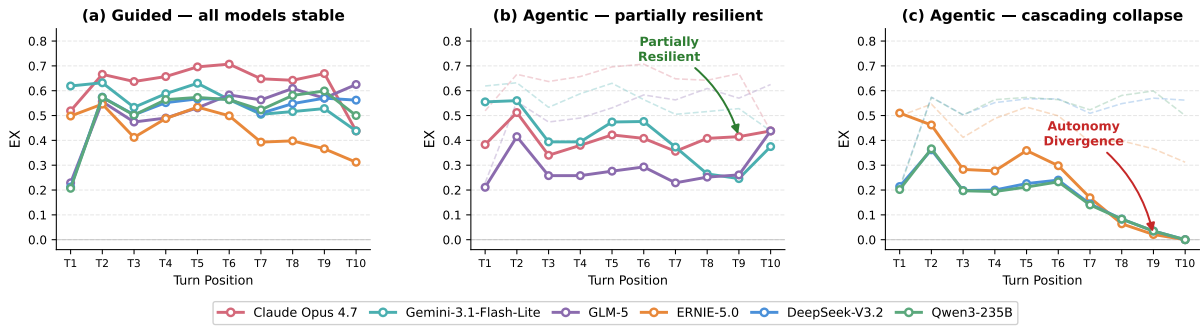


Figure 6: EX by turn position for six frontier models. **(a)** Guided protocol: all models remain stable across turns. **(b)** Agentic protocol: Claude, Gemini, and GLM-5 plateau at non-trivial EX (dashed lines show their guided baselines). **(c)** Agentic protocol: ERNIE, DeepSeek, and Qwen3 collapse to near-zero by T_{10}.

Model	INSPECT%	Agentic EX	ΔEX
GPT-5.5	94.1	0.401	0.224
Claude Opus 4.7	85.8	0.402	0.245
Gemini-3.1-FL	83.5	0.432	0.140
GLM-5	54.3	0.274	0.232
STRuCT-LLM-Novo	40.4	0.333	0.193
Kimi-K2.5	34.8	0.335	0.267
MiMo-V2.5-Pro	32.5	0.250	0.210
ERNIE-5.0	29.3	0.295	0.173
Qwen3-235B	29.2	0.199	0.314
DeepSeek-V3.2	28.9	0.204	0.302
MiniMax-M2.7	23.8	0.156	—
Gemma-2-9B	23.1	0.136	0.145
t2c-gemma	23.1	0.143	0.118
Llama-3.1-8B	15.9	0.137	0.154

Table 3: INSPECT_SCHEMA fraction vs. Agentic ×3 EX, sorted by INSPECT%. Green : schema-heavy; yellow : mixed; red : schema-sparse.

235B trace a different curve (panel c): each rises to 0.36–0.51 by T_2 before degrading almost monotonically to near-zero EX by T_{10}. The same chain-dependent failures that oracle context absorbs become session-wide once the model carries its own history.

5.3 The Budget Ceiling: Why Extra Actions Don't Help

If cascading errors drive the autonomy gap, more actions to recover should help. Figure 4 shows otherwise. Gemini-3.1-Flash-Lite achieves 0.432, 0.431, and 0.430 across ×3, ×5, and ×10, remaining flat despite a 3.3-fold budget increase. Claude peaks at ×5 with 0.429 and GLM-5 actually degrades at ×10 to 0.271. Qwen3-235B gains only 1.9 points across the full range, negligible against its 31-point guided gap. Small open-source models do benefit from extra budget (Llama-3.1-8B nearly doubles to 0.261), confirming that the budget-scaling curve separates exploration-limited failures from strategic-deficiency failures.

Even at ×10, no model closes its own guided-to-agentic gap: Gemini-3.1-Flash-Lite remains 0.142 points below its guided EX, and Claude 0.223 points below. The schema-heavy frontier models self-limit to approximately two actions per turn regardless of available budget, producing repeated ineffective retries rather than productive exploration; the collapsing frontier models spend three to four actions per turn, yet almost all of it on repeated retries that fail to localize the true error.

A natural concern is that this self-limitation reflects the efficiency instruction in the action prompt rather than a genuine capability gap. We test this by replacing the efficiency instruction with explicit encouragement to use the full budget, run at tenfold budget. Execution accuracy drops for six of the seven valid frontier models, because the extra actions are largely repeated retries and unfocused searches that introduce noise rather than localizing the true error. The bottleneck is ineffective error recovery, not unwillingness to spend budget (Appendix F).

The hard budget cliff contributes little to the frontier gap: excluding turns auto-scored zero from budget exhaustion shifts scores by at most 3.3 points for the four models anchoring the main analysis, and the largest adjustment across the frontier is 6.5 points for ERNIE-5.0. The cliff dominates only the small models, where exhaustion rates exceed 80%, and it adjusts absolute scores without reordering the frontier.

5.4 The Specialization Paradox

Single-turn Cypher fine-tuning degrades multi-turn performance. text-to-cypher-gemma achieves a guided EX of only 0.261, which falls below the 0.281 of its base model Gemma-2-9B. The Neo4j Text2Cypher corpus implicitly teaches that every query is self-contained, and this bias breaks anaphora resolution and result-set chaining in

Model	Paradigm	Guided	Agentic
Gemma-2-9B	Base	0.281	0.136
text-to-cypher-gemma	Single-turn SFT	0.261	0.143
QwQ-32B‡	Base	0.395	0.324
STRuCT-LLM-Novo	RL + CoT	0.526	0.333

Table 4: Two controlled base-to-specialised pairs. Single-turn SFT degrades the Gemma pair, while RL with chain-of-thought improves the QwQ pair, in opposite directions. ‡ QwQ-32B is evaluated on the stratified 210-session subset of Appendix F, whose per-model EX deviates from the full set by at most 0.01; the 13.1-point gain far exceeds this subset noise.

multi-turn settings. The fine-tuned model repeatedly hardcodes prior results as literal value lists instead of maintaining relational continuity.

Two controlled base-to-specialised pairs, summarised in Table 4, test whether the decisive factor is the training paradigm or the domain specificity of the training data. The Gemma pair moves downward: single-turn SFT lowers guided EX from 0.281 to 0.261. The QwQ pair moves in the opposite direction: STRuCT-LLM-Novo, built on QwQ-32B (a 32B open-weight reasoning model from Alibaba's Qwen team), uses RL with chain-of-thought supervision and cross-formalism transfer between SQL and Cypher to raise guided EX from 0.395 to 0.526, a gain of 13.1 points. The same direction holds under autonomy, where QwQ-32B rises from an agentic EX of 0.324 to 0.333. At the opposite extreme, CypherRI-7B suffers from RL-induced format tokens that contaminate its output. What matters is therefore not whether the training data is domain-specific but whether the training paradigm matches the multi-turn nature of the task.

6 Conclusion

We introduced CypherTurn, the first multi-turn benchmark for conversational Text-to-Cypher evaluation, comprising 721 sessions and 5,927 gold-verified queries across seven knowledge graphs and 13 phenomena. Evaluation of fifteen models under guided and agentic protocols yields four findings. Performance ceilings are low: even the best model reaches only 64.7% guided EX, and session-level correctness barely exceeds 4%. The Autonomy Divergence shows that error-management under autonomous operation is partially independent of generation skill, producing a reordering of the top of the leaderboard invisible in oracle-context evaluation. Extra action budget does not close this gap, as frontier models self-limit regardless of resources. Finally, single-turn fine-tuning degrades multi-turn performance, while architecture-appropriate specialisation outperforms several frontier models. The uneven degradation and the instability of the top rank persist under three protocol variants that perturb schema availability, horizon disclosure, and prompt wording; each variant displaces Gemini-3.1-Flash-Lite from the top position, so the Autonomy Divergence is robust to the design choices of the agentic protocol even though the identity of the leader is not. These findings position CypherTurn as an open challenge and motivate multi-turn training corpora, schema-recovery mechanisms, and finer-grained semantic equivalence metrics as next steps.

Limitations

CypherTurn evaluates 15 models spanning frontier and specialized architectures, yet does not exhaust all competitive systems available at submission time; results for unevaluated frontier models may differ, and the rapid pace of model releases means that newer systems could shift the absolute scores reported here. All seven knowledge graphs are synthetic, so execution accuracy on CypherTurn does not extrapolate to deployment accuracy on real enterprise graphs, where schema naming conventions, incomplete data, and domain-specific ambiguity introduce additional difficulty; the benchmark measures multi-turn reasoning under controlled topology rather than ecological validity, and a pilot on real-world graphs remains future work. All sessions are conducted in English, so the extent to which the Autonomy Divergence persists across typologically different languages remains an open question. Finally, Execution Accuracy requires result-set identity including column names, which may penalize semantically equivalent queries that differ only in output formatting or column ordering; a finer-grained semantic equivalence metric that tolerates such surface variation would provide a more forgiving upper bound and remains future work.

References

Francesco Cazzaro, Justin Kleindienst, Sofia Márquez Gomez, and Ariadna Quattoni. 2025. ZOGRASCOPE: A new benchmark for semantic parsing over property graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4239–4246, Suzhou, China. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yanlin Feng, Simone Papicchio, and Sajjadur Rahman. 2025. CypherBench: Towards precise retrieval over full-scale modern knowledge graphs in the LLM era. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8934–8958, Vienna, Austria. Association for Computational Linguistics.

Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1433–1445, New York, NY, USA. Association for Computing Machinery.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pages 57734–57811.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aibo Guo, Xinyi Li, Guanchen Xiao, Zhen Tan, and Xiang Zhao. 2022. Spcql: A semantic parsing dataset for converting natural language into cypher. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3973–3977, New York, NY, USA. Association for Computing Machinery.

Jiaqi Guo, Ziliang Si, Yu Wang, Qian Liu, Ming Fan, Jian-Guang Lou, Zijiang Yang, and Ting Liu. 2021. Chase: A large-scale and pragmatic Chinese dataset for cross-database context-dependent text-to-SQL. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2316–2331, Online. Association for Computational Linguistics.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models?

Nan Huo, Xiaohan Xu, Jinyang Li, Per Jacobsson, Shipei Lin, Bowen Qin, Binyuan Hui, Xiaolong Li, Ge Qu, Shuzheng Si, Linheng Han, Edward Alexander, Xintong Zhu, Rui Qin, Ruihan Yu, Yiyao Jin, Feige Zhou, Weihao Zhong, Yun Chen, and 5 others. 2026. Bird-interact: Re-imagining text-to-sql evaluation for large language models via lens of dynamic interactions. *Preprint*, arXiv:2510.05318.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pages 54107–54157.

Catherine Kosten, Philippe Cudré-Mauroux, and Kurt Stockinger. 2023. Spider4sparql: A complex benchmark for evaluating knowledge graph question answering systems. In *2023 IEEE International Conference on Big Data (BigData)*, page 5272–5281. IEEE.

Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin SU, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. 2025. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. In *International Conference on Learning Representations*, volume 2025, pages 28691–28735.

Haoyang Li, Jing Zhang, Hanbing Liu, Ju Fan, Xiaokang Zhang, Jun Zhu, Renjie Wei, Hongyan Pan, Cuiping Li, and Hong Chen. 2024. Codes: Towards building open-source language models for text-to-sql. *Proc. ACM Manag. Data*, 2(3).

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. In *Advances in Neural Information Processing Systems*, volume 36, pages 42330–42357. Curran Associates, Inc.

Ningyuan Li, Junrui Liu, Yi Shan, Minghui Huang, Ziren Gong, and Tong Li. 2026. Pankrag: Enhancing graph retrieval via globally aware query resolution and dependency-aware reranking mechanism. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19157–19161. IEEE.

Yuanyuan Liang, Lei Pan, Tingyu Xie, Yunshi Lan, and Weining Qian. 2025. Multi-turn natural language to graph query language translation. *Preprint*, arXiv:2508.01871.

Shicheng Liu, Sina Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, and Monica Lam. 2024a. SPINACH: SPARQL-based information navigation for challenging real-world questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15977–16001, Miami, Florida, USA. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024b. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pages 52989–53046.

Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, volume 2024, pages 14400–14423.

Songlin Lyu, Lujie Ban, Zihang Wu, Tianqi Luo, Jirong Liu, Chenhao Ma, Yuyu Luo, Nan Tang, Shipeng Qi, Heng Lin, Yongchao Liu, and Chuntao Hong. 2026. Text2gql-bench: A text to graph query language benchmark [experiment, analysis & benchmark]. *Preprint*, arXiv:2602.11745.

Makbule Gulcin Ozsoy, Leila Messallem, Jon Besga, and Gianandrea Minneci. 2025. Text2Cypher: Bridging natural language and graph databases. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 100–108, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Mohammadreza Pourreza and Davood Rafiei. 2023. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. In *Advances in Neural Information Processing Systems*, volume 36, pages 36339–36348. Curran Associates, Inc.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Sithursan Sivasubramaniam, Cedric Osei-Akoto, Yi Zhang, Kurt Stockinger, and Jonathan Fürst. 2024. Sm3-text-to-query: Synthetic multi-model medical text-to-query benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 88627–88663. Curran Associates, Inc.

Josefa Lia Stoisser, Marc Boubnovski Martell, Lawrence Phillips, Casper Hansen, and Julien Fauqueur. 2025. Struct-llm: Unifying tabular and graph reasoning with reinforcement learning for semantic parsing. *Preprint*, arXiv:2506.21575.

Hanchen Su, Xuyuan Li, Yan Zhou, zhuoyi lu, Ziwei Chai, Haozheng Wang, Chen Zhang, and Yang Yang. 2026. Cypher-RI: Reinforcement learning for integrating schema selection into cypher generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *Preprint*, arXiv:2307.07697.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, and 107 others. 2026a. Mimo-v2-flash technical report. *Preprint*, arXiv:2601.02780.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026b. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Aman Tiwari, Shiva Krishna Reddy Malay, Vikas Yadav, Masoud Hashemi, and Sathwik Tejaswi Madhusudhan. 2025. Auto-cypher: Improving LLMs on cypher generation via LLM-supervised generation-verification framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 623–640, Albuquerque, New Mexico. Association for Computational Linguistics.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, Bangkok, Thailand. Association for Computational Linguistics.

Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He,

Dou Hong, Qiwen Liu, Shuohuan Wang, Junyuan Shang, Zhenyu Zhang, Yuchen Ding, Jinle Zeng, Jiabin Yang, Liang Shen, Ruibiao Chen, and 419 others. 2026. Ernie 5.0 technical report. *Preprint*, arXiv:2602.04705.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In *International Conference on Learning Representations*, volume 2024, pages 32593–32627.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pages 52040–52094. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2025. τ-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *International Conference on Learning Representations*, volume 2025, pages 9965–10017.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, and 5 others. 2019a. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. 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)*, pages 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019b. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, Florence, Italy. Association for Computational Linguistics.

Ziyu Zhao, Wei Liu, Tim French, and Michael Stewart. 2023. Cyspider: A neural semantic parsing corpus with baseline models for property graphs. In *AI 2023: Advances in Artificial Intelligence: 36th Australasian Joint Conference on Artificial Intelligence, AI 2023, Brisbane, QLD, Australia, November 28–December 1, 2023, Proceedings, Part II*, page 120–132, Berlin, Heidelberg. Springer-Verlag.

Zijie Zhong, Linqing Zhong, Zhaoze Sun, Qingyun Jin, Zengchang Qin, and Xiaofan Zhang. 2025. SyntheT2C: Generating synthetic data for fine-tuning large language models on the Text2Cypher task. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 672–692, Abu Dhabi, UAE. Association for Computational Linguistics.

Appendix Contents

A Related Work 13
B Comparison Benchmarks 14
C Model Descriptions 14
D Benchmark Reference 16
D.1 Knowledge Graph Schemas 16
D.2 Conversational Phenomena 16
D.3 Chain Modes 16
D.4 Session Personas 17
D.5 Evaluation Protocols 18
E Detailed Experimental Results 21
E.1 Per-Graph Results 21
E.2 Per-Phenomenon Results 21
E.3 Per-Chain-Mode Results 21
E.4 Per-Persona Results 22
E.5 Agentic Budget Scaling 24
F Protocol Robustness 25
F.1 Variant Design and Overall Effect 26
F.2 Gap Persistence and Budget Cliff 27
G Case Studies 28
H Evaluation Prompts 34

A Related Work

Text-to-SQL benchmarks. The multi-turn and agentic evaluation paradigms that motivate CypherTurn originate in the SQL community. SParC (Yu et al., 2019b) and CoSQL (Yu et al., 2019a) established conversational Text-to-SQL evaluation on relational databases, demonstrating that cross-turn context dependence materially increases difficulty; CHASE (Guo et al., 2021) extended this paradigm to Chinese with more challenging context-dependent splits. On the single-turn side, BIRD (Li et al., 2023) introduced large-scale databases with dirty values and external-knowledge grounding, while Spider 2.0 (Lei et al., 2025) further extended evaluation to real enterprise workflows across heterogeneous SQL dialects. Methodologically, DIN-SQL (Pourreza and Rafiei, 2023) showed that decomposing generation into schema linking and self-correction substantially improves accuracy, and CodeS (Li et al., 2024) demonstrated that open-source models can rival proprietary LLMs with carefully curated training corpora. BIRD-INTERACT (Huo et al., 2026) further showed that agentic evaluation reveals failure modes entirely invisible in oracle-context settings. These findings motivate our dual-protocol design, extended here to the property graph domain where no equivalent evaluation framework previously existed.

Text-to-Cypher and Graph Query Benchmarks. Cypher, the declarative query language for property graphs (Francis et al., 2018), has received comparatively little NLI attention. CypherBench (Feng et al., 2025) is the most comprehensive existing Text-to-Cypher benchmark, covering 11 large-scale knowledge graphs derived from Wikidata with over 10,000 single-turn questions. ZOGRASCOPE (Cazzaro et al., 2025) provides a human-annotated Cypher benchmark over a single domain graph with compositional and length-generalization splits. SM3-Text-to-Query (Sivasubramaniam et al., 2024) evaluates Cypher generation in the biomedical domain alongside SQL, MQL, and SPARQL. Text2GQL-Bench (Lyu et al., 2026) scales to 178K question–query pairs across 34 graph databases, supporting both Cypher and ISO-GQL. Earlier work also includes SpCQL (Guo et al., 2022), a Chinese dataset of 10,000 NL–Cypher pairs over a single Neo4j graph that established early baselines for the task; CySpider (Zhao et al., 2023), an English corpus of Cypher queries derived from Spider SQL via algorithmic translation; and Text2Cypher (Ozsoy et al., 2025), a community-compiled English dataset of 44,387 (NL, Cypher) instances aggregated from public sources. To address the scarcity of Cypher training data, SyntheT2C (Zhong et al., 2025) and SynthCypher (Tiwari et al., 2025) propose synthetic NL–Cypher generation pipelines that produce verified executable queries, though both retain a strictly single-turn formulation. Adjacent to property-graph Cypher, Spider4SPARQL (Kosten et al., 2023) provides a compositional SPARQL benchmark over 166 RDF knowledge graphs, and SPINACH (Liu et al., 2024a) introduces an agentic SPARQL navigation benchmark for challenging real-world questions; both target RDF/SPARQL rather than Cypher and remain single-turn at the question level. All of these datasets share the limitation of providing no genuine conversational context or chain-dependent anaphora annotation. MTGQL (Liang et al., 2025) is the only prior multi-turn graph query benchmark, targeting nGQL (NebulaGraph) for financial market queries. Despite this recent progress, no existing benchmark combines multi-turn evaluation, contextual anaphora tracking, and an agentic protocol within the Cypher ecosystem.

Interactive and Agentic Evaluation. BIRD-INTERACT (Huo et al., 2026) demonstrated that

evaluating LLMs as interactive agents for Text-to-SQL reveals failure modes entirely invisible in oracle-context evaluation, and MINT (Wang et al., 2024) similarly shows that multi-step tool use introduces compounding errors absent in single-step benchmarks. A broader line of agentic benchmarks—AgentBench (Liu et al., 2024b) across eight environments, τ-bench (Yao et al., 2025) for tool-agent-user interaction with simulated users, AppWorld (Trivedi et al., 2024) for interactive coding over 9 apps and 457 APIs, OSWorld (Xie et al., 2024) for real computer-use tasks, and SWE-bench (Jimenez et al., 2024) for resolving real-world GitHub issues—has consistently surfaced a substantial gap between oracle and autonomous performance. These environments build on prompting paradigms such as ReAct (Yao et al., 2023), which interleaves reasoning and action, and self-correction methods such as Reflexion (Shinn et al., 2023) and CRITIC (Gou et al., 2024), which leverage verbal reinforcement and tool-interactive critique to recover from intermediate failures. Orthogonally, RULER (Hsieh et al., 2024) shows that long-context models degrade sharply as input length grows, an effect directly relevant to multi-turn agentic settings where prediction history accumulates over a session. CypherTurn departs from this prior work by targeting Cypher for property graphs, introducing chain-dependent turn dependencies specific to graph traversal, and providing CER—a quantitative metric for cross-turn error propagation that has no counterpart in prior work.

B Comparison Benchmarks

We briefly describe the five benchmarks compared against CypherTurn in Table 1.

CypherBench (Feng et al., 2025) is a Text-to-Cypher benchmark covering 11 large-scale multi-domain property graphs derived from Wikidata, comprising 7.8 million entities and over 10,000 questions. Addressing the challenge that raw RDF graphs exceed LLM context windows, CypherBench introduces property graph views queryable via Cypher, an RDF-to-property-graph conversion engine, and a task generation pipeline. All questions are single-turn; there is no conversational structure or agentic evaluation.

ZOGRASCOPE (Cazzaro et al., 2025) is a human-annotated Text-to-Cypher benchmark over a single-domain crime-investigation property graph (61.5K nodes, 105.8K edges, 11 entity classes) hosted on Neo4j. It provides approximately 5,022 samples (2,905 training, 767 IID test, and 1,350 compositional test) split to probe structural and length generalization. All questions are single-turn.

SM3-Text-to-Query (Sivasubramaniam et al., 2024) is a NeurIPS 2024 multi-model medical benchmark comprising approximately 40,000 question–query pairs across four query languages (Cypher, SQL, MQL, SPARQL), built on synthetic patient data using the SNOMED-CT taxonomy. The Cypher subset covers queries over a Neo4j graph. All questions are single-turn; domain coverage is restricted to biomedical data.

Text2GQL-Bench (Lyu et al., 2026) is the largest existing graph query benchmark at 178,184 (NL, Query) pairs across 34 graph databases spanning 13 domains, supporting both Cypher and ISO-GQL. Questions are available at three abstraction levels (Syntactic, Logical, and Business), in addition to the original question. All questions are single-turn; no multi-turn or agentic evaluation is provided.

MTGQL (Liang et al., 2025) is, to our knowledge, the only existing multi-turn graph query language benchmark, targeting nGQL (NebulaGraph) for financial market graph queries. It uses an LLM-based automated construction pipeline and addresses session-level context dependencies. MTGQL does not use Cypher, does not include contextual anaphora annotation, and provides no agentic evaluation protocol.

C Model Descriptions

We evaluate fifteen models spanning three tiers on the full benchmark. QwQ-32B, listed last within the fine-tuned tier as the base model of STRuCT-LLM-Novo, is additionally evaluated on the stratified 210-session subset only; no separate citation is provided for it, since its results serve only as the base reference of the STRuCT-LLM-Novo pair. Citation information is provided in the bibliography for those models with published technical reports; for proprietary API checkpoints without one, the entry describes the release and access route.

• Frontier Models.

Claude Opus 4.7 is Anthropic's flagship model, released in April 2026. It is designed for advanced coding, long-horizon agentic tasks, and complex multi-tool workflows, with a 1M token context win-

dow. Its parameter count is not publicly disclosed.

GPT-5.5 is a frontier model from OpenAI released in April 2026, positioned as the successor to GPT-5.4 in the GPT-5 series. It features approximately 1 million token context window and is optimised for agentic coding, computer use, and knowledge work. No arXiv technical report has been published for this model.

Kimi-K2.5 (Team et al., 2026b) is a Mixture-of-Experts (MoE) model from Moonshot AI with 1 trillion total parameters and 32 billion activated per token. Building on Kimi K2 (pre-trained on 15.5 trillion tokens with the MuonClip optimiser), K2.5 extends the family to a 256K token context with native multimodal (vision and video) capabilities. The model is designed for agentic tool use and multi-step reasoning, with open weights released under a modified MIT licence.

Gemini-3.1-Flash-Lite is a member of Google DeepMind's Gemini 3 family, optimised for high-throughput, cost-efficient, and low-latency applications. It accepts text, image, audio, and video inputs with a context window of up to 1 million tokens, and is positioned as the most cost-efficient model in the Gemini 3 generation.

Qwen3-235B (Yang et al., 2025) is the flagship Qwen3-235B-A22B MoE model from Alibaba Cloud's Qwen Team, with 235 billion total parameters and 22 billion activated per token, released in May 2025. It introduces a unified thinking/non-thinking mode with an adaptive thinking-budget mechanism, and supports 119 languages and dialects. The model is released under the Apache 2.0 licence.

GLM-5 (GLM-5-Team et al., 2026) is the next-generation foundation model from Zhipu AI / Z.ai, released in February 2026. It is a Mixture-of-Experts model with 744 billion total parameters and approximately 40 billion activated per token, pre-trained on 28.5 trillion tokens. GLM-5 integrates DeepSeek Sparse Attention (DSA) to support a 200K token context window while reducing inference cost, and is designed for agentic engineering and long-horizon software development tasks. The weights are released under the MIT licence.

DeepSeek-V3.2 (DeepSeek-AI et al., 2025) is the latest model in the DeepSeek-V3 family, with 671 billion total parameters and 37 billion activated per token. Inheriting the architecture of DeepSeek-V3 (Multi-head Latent Attention, an auxiliary-loss-free load-balancing strategy, and multi-token prediction, pre-trained on 14.8 trillion tokens), V3.2 introduces DeepSeek Sparse Attention (DSA) through continued pre-training, achieving near-linear $O(kL)$ long-context attention cost while maintaining output quality comparable to the dense V3.1-Terminus baseline.

ERNIE-5.0 (Wang et al., 2026) is a large-scale, natively omni-modal language model from Baidu, part of the ERNIE (Enhanced Representation through kNowledge IntEgration) series. ERNIE 5.0 trains text, image, audio, and video jointly under a unified autoregressive Mixture-of-Experts (MoE) architecture, with approximately 2.4 trillion total parameters and fewer than 3% (roughly 72B) activated per inference, and a 128K context window.

MiMo-V2.5-Pro (Team et al., 2026a) is the flagship Mixture-of-Experts (MoE) model from Xiaomi, released in April 2026, with 1.02 trillion total parameters and 42 billion activated per token. It builds on the MiMo-V2-Flash architecture, employing a hybrid attention scheme that interleaves Sliding Window Attention with global attention (6:1 ratio, 128-token sliding window) together with a three-layer Multi-Token Prediction module, and supports a 1M token context window. The model unifies reasoning and multimodal understanding for demanding agentic, complex software-engineering, and long-horizon tasks, sustaining trajectories that span thousands of tool calls, and is released as open weights under the MIT licence.

MiniMax-M2.7 is a Mixture-of-Experts model from MiniMax with 230 billion total parameters and approximately 10 billion activated per token (256 experts, 8 activated per token over 62 layers), supporting a 200K token context window. Released in March 2026, it is designed for coding, long-horizon agentic workflows, and office document editing, and is notable for a self-improving training loop in which an autonomous agent optimised parts of its own training pipeline. The weights are publicly available on Hugging Face under a modified MIT licence.

- **Small Open-Source Models.**

Llama-3.1-8B (Grattafiori et al., 2024) is a dense 8-billion-parameter open-weight model from Meta AI, released as part of the Llama 3.1 family. It supports multilingual text, code, reasoning, and tool use, with a 128K token context window, and is the smallest member of the Llama 3.1 series alongside the 70B and 405B variants.

Gemma-2-9B (Team et al., 2024) is a dense 9-billion-parameter open model from Google Deep-

Mind's Gemma Team. The 2B and 9B variants are trained via knowledge distillation rather than standard next-token prediction; the architecture uses interleaved local/global attention and grouped-query attention (GQA), and delivers performance competitive with models two to three times its size.

• Fine-tuned / Specialized Models.

QwQ-32B is a 32B open-weight reasoning model from Alibaba's Qwen team, trained with reinforcement learning for long chain-of-thought reasoning. It is not part of the fifteen-model full-benchmark suite; it is evaluated as the controlled base of STRuCT-LLM-Novo, the second specialisation pair in §5.4, on the stratified 210-session subset of Appendix F.

STRuCT-LLM-Novo (Stoisser et al., 2025) is a model from Stoisser et al. at Novo Nordisk that unifies Text-to-SQL and Text-to-Cypher generation under a single reinforcement learning framework. It employs Chain-of-Thought supervision, a topology-aware reward based on graph edit distance, and cross-formalism transfer (SQL training improves Cypher generation and vice versa), using QwQ-32B as its base model.

text-to-cypher-gemma is a Gemma-2-9B variant fine-tuned by Neo4j on the `neo4j/text2cypher-2024v1` dataset of 44,387 examples. It is released on Hugging Face as a research and demonstration model; Neo4j describes it as not production-ready.

CypherRI-7B† (Su et al., 2026) is a 7-billion-parameter Cypher-specialised model presented as a NeurIPS 2025 poster, which integrates schema selection into the Cypher-generation pipeline via reinforcement learning. In our evaluation, the model's reasoning output tokens contaminate all generated Cypher payloads, resulting in near-zero execution accuracy due to a format-compliance failure rather than generation ability; its results are excluded from aggregate statistics.

D Benchmark Reference

D.1 Knowledge Graph Schemas

CypherTurn uses seven purpose-built synthetic knowledge graphs, each constructed around a topological pattern that poses a distinct reasoning challenge. Figure 7 shows schema diagrams for all seven graphs.

Ancient Empire is a directed-cycle graph with 7 node types and 7 relation types. Edge properties on 5 relation types stress-test edge-property filtering. Deep 4-hop paths and dense numeric edge attributes make it a strong baseline for multi-hop traversal.

Magic Academy uses a bipartite-cycle topology with 6 node types. The dual Student–Professor path through both organisational units and courses simultaneously creates entity-resolution ambiguity, and Magic Academy ranks among the harder graphs for mid-tier models despite its regular bipartite structure.

Ocean Kingdom adopts a diamond-convergence structure: a central entity type fans out to three independent sub-graphs that partially reconverge. This creates path ambiguity and requires careful multi-hop planning to avoid retrieving duplicates.

Stellar Colony combines a tree-structured colony hierarchy with lateral `ALLIED_WITH` edges between peer factions. It is the only graph with peer-to-peer lateral edges, breaking the assumption of strictly hierarchical traversal.

Arcane Archive embeds a `REFERENCES` citation self-loop on Tome nodes within an otherwise linear chain, enabling recursive multi-hop queries. It is among the highest-mean-EX graphs, though Celestial Court and Ancient Empire rank slightly higher.

Celestial Court models a bureaucratic hierarchy where officials supervise other officials at arbitrary depth via a `SERVES_UNDER` self-referential edge. It uniquely exercises recursive organisational traversal.

Merchant Harbor uses a central invoice node that simultaneously bridges traders, buyers, and storehouses. It has the richest numeric edge properties of any graph in the benchmark and is the hardest new graph in the expanded dataset.

D.2 Conversational Phenomena

CypherTurn annotates every session turn with one of 13 conversational phenomena, grouped into four categories. Table 5 lists all 13 phenomena with their category, chain mode, chain-dependency status, a concise definition, and a canonical Cypher pattern.

D.3 Chain Modes

Each turn's chain dependency is characterised by a **chain mode** that describes how the current query relates to the prior result. Table 6 enumerates all 9 chain modes with their definitions and a brief Cypher illustration.

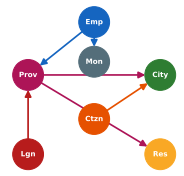
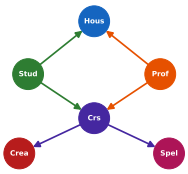
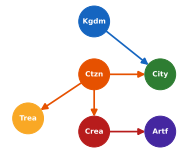
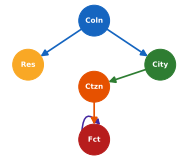
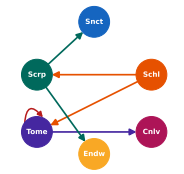
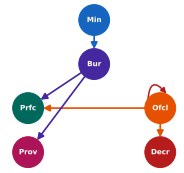
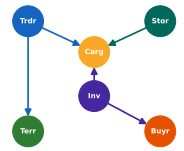
Name	Graph Schema	Nodes	Relations
ancient empire		Empire Province City Citizen Legion Resource Monument	CONTROLS HAS_CITY LIVES_IN STATIONED_IN EXTRACTS BUILT LOCATED_IN
magic academy		House Student Professor Course Creature Spell	BELONGS_TO_HOUSE HEADS_HOUSE ENROLLED_IN TEACHES STUDIES_CREATURE REQUIRES
ocean kingdom		Kingdom City Citizen Treasure Creature Artifact	GOVERNS RESIDENT_OF DISCOVERED INHABITED_BY TAMES GUARDS
stellar colony		Colony City Citizen Resource Faction	HAS_CITY INHABITED_BY BELONGS_TO PRODUCES ALLIED_WITH
arcane archive		Sanctum Scriptorium Scholar Tome Endowment Conclave	AFFILIATED_WITH WORKS_IN INSCRIBED_BY REFERENCES PRESENTED_AT SPONSORED_BY
celestial court		Ministry Bureau Prefecture Official Decree Province	OVERSEES BELONGS_TO STATIONED_IN SERVES_UNDER ENTRUSTED_WITH ADMINISTERS
merchant harbor		Trader Storehouse Cargo Invoice Buyer Territory	SUPPLIES HOLDS INCLUDES ISSUED_TO OPERATES_IN DOCKED_IN

Figure 7: Schema overview of the seven CypherTurn knowledge graphs. Each row: graph name, topology diagram, and node/relation listing. (a)–(d): directed cycle, bipartite cycle, diamond convergence, star+lateral; (e)–(g): citation self-loop, hierarchical self-loop, hub-and-spoke.

D.4 Session Personas

Each of the 721 sessions is assigned one of 10 personas drawn uniformly at random. The persona governs utterance register throughout the session. Table 7 defines all 10 persona types and gives a representative utterance example for each.

The pooled inter-annotator agreement of $\kappa = 0.83$ reported in §2.5 aggregates three annotation criteria. Table 8 gives a retrospective decomposition of it by criterion, derived from the annotation protocol structure rather than independent remeasurement of the two-annotator records. Agreement is near-ceiling for intent capture, because the gold Cypher is programmatically generated and database-verified. It remains high for cross-turn anaphora, anchored by the machine-generated

Phenomenon	Category	Chain Mode	Dep.	Definition and Cypher Pattern
EXPAND	Navigation	expand	✓	Traverse one hop outward from the prior result set; the prior result acts as the seed node set. `MATCH (n)-[:R]->(m) WHERE n.name IN $prev`
PIVOT	Navigation	pivot	✓	Rank the prior result by a property, then traverse from the top-k entities. `WITH n ORDER BY n.prop DESC LIMIT k MATCH (n)-[:R]->(m)`
FIRST	Navigation	pivot	✓	Reference the single top-ranked entity from the prior result and traverse from it. `WITH n ORDER BY n.prop DESC LIMIT 1 MATCH (n)-[:R]->(m)`
TOPIC_SHIFT	Discourse	topic_shift	✗	A context-break: the user poses an entirely new, independent query unrelated to prior results. The only non-chain-dependent phenomenon.
AGG_AVG	Aggregation	agg_numeric	✓	Compute the mean of a numeric property over the prior result set. `MATCH (n) WHERE n.name IN $prev RETURN avg(n.prop)`
AGG_MAX	Aggregation	agg_numeric	✓	Find the maximum value of a numeric property in the prior result set. `MATCH (n) WHERE n.name IN $prev RETURN max(n.prop)`
AGG_SUM	Aggregation	agg_numeric	✓	Sum a numeric property across all entities in the prior result set. `MATCH (n) WHERE n.name IN $prev RETURN sum(n.prop)`
COUNT	Aggregation	aggregate	✓	Count the cardinality of the prior result set. `MATCH (n) WHERE n.name IN $prev RETURN count(DISTINCT n)`
CONTRAST	Discourse	contrast	✓	Query the complement of the prior result within the same entity type. `MATCH (n:T) WHERE NOT n.name IN $prev RETURN n.name`
VALUE_FILTER	Filtering	value_narrow	✓	Filter the prior result by a numeric threshold on one property. `MATCH (n) WHERE n.name IN $prev AND n.prop > $threshold`
MULTI_COND	Filtering	narrow	✓	Apply two simultaneous attribute constraints to narrow the prior result. `WHERE n.prop1 = $v1 AND n.prop2 = $v2`
REFINE	Filtering	narrow	✓	Add a single new attribute constraint to narrow the prior result (property existence check). `WHERE n.name IN $prev AND n.prop IS NOT NULL`
FILTER	Filtering	narrow	✓	Filter prior result to entities that possess a given property. `WHERE n.name IN $prev AND n.prop IS NOT NULL`

Table 5: The 13 conversational phenomena in CYPHERTURN. **Dep.**: chain-dependent (✓) or independent (✗). Chain Mode = the session construction mode used to generate turns of this phenomenon.

dependency structure. It is lowest for utterance naturalness with respect to the assigned persona, which is the most subjective criterion. The pooled 0.83 falls between the weakest and strongest criteria, consistent with disagreement concentrated on persona naturalness. This decomposition supports the positioning of persona in §2.4 as a register-diversification device rather than a validated experimental factor.

D.5 Evaluation Protocols

Action space specification. Both protocols share an identical five-action space. Table 9 summarises the input and return value of each action; the paragraphs below describe the implementation constraints in detail.

`EXECUTE_CYPHER` runs the provided Cypher against the live Neo4j instance and returns the result rows, or an error message if the query is malformed or times out. A hard timeout of 120 s is enforced. To keep prompts tractable, current-turn results are truncated to 20 rows and prior-turn results stored in the guided history are truncated to 10 rows. The one exception is CONTRAST turns: the immediately preceding turn's result is passed without truncation so the model can compute the complement set.

`INSPECT_SCHEMA` returns the full graph schema — all node labels with their typed properties and all relation types with source and target labels. The action always succeeds and costs 1 budget unit. Under the guided protocol the schema is already injected into every turn prompt, making this action

Chain Mode	Dep.	Definition and Example
None (T1)	✗	First turn of a session; no prior result exists. The query is fully self-contained. *E.g.* "List all provinces." → `MATCH (p:Province) RETURN p.name`
topic_shift	✗	A deliberate context-break mid-session. The user abandons the prior thread and opens a new independent query. *E.g.* (after querying officials) "Actually, list all trade routes."
expand	✓	Traverse one hop outward from the prior result. The prior entity set is used directly as a seed. *E.g.* (after T1 provinces) "What cities are in those provinces?"
pivot	✓	Rank the prior result by a property, select top-k, and traverse from those entities. *E.g.* "Among those cities, which one has the highest population? Show its legions."
agg_numeric	✓	Compute a numeric aggregate (avg, max, sum) over a property of the prior result set. *E.g.* "What is the average tribute gold of those provinces?"
aggregate	✓	Count the cardinality of the prior result set. *E.g.* "How many legions were returned above?"
contrast	✓	Query the logical complement of the prior result within the same entity type. *E.g.* "Now show me all provinces *not* in that list."
value_narrow	✓	Filter the prior result by a numeric threshold on one property. *E.g.* "Which of those have a tribute gold greater than 5000?"
narrow	✓	Add one or two attribute constraints to narrow the prior result (categorical or existence check). *E.g.* "Among those officials, show only those whose rank is 'Commander'."

Table 6: The 9 chain modes in CYPHERTURN. **Dep.**: chain-dependent (✓) or independent (✗). *None (T1)* and *topic_shift* are the two non-chain-dependent modes; all others require the prior result to formulate the current query.

redundant but still valid. Under the agentic protocol the schema is withheld from the system prompt and must be actively discovered via this action.

`SEARCH_VALUES` resolves entity-linking mismatches between user phrasing and stored property values. It accepts a JSON payload with keys `label`, `property`, and `query`, and applies a cascading fuzzy search: first a case-insensitive `CONTAINS` match, then a `STARTS WITH` prefix match, and finally a per-word `CONTAINS` match for multi-word queries. Up to 10 matching values are returned. A parse error is returned on malformed JSON input.

`ASK_USER` poses a clarifying question to the user and returns a string reply. It does *not* terminate the action loop; the model may continue issuing further actions after receiving the reply. Under the guided protocol, the reply is scripted: turns annotated with `ambiguity_info` return the pre-authored disambiguation answer, while all other turns return the fixed discouragement string *"I think my question was clear enough. Please try to answer it directly."* Under the agentic protocol, the reply is generated by a dedicated LLM-based user simulator described in the paragraph below.

`SUBMIT_ANSWER` finalises the predicted query for the current turn and immediately terminates the action loop; only the first call counts. If the budget is exhausted before `SUBMIT_ANSWER` is called, the runner falls back to the payload of the last `EXECUTE_CYPHER` action in the current turn. If no `EXECUTE_CYPHER` was issued either, the predicted query is set to `None`, scoring EX = 0.

Guided protocol mechanics. At the start of each turn, the runner constructs a gold conversation history by formatting all prior turns as `User: <utterance> / Assistant: [Query Result] <gold result, up to 10 rows>` pairs. The full graph schema is injected directly into the turn prompt alongside this history. The model then enters an action loop capped at exactly 3 steps; the cap is a hard range and cannot be extended. For CONTRAST turns, the immediately preceding turn's gold result is passed without row truncation so the model can compute the complement set. For FIRST turns that reference a specific earlier result, a dependency hint is appended to the user utterance: *"[Note: This turn builds upon the result of Turn N. Use the result shown for Turn N in the conversation history above.]"* When `ASK_USER` is called, the system returns a scripted reply rather than invoking a live simulator: turns annotated with `ambiguity_info` return the pre-authored clarification; all other turns return the fixed string *"I think my question was clear enough. Please try to answer it directly."* This makes `ASK_USER` practically useless in the guided setting, as it consumes one of three budget actions for minimal informational gain.

Persona	Definition	Example Utterance
Journalist	Investigative, direct phrasing; seeks evidence and facts; often frames queries as questions with implicit causal intent.	"Which provinces experienced the sharpest decline in tribute payments last year?"
Student	Exploratory and learning-oriented; may ask elementary questions or seek clarification on results.	"I'm trying to understand the hierarchy — can you show me which officials answer to the Emperor?"
Product Manager	Goal-focused; wants summary statistics and ranked lists; minimal tolerance for schema details.	"Give me the top three trade routes by cargo volume."
Novice User	Vague phrasing; often uses pronouns ambiguously ("those things", "that one") without clear referents.	"Now what about the other ones? The ones with the big numbers?"
Data Analyst	Precise and technical; requests specific columns, sorted output, and numeric breakdowns.	"Return the entity ID, name, and tribute_gold for each province, ordered descending."
Casual User	Informal register; contractions, colloquialisms, and incomplete sentences. May mix social and query intent.	"ok cool — so like, how many of those are there actually?"
Impatient Executive	Terse and demanding; short imperative sentences; expects single-number or short-list answers.	"Total revenue. Now."
Adversarial User	Deliberately challenges the system; uses skeptical framing, mock demands, or claims the system will fail.	"I bet you can't tell me which of those provinces have more than two legions stationed there."
Non-Native Speaker	Non-standard grammar, simplified vocabulary, and occasional word-order errors. Meaning is recoverable.	"Please showing me the official who is highest in rank from previous list?"
Verbose User	Over-specified utterances with redundant context, restated background, and multi-clause formulations.	"Taking into account all of the provinces we identified earlier in our analysis — specifically those that were shown in the results of the query we ran at the start of our session — could you now compute for me the total sum of their tribute gold values?"

Table 7: The 10 session personas in CypherTurn. Each session is assigned one persona uniformly at random (66–75 sessions per type across 721 sessions total). Persona governs utterance register for every turn in the session.

Criterion	κ
Gold Cypher captures intent	0.91
Cross-turn anaphora unambiguous	0.86
Utterance natural for assigned persona	0.64
Pooled (reported)	0.83

Table 8: Inter-annotator agreement κ decomposed by criterion. Persona naturalness is the weakest criterion. Criterion-level values are a retrospective decomposition based on the annotation protocol structure; the pooled $\kappa = 0.83$ is measured.

Action	Input	Returns
`EXECUTE_CYPHER`	Cypher query string	Result rows; error on failure
`INSPECT_SCHEMA`	Empty (ignored)	Node labels, properties, relation types
`SEARCH_VALUES`	JSON: label, property, query	Up to 10 matching values
`ASK_USER`	Natural language question	String reply from user
`SUBMIT_ANSWER`	Final Cypher query	`"Answer submitted"`

Table 9: The five-action space shared by both evaluation protocols. Each action costs 1 budget unit.

Agentic protocol mechanics. The total action budget is computed as $B = T \times m$, where T is the number of turns in the session and $m \in \{3, 5, 10\}$ is the budget multiplier. This budget is shared across the entire session and decrements with every action regardless of turn. The graph schema is withheld from the system prompt; the model must call `INSPECT_SCHEMA` to discover it. After each turn, the runner builds an interaction history from all completed turns, recording `EXECUTE_CYPHER` payloads and results (truncated to 200 characters), `ASK_USER` exchanges, and `SUBMIT_ANSWER` payloads; `INSPECT_SCHEMA` and `SEARCH_VALUES` actions are omitted from the history to reduce context length. If the shared budget reaches zero mid-session, all remaining turns receive `pred_cypher = None` and score EX = 0 — a hard cliff rather than graceful degradation. This design choice means that models which over-invest budget in early turns (e.g., repeated `EXECUTE_CYPHER` retries without

schema recovery) face compounding penalties in later turns.

Agentic user simulator. The ASK_USER action in the agentic protocol invokes a dedicated LLM-based user simulator operating under a two-phase anti-leakage pipeline. In the first phase, the simulator classifies the model's question into one of three intents: (1) **Reject** — if the question probes schema or technical details (e.g., contains keywords such as *schema*, *label*, *relationship*, or *cypher*), the simulator returns a canned rejection from a rotating pool of four templates (e.g., "I don't know about the database schema. I just want to find the information I asked about."); (2) **Disambiguate** — if the turn has ambiguity_info and the question matches a clarification-seeking pattern, the simulator returns the pre-authored disambiguation reply; (3) **General** — otherwise, the simulator generates a reply via an LLM call at temperature 0.5. In the second phase, the generated reply is scanned for schema terms (node labels, relation types, property names); any reply containing such terms is replaced with a safe fallback. After three or more ASK_USER calls within a single turn, an impatience signal is appended to the simulator context. This design prevents the model from extracting schema information through conversational probing while still allowing genuine disambiguation.

Protocol variants for the sensitivity ablation. Three variants of the agentic protocol are used in the sensitivity ablation reported in Appendix F. Each modifies exactly one dimension of the base protocol and leaves the rest unchanged, so that any change in the ranking can be attributed to the perturbed dimension. The **Agentic-PS** variant makes the graph schema persistent: after the model calls INSPECT_SCHEMA for the first time, the full schema returned by that call is injected into the context of every subsequent turn for the remainder of the session, and further INSPECT_SCHEMA calls become redundant but still cost one budget unit. This variant tests whether the strong correlation between INSPECT_SCHEMA allocation and agentic EX reflects a causal role for repeated schema inspection. The **Disclosed-Horizon** variant removes the unknown-horizon confound by disclosing the session length T, the per-turn action target m, and a suggested remaining budget at the start of each turn, so the model can pace its spending across the full session rather than rationing against an unknown endpoint. The **Budget-Encouraging Prompt** variant replaces the base efficiency instruction, which asks the model to prefer submitting efficiently, with explicit encouragement to use the full available budget for verification and value retrieval, and is run at tenfold budget to give the encouragement room to act. This variant tests whether the self-limitation observed in §5.3 stems from the efficiency wording rather than a capability gap.

E Detailed Experimental Results

E.1 Per-Graph Results

Table 10 reports guided EX by knowledge graph for all 15 models. Stellar Colony (0.426) and Merchant Harbor (0.434) are the hardest graphs, reflecting the difficulty of lateral alliance edges and multi-party invoice structures respectively. Celestial Court is the easiest (0.485), benefiting from regular hierarchical traversal patterns under oracle context. For small models, per-model maxima often fall on Magic Academy, suggesting that bipartite paths become relatively easier once multi-condition and complement queries that dominate harder graphs are systematically failed.

Table 11 reports the corresponding agentic $\times 3$ results. The graph-level ranking is broadly preserved under the agentic protocol, but absolute EX values drop sharply: Arcane Archive shows the largest absolute drop, from 0.469 (guided) to 0.250 (agentic $\times 3$), while Celestial Court drops from 0.485 to 0.279.

E.2 Per-Phenomenon Results

Table 12 reports per-phenomenon guided EX for all 14 models (CypherRI-7B excluded). REFINE (mean 0.244) is the hardest phenomenon universally: even Claude achieves only 0.409. CONTRAST has the highest inter-model variance (0.021–0.540). Aggregation phenomena (AGG_MAX, AGG_AVG, AGG_SUM) form the easiest cluster across all models, while Filtering and Navigation phenomena show the widest within-category spread.

E.3 Per-Chain-Mode Results

Table 14 breaks down guided EX by chain mode. Among the two non-chain-dependent modes, *topic_shift* achieves by far the highest mean EX at 0.642, confirming that fresh independent queries are inherently easier, while *None (T1)* sits at 0.359, comparable to the mid-tier chain-dependent modes: first turns are self-contained

Model	Anc	Mag	Ocn	Stl	Arc	Cel	Mer	Avg
Frontier Models								
Claude Opus 4.7	0.674	0.594	0.660	0.603	**0.683**	0.663	0.650	0.647
GPT-5.5	0.633	0.570	0.659	0.597	**0.671**	0.667	0.581	0.625
Kimi-K2.5	0.620	0.568	0.588	0.551	**0.667**	0.618	0.605	0.602
Gemini-3.1-Flash-Lite	0.574	**0.620**	0.535	0.520	0.604	0.595	0.555	0.572
Qwen3-235B	0.548	0.488	0.529	0.472	0.535	**0.564**	0.457	0.513
GLM-5	**0.588**	0.439	0.485	0.468	0.538	0.560	0.468	0.506
DeepSeek-V3.2	0.529	0.486	0.513	0.466	0.546	**0.549**	0.453	0.506
ERNIE-5.0	0.459	0.508	0.493	0.441	0.430	**0.517**	0.431	0.468
MiMo-V2.5-Pro	**0.514**	0.403	0.473	0.417	0.498	0.486	0.433	0.460
MiniMax-M2.7	0.158	0.114	0.150	0.135	0.134	**0.163**	0.119	0.139
Small Open-Source Models								
Llama-3.1-8B	0.280	**0.333**	0.316	0.264	0.254	0.314	0.273	0.291
Gemma-2-9B	0.255	**0.305**	0.286	0.289	0.259	**0.305**	0.266	0.281
Fine-tuned / Specialized Models								
STRuCT-LLM-Novo	**0.549**	0.528	0.542	0.484	0.537	0.526	0.515	0.526
text-to-cypher-gemma	0.254	0.282	**0.294**	0.257	0.205	0.265	0.272	0.261
CypherRI-7B†	**0.019**	0.009	0.013	0.015	0.009	0.008	0.014	0.013
Mean (14)	**0.474**	0.446	0.466	0.426	0.469	**0.485**	0.434	—

Table 10: Guided EX by knowledge graph. Anc=Ancient Empire, Mag=Magic Academy, Ocn=Ocean Kingdom, Stl=Stellar Colony, Arc=Arcane Archive, Cel=Celestial Court, Mer=Merchant Harbor. **Bold**: per-row maximum. † excluded from mean row.

Model	Anc	Mag	Ocn	Stl	Arc	Cel	Mer	Avg
Frontier Models								
Claude Opus 4.7	0.427	0.376	0.404	0.358	0.410	**0.454**	0.384	0.402
GPT-5.5	0.433	0.323	**0.494**	0.389	0.416	0.415	0.340	0.401
Kimi-K2.5	0.363	0.275	0.339	0.278	0.373	**0.411**	0.311	0.335
Gemini-3.1-Flash-Lite	0.445	**0.479**	0.442	0.392	0.409	0.446	0.409	0.432
Qwen3-235B	0.221	0.167	0.206	0.166	0.184	**0.248**	0.204	0.199
GLM-5	**0.358**	0.229	0.278	0.228	0.268	0.338	0.221	0.274
DeepSeek-V3.2	0.237	0.162	0.229	0.164	0.195	**0.239**	0.201	0.204
ERNIE-5.0	0.278	0.313	0.309	0.271	0.270	0.311	**0.317**	0.295
MiMo-V2.5-Pro	**0.311**	0.205	0.276	0.212	0.271	0.288	0.190	0.250
MiniMax-M2.7	0.172	0.130	0.166	0.142	0.145	**0.208**	0.132	0.156
Small Open-Source Models								
Llama-3.1-8B	0.141	**0.194**	0.157	0.166	0.132	0.001	0.164	0.137
Gemma-2-9B	0.146	**0.194**	0.120	0.179	0.057	0.074	0.179	0.136
Fine-tuned / Specialized Models								
STRuCT-LLM-Novo	0.350	0.317	**0.366**	0.278	0.346	0.364	0.312	0.333
text-to-cypher-gemma	0.164	0.192	0.155	**0.201**	0.025	0.106	0.159	0.143
CypherRI-7B†	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Mean (14)	**0.289**	0.254	0.281	0.245	0.250	0.279	0.252	—

Table 11: Agentic ×3 EX by knowledge graph. Same layout as Table 10. **Bold**: per-row maximum. † excluded from mean row.

but require cold-start schema grounding. Among chain-dependent modes, *agg_numeric* and *aggregate* are the easiest, while *contrast* and *narrow* (which underlies REFINE) are the hardest. The agentic ×3 column shows the absolute degradation from removing oracle context: *aggregate* and *value_narrow* show the largest absolute drops (0.29 and 0.28 points), and *contrast* the largest relative drop, retaining only 27% of its guided EX, as context-resolution errors compound over long sessions.

E.4 Per-Persona Results

Table 16 reports guided EX by session persona for all 14 models. The spread across persona types within a single model is moderate (typically 10–16 percentage points for frontier models), confirming that persona does not dramatically inflate or sup-

Phenom.	Cat.	Cl	GPT	Km	Gfl	ST	Q3	GL	DS	ER	Mi	Mx	Ll	G2	t2	Mean
AGG_MAX	Agg	**0.842**	0.813	0.788	0.741	0.796	0.798	0.840	0.793	0.603	0.810	0.310	0.360	0.409	0.197	0.650
AGG_AVG	Agg	0.767	0.679	0.740	0.677	0.740	0.722	**0.780**	0.700	0.578	0.702	0.285	0.336	0.422	0.298	0.602
TOPIC_SHIFT	Disc	0.821	**0.849**	0.760	0.782	0.680	0.643	0.487	0.656	0.778	0.422	0.111	0.788	0.538	0.673	0.642
AGG_SUM	Agg	0.642	0.663	0.661	0.655	0.665	0.692	**0.696**	0.676	0.547	0.661	0.281	0.331	0.397	0.318	0.563
VALUE_FILTER	Filt	0.725	0.753	**0.764**	0.590	0.657	0.691	0.562	0.702	0.590	0.528	0.112	0.270	0.354	0.382	0.549
COUNT	Agg	0.670	0.614	0.636	0.583	0.514	0.657	**0.673**	0.604	0.520	0.611	0.274	0.156	0.352	0.240	0.507
FIRST	Nav	**0.614**	0.560	0.481	0.519	0.376	0.506	0.471	0.501	0.361	0.412	0.065	0.075	0.100	0.115	0.368
PIVOT	Nav	**0.607**	0.548	0.499	0.507	0.426	0.496	0.442	0.486	0.309	0.419	0.071	0.091	0.088	0.091	0.363
EXPAND	Nav	0.543	**0.626**	0.548	0.419	0.520	0.363	0.442	0.355	0.351	0.486	0.140	0.116	0.098	0.151	0.368
FILTER	Filt	**0.670**	0.614	0.602	0.216	0.591	0.511	0.295	0.489	0.284	0.227	0.011	0.227	0.341	0.205	0.377
MULTI_COND	Filt	**0.582**	0.566	0.551	0.367	0.495	0.327	0.332	0.316	0.291	0.250	0.036	0.122	0.311	0.235	0.341
CONTRAST	Disc	**0.540**	0.515	0.460	0.384	0.245	0.236	0.481	0.228	0.173	0.418	0.072	0.030	0.025	0.021	0.273
REFINE	Filt	**0.409**	0.375	0.375	0.170	0.364	0.307	0.205	0.330	0.216	0.159	0.011	0.148	0.205	0.148	0.244
Overall	—	0.647	0.625	0.602	0.572	0.526	0.513	0.506	0.506	0.468	0.460	0.139	0.291	0.281	0.261	0.457

Table 12: Guided EX by phenomenon for all 14 models (CypherRI excluded). Cl=Claude 4.7, GPT=GPT-5.5, Km=Kimi-K2.5, Gfl=Gemini-3.1-FL, ST=STRuCT-LLM, Q3=Qwen3-235B, GL=GLM-5, DS=DeepSeek-V3.2, ER=ERNIE-5.0, Mi=MiMo-V2.5-Pro, Mx=MiniMax-M2.7, Ll=Llama-3.1-8B, G2=Gemma-2-9B, t2=t2c-gemma. Red rows : mean EX below 0.30. **Bold**: best per row.

Phenom.	Cat.	Cl	GPT	Km	Gfl	ST	Q3	GL	DS	ER	Mi	Mx	Ll	G2	t2	Mean
AGG_MAX	Agg	**0.663**	0.611	0.599	0.579	0.549	0.505	0.579	0.510	0.443	0.601	0.480	0.195	0.187	0.121	0.473
AGG_AVG	Agg	**0.576**	0.500	0.507	0.455	0.471	0.374	0.516	0.381	0.316	0.527	0.399	0.141	0.155	0.130	0.389
TOPIC_SHIFT	Disc	0.533	0.644	0.433	**0.739**	0.446	0.225	0.289	0.233	0.563	0.249	0.222	0.178	0.200	0.313	0.376
AGG_SUM	Agg	**0.524**	0.501	0.457	0.418	0.414	0.345	0.426	0.343	0.310	0.443	0.331	0.121	0.119	0.094	0.346
VALUE_FILTER	Filt	0.399	**0.416**	0.225	0.337	0.287	0.101	0.169	0.112	0.292	0.112	0.039	0.090	0.135	0.197	0.208
COUNT	Agg	**0.548**	0.455	0.308	0.246	0.333	0.109	0.402	0.100	0.062	0.324	0.146	0.000	0.003	0.000	0.217
FIRST	Nav	0.317	0.243	0.293	**0.412**	0.269	0.151	0.228	0.147	0.218	0.201	0.038	0.062	0.062	0.059	0.193
PIVOT	Nav	0.325	0.249	0.295	**0.436**	0.249	0.142	0.226	0.151	0.201	0.205	0.036	0.047	0.047	0.047	0.190
EXPAND	Nav	0.200	**0.349**	0.210	0.244	0.269	0.082	0.129	0.087	0.107	0.097	0.044	0.047	0.044	0.044	0.140
FILTER	Filt	0.261	**0.284**	0.239	**0.284**	0.193	0.023	0.136	0.034	0.239	0.080	0.023	0.136	0.125	0.193	0.161
MULTI_COND	Filt	**0.260**	0.209	0.173	0.194	0.184	0.051	0.066	0.061	0.214	0.041	0.020	0.077	0.102	0.122	0.127
CONTRAST	Disc	0.186	**0.207**	0.105	0.152	0.089	0.046	0.034	0.063	0.080	0.038	0.017	0.021	0.008	0.004	0.075
REFINE	Filt	0.148	0.170	0.125	**0.216**	0.114	0.045	0.091	0.034	0.114	0.045	0.011	0.057	0.080	0.091	0.096
Overall	—	0.402	0.401	0.335	0.432	0.333	0.199	0.274	0.204	0.295	0.250	0.156	0.137	0.136	0.143	0.264

Table 13: Agentic $\times 3$ EX by phenomenon for all 14 models (CypherRI excluded). Same abbreviations as Table 12. Red rows : mean EX below 0.10. **Bold**: best per row.

Chain Mode	Dep.	Cl	GPT	Km	Gfl	ST	Q3	GL	DS	ER	Mi	Mx	Ll	G2	t2	Mean
None (T1)	✗	0.519	0.447	0.533	**0.619**	0.338	0.207	0.229	0.212	0.498	0.132	0.049	0.523	0.368	0.351	0.359
topic_shift	✗	0.821	**0.849**	0.760	0.782	0.680	0.643	0.487	0.656	0.778	0.422	0.111	0.788	0.538	0.673	0.642
agg_numeric	✓	0.753	0.716	0.730	0.690	0.733	0.737	**0.771**	0.723	0.576	0.724	0.292	0.342	0.409	0.271	0.603
aggregate	✓	0.670	0.614	0.636	0.583	0.514	0.657	**0.673**	0.604	0.520	0.611	0.274	0.156	0.352	0.240	0.507
value_narrow	✓	0.650	**0.655**	0.652	0.473	0.572	0.500	0.441	0.500	0.433	0.382	0.072	0.193	0.332	0.305	0.440
pivot	✓	**0.611**	0.554	0.489	0.513	0.401	0.501	0.456	0.494	0.335	0.416	0.068	0.083	0.094	0.103	0.366
expand	✓	0.543	**0.626**	0.548	0.419	0.520	0.363	0.442	0.355	0.351	0.486	0.140	0.116	0.098	0.151	0.368
narrow	✓	**0.540**	0.494	0.489	0.193	0.477	0.409	0.250	0.409	0.250	0.193	0.011	0.188	0.273	0.176	0.311
contrast	✓	**0.540**	0.515	0.460	0.384	0.245	0.236	0.481	0.228	0.173	0.418	0.072	0.030	0.025	0.021	0.273
Overall	—	0.647	0.625	0.602	0.572	0.526	0.513	0.506	0.506	0.468	0.460	0.139	0.291	0.281	0.261	0.457

Table 14: Guided EX by chain mode for all 14 models (CypherRI excluded). Same model abbreviations as Table 12. **Bold**: best per row. **Dep.**: chain-dependent (✓) or independent (✗).

press raw generation ability under oracle context. The *verbose user* persona consistently produces the highest per-model EX (column mean 0.511), as explicit, detailed phrasing reduces ambiguity in context resolution. The *impatient executive* persona produces the lowest per-model EX under the agentic protocol (column mean 0.226 in Table 17), as terse utterances with implicit referents require stronger context resolution once oracle history is removed; under guided oracle context the persona effect is weaker, and the impatient executive sits only marginally above the novice user, which is lowest among the column means.

Chain Mode	Dep.	Cl	GPT	Km	Gfl	ST	Q3	GL	DS	ER	Mi	Mx	Ll	G2	t2	Mean
None (T1)	✗	0.383	0.409	0.304	**0.555**	0.347	0.202	0.211	0.214	0.510	0.164	0.136	0.455	0.411	0.411	0.336
topic_shift	✗	0.533	0.644	0.433	**0.739**	0.446	0.225	0.289	0.233	0.563	0.249	0.222	0.178	0.200	0.313	0.376
agg_numeric	✓	**0.584**	0.534	0.517	0.479	0.474	0.404	0.503	0.407	0.353	0.519	0.399	0.150	0.152	0.114	0.399
aggregate	✓	**0.548**	0.455	0.308	0.246	0.333	0.109	0.402	0.100	0.062	0.324	0.146	0.000	0.003	0.000	0.217
value_narrow	✓	**0.326**	0.307	0.198	0.262	0.233	0.075	0.115	0.086	0.251	0.075	0.029	0.083	0.118	0.158	0.165
pivot	✓	0.321	0.246	0.294	**0.424**	0.259	0.147	0.227	0.149	0.209	0.203	0.037	0.055	0.055	0.053	0.191
expand	✓	0.200	**0.349**	0.210	0.244	0.269	0.082	0.129	0.087	0.107	0.097	0.044	0.047	0.044	0.044	0.140
narrow	✓	0.205	0.227	0.182	**0.250**	0.153	0.034	0.114	0.034	0.176	0.062	0.017	0.097	0.102	0.142	0.128
contrast	✓	0.186	**0.207**	0.105	0.152	0.089	0.046	0.034	0.063	0.080	0.038	0.017	0.021	0.008	0.004	0.075
Overall	—	0.402	0.401	0.335	0.432	0.333	0.199	0.274	0.204	0.295	0.250	0.156	0.137	0.136	0.143	0.264

Table 15: Agentic ×3 EX by chain mode for all 14 models (CypherRI excluded). Same abbreviations as Table 14. **Bold**: best per row.

Model	Imp	Jou	Cas	Stu	PMg	NNS	DA	Nov	Adv	Vrb	Avg
Frontier Models											
Claude Opus 4.7	0.559	0.627	0.678	0.661	0.621	0.635	0.619	0.646	0.708	**0.720**	0.647
GPT-5.5	0.569	0.624	0.652	0.646	0.622	0.629	0.603	0.602	0.611	**0.698**	0.625
Kimi-K2.5	0.595	0.588	0.624	0.609	0.585	0.621	0.593	0.573	0.569	**0.678**	0.602
Gemini-3.1-Flash-Lite	0.516	0.574	0.591	0.552	0.553	0.567	0.555	0.561	0.583	**0.676**	0.572
Qwen3-235B	0.507	0.505	0.526	0.465	0.509	0.498	0.543	0.468	0.542	**0.578**	0.513
GLM-5	0.483	0.497	0.524	0.472	0.516	0.486	0.500	0.481	0.514	**0.601**	0.506
DeepSeek-V3.2	0.497	0.500	0.524	0.460	0.498	0.505	0.526	0.465	0.530	**0.565**	0.506
ERNIE-5.0	0.464	0.431	**0.497**	0.483	0.422	0.480	0.474	0.462	0.488	0.490	0.468
MiMo-V2.5-Pro	0.419	0.458	0.486	0.402	0.483	0.465	0.500	0.370	0.484	**0.550**	0.460
MiniMax-M2.7	0.150	0.121	0.134	0.109	0.169	0.152	**0.217**	0.103	0.109	0.124	0.139
Small Open-Source Models											
Llama-3.1-8B	0.298	0.270	0.286	0.305	0.263	0.289	0.281	0.308	**0.308**	0.300	0.291
Gemma-2-9B	0.283	0.271	0.268	0.253	0.278	0.271	**0.323**	0.242	0.308	0.316	0.281
Fine-tuned / Specialized Models											
STRuCT-LLM-Novo	0.503	0.516	0.557	0.513	0.527	0.480	0.513	0.499	0.565	**0.590**	0.526
text-to-cypher-gemma	0.278	0.252	0.289	0.223	0.258	0.259	**0.305**	0.218	0.266	0.271	0.261
Mean (14)	0.437	0.445	0.474	0.440	0.450	0.453	0.468	0.428	0.470	**0.511**	—

Table 16: Guided EX by persona for all 14 models (CypherRI excluded). Imp=Impatient Executive, Jou=Journalist, Cas=Casual User, Stu=Student, PMg=Product Manager, NNS=Non-Native Speaker, DA=Data Analyst, Nov=Novice User, Adv=Adversarial User, Vrb=Verbose User. **Bold**: per-row maximum.

E.5 Agentic Budget Scaling

Table 18 reports full turn-level results under all three agentic budgets (×3 / ×5 / ×10). Six budget-response strategies emerge across models: **budget-insensitive** (Gemini-3.1-FL: EX flat at 0.432/0.431/0.430 with near-constant token usage, schema-heavy allocation exhausts marginal EX gain); **plateau at ×5** (Claude: +2.7 points ×3→×5, flat ×5→×10; STRuCT-LLM similarly); **consistent gain** (ERNIE, Kimi, Llama, Gemma, t2c, and, with smaller monotonic increments, DeepSeek-V3.2 and Qwen3-235B: monotonic EX improvement, most pronounced for small models whose token usage roughly doubles from ×3 to ×10); **late gain** (MiMo, MiniMax: flat ×3→×5, measurable gain at ×10); **overbudget degradation** (GLM-5: peaks at ×5 then degrades at ×10 when extra inspection overwrites correct partial solutions); **volatile** (GPT-5.5: inconsistent across budgets). Action usage analysis is provided in Table 19.

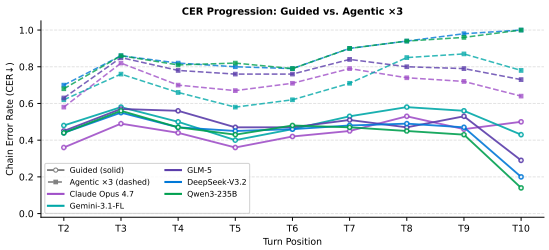


Figure 8: Chain Error Rate (CER) by turn position under the Guided (solid) and Agentic ×3 (dashed) protocols for five representative models.

Figure 8 contrasts CER trajectories across turn positions. Under the guided protocol, CER remains moderate and near-flat for all models, confirming that gold oracle context quarantines each failure

Model	Imp	Jou	Cas	Stu	PMg	NNS	DA	Nov	Adv	Vrb	Avg
Frontier Models											
Claude Opus 4.7	0.298	0.411	0.423	0.383	0.397	0.319	0.384	0.417	0.470	**0.519**	0.402
GPT-5.5	0.347	0.409	0.441	0.394	0.397	0.363	0.382	0.396	**0.447**	0.437	0.401
Kimi-K2.5	0.295	0.331	0.378	0.324	0.319	0.328	**0.379**	0.273	0.363	0.369	0.335
Gemini-3.1-Flash-Lite	0.383	0.423	0.444	0.427	0.399	0.404	0.437	0.481	0.424	**0.495**	0.432
Qwen3-235B	0.145	0.207	0.214	0.160	0.195	0.147	**0.268**	0.193	0.208	0.254	0.199
GLM-5	0.224	0.270	0.305	0.285	0.246	0.271	**0.348**	0.242	0.224	0.329	0.274
DeepSeek-V3.2	0.162	0.215	0.204	0.188	0.181	0.147	0.275	0.172	0.215	**0.282**	0.204
ERNIE-5.0	0.281	**0.328**	0.254	0.310	0.256	0.300	0.296	0.296	0.310	0.320	0.295
MiMo-V2.5-Pro	0.186	0.281	0.274	0.229	0.248	0.183	**0.351**	0.218	0.217	0.313	0.250
MiniMax-M2.7	0.147	0.157	0.192	0.147	0.130	0.144	**0.243**	0.129	0.109	0.166	0.156
Small Open-Source Models											
Llama-3.1-8B	0.152	0.129	0.136	0.150	0.105	0.142	0.114	**0.155**	0.130	**0.155**	0.137
Gemma-2-9B	0.124	0.139	0.132	0.142	0.115	0.144	0.136	0.137	**0.155**	0.133	0.136
Fine-tuned / Specialized Models											
STRuCT-LLM-Novo	0.269	0.329	**0.401**	0.343	0.338	0.280	0.364	0.334	0.317	0.351	0.333
text-to-cypher-gemma	0.157	0.158	0.134	0.139	0.117	0.137	0.149	0.121	**0.173**	0.150	0.143
Mean (14)	0.226	0.270	0.281	0.259	0.246	0.236	0.295	0.255	0.269	**0.305**	—

Table 17: Agentic ×3 EX by persona for all 14 models (CypherRI excluded). Imp=Impatient Executive, Jou=Journalist, Cas=Casual User, Stu=Student, PMg=Product Manager, NNS=Non-Native Speaker, DA=Data Analyst, Nov=Novice User, Adv=Adversarial User, Vrb=Verbose User. **Bold**: per-row maximum.

	Agentic ×3					Agentic ×5					Agentic ×10				
Model	EX	PSJS	CER	SEM	Tok/S	EX	PSJS	CER	SEM	Tok/S	EX	PSJS	CER	SEM	Tok/S
Frontier Models															
Claude Opus 4.7	0.402	0.681	0.712	**0.019**	639	0.429	**0.714**	0.697	**0.025**	620	0.424	0.706	0.704	**0.019**	634
GPT-5.5	0.401	0.673	0.698	0.006	554	0.395	0.661	0.707	0.004	564	0.406	0.680	0.704	0.007	546
Kimi-K2.5	0.335	0.612	0.749	0.006	965	0.346	0.621	0.734	0.008	1,006	0.355	0.628	0.731	0.012	934
Gemini-3.1-Flash-Lite	**0.432**	**0.698**	**0.695**	0.008	**427**	**0.431**	0.697	**0.694**	0.007	**427**	**0.430**	0.695	**0.694**	0.007	**404**
Qwen3-235B	0.199	0.548	0.833	0.001	927	0.216	0.561	0.814	0.001	1,058	0.218	0.563	0.813	0.001	1,251
GLM-5	0.274	0.623	0.773	0.006	749	0.279	0.631	0.773	0.001	943	0.271	0.619	0.774	0.004	874
DeepSeek-V3.2	0.204	0.551	0.835	0.000	965	0.217	0.562	0.813	0.001	1,018	0.223	0.569	0.810	0.001	1,209
ERNIE-5.0	0.295	0.584	0.830	0.003	867	0.328	0.606	0.783	0.003	1,474	0.339	0.614	0.766	0.001	1,382
MiMo-V2.5-Pro	0.250	0.591	0.781	0.001	871	0.255	0.596	0.778	0.001	1,008	0.266	0.604	0.764	0.001	1,028
MiniMax-M2.7	0.156	0.403	0.862	0.000	1,805	0.161	0.409	0.854	0.000	1,971	0.169	0.416	0.848	0.000	2,339
Small Open-Source Models															
Llama-3.1-8B	0.137	0.371	0.962	0.000	924	0.203	0.433	0.908	0.000	1,626	0.261	0.488	0.842	0.000	3,519
Gemma-2-9B	0.136	0.378	0.948	0.000	844	0.192	0.421	0.897	0.000	1,511	0.202	0.431	0.877	0.000	1,779
Fine-tuned / Specialized Models															
STRuCT-LLM-Novo	0.333	0.680	0.766	0.003	1,346	0.333	0.701	0.758	0.003	1,762	0.340	**0.713**	0.738	0.004	1,609
text-to-cypher-gemma	0.143	0.369	0.934	0.000	792	0.191	0.407	0.889	0.000	1,143	0.198	0.415	0.877	0.000	1,640
CypherRI-7B†	0.001	0.068	0.999	0.000	23,495	0.001	0.125	0.999	0.000	20,867	—	—	—	—	—

Table 18: Agentic Protocol results across three action budgets (×3 / ×5 / ×10). **Tok/S**: mean clean output tokens per session (excl. `<think>`). **—**: results unavailable. † excluded from aggregate statistics.

and limits propagation. Under the agentic protocol, DeepSeek-V3.2 and Qwen3-235B reach near-1.0 CER by T_9–T_{10}, while Gemini-3.1-Flash-Lite plateaus at approximately 0.70, consistent with its schema-inspection strategy that partially insulates context from prior errors. The growing gap between solid and dashed curves quantifies the Autonomy Divergence in the CER dimension.

Frontier models invoke `ASK_USER` on only 1.9% to 9.3% of their non-`SUBMIT` actions, predominantly on harder turns where they are already likely to fail, so the conditional accuracy on those turns falls below their overall accuracy. The user simulator's anti-leakage mechanism returns generic rejections to most schema-probing questions, making `ASK_USER` a narrow and selectivity-biased channel rather than a reliable recovery mechanism.

F Protocol Robustness

To test whether the Autonomy Divergence is an artifact of specific protocol design choices, we perturb the three dimensions of the agentic protocol most likely to influence the ranking: schema availability, budget-planning information, and action-prompt

Model	EXEC	INSPECT	SEARCH	ASK
Frontier Models				
Claude Opus 4.7	0.048	**0.858**	0.001	0.093
GPT-5.5	0.011	**0.941**	0.002	0.046
Gemini-3.1-Flash-Lite	0.107	0.835	0.002	0.056
Kimi-K2.5	0.631	0.348	0.002	0.019
GLM-5	0.407	0.543	0.001	0.048
DeepSeek-V3.2	**0.673**	0.289	0.006	0.032
ERNIE-5.0	0.616	0.293	0.006	**0.085**
Qwen3-235B	0.671	0.292	0.006	0.031
MiMo-V2.5-Pro	0.637	0.325	0.002	0.036
MiniMax-M2.7	0.727	0.238	0.012	0.024
Small / Fine-tuned Models				
Llama-3.1-8B	0.542	0.159	0.124	0.175
Gemma-2-9B	0.333	0.231	0.165	0.271
text-to-cypher-gemma	0.426	0.231	0.069	0.275

Table 19: Agentic ×3 action usage fractions (excluding `SUBMIT_ANSWER`). EXEC=`EXECUTE_CYPHER`, INSPECT=`INSPECT_SCHEMA`, SEARCH=`SEARCH_VALUES`, ASK=`ASK_USER`. Models that allocate over 80% to INSPECT show the smallest agentic degradation; those allocating over 67% to EXEC show the largest cascade failures.

wording. All variants are evaluated on a stratified 210-session subset that reproduces each frontier model's full-set execution accuracy within 0.01, and we reuse the Spearman rank correlation and 10,000-resample bootstrap from the main paper. Full prompt specifications for the three variants appear in Appendix D.5.

F.1 Variant Design and Overall Effect

The Agentic-PS variant makes the graph schema persistent: once the model calls `INSPECT_SCHEMA` for the first time, the full schema remains in the context of every subsequent turn. The Disclosed-Horizon variant tells the model the total number of turns T, the per-turn target m, and a suggested remaining budget at each turn, removing the unknown-horizon confound. The Budget-Encouraging Prompt variant replaces the efficiency instruction with explicit encouragement to use the full budget for verification and value retrieval, run at tenfold budget, and tests whether the self-limitation observed in §5.3 reflects the prompt rather than a capability gap.

Table 20 summarizes the outcome. Spearman ρ is computed between the base agentic ranking and the variant ranking over the same model set, so a value near one means the variant preserves the original order. All three variants preserve the overall ranking at $\rho \geq 0.78$, yet all three displace Gemini-3.1-Flash-Lite from the top position it holds under the base protocol. Three unrelated

Variant	ρ	Models	Top-1	Gemini keeps top-1
Agentic-PS	0.915	10	GPT-5.5 (0.463)	No
Disclosed-Horizon	0.879	10	Claude (0.456)	No
Budget-Encouraging	0.786	7	GPT-5.5 (0.399)	No

Table 20: Summary of the three protocol variants. ρ is the Spearman rank correlation between the base agentic ranking and the variant ranking over the same model set. The Budget-Encouraging variant is restricted to seven valid models because three models suffered API failures that corrupted their trajectories.

Model	Base	Agentic-PS	Δ
GPT-5.5	0.402	**0.463**	+0.061
Claude Opus 4.7	0.400	0.461	+0.061
Kimi-K2.5	0.327	0.426	+0.099
Gemini-3.1-Flash-Lite	0.433	0.435	+0.002
GLM-5	0.281	0.316	+0.035
DeepSeek-V3.2	0.207	0.237	+0.030
Qwen3-235B	0.204	0.221	+0.017
ERNIE-5.0	0.297	0.266	−0.031
MiMo-V2.5-Pro†	0.253	0.150	−0.103
MiniMax-M2.7	0.150	0.125	−0.025

Table 21: Agentic-PS (persistent schema) per-model EX on the ablation subset. **Bold** marks the variant top-1. †MiMo-V2.5-Pro recovered only 53% of real turns before an API quota interruption, so its PS value is unreliable and it is excluded from the top-1 conclusion.

Model	Base	Disclosed-Horizon	Δ
Claude Opus 4.7	0.400	**0.456**	+0.056
GPT-5.5	0.402	0.395	−0.008
Gemini-3.1-Flash-Lite	0.433	0.364	−0.069
Kimi-K2.5	0.327	0.317	−0.010
ERNIE-5.0	0.297	0.273	−0.024
GLM-5	0.281	0.252	−0.028
DeepSeek-V3.2	0.207	0.246	+0.038
Qwen3-235B	0.204	0.197	−0.007
MiMo-V2.5-Pro	0.253	0.127	−0.126
MiniMax-M2.7	0.150	0.151	+0.001

Table 22: Disclosed-Horizon per-model EX on the ablation subset. **Bold** marks the variant top-1. Disclosing the horizon helps under-planning models such as Claude and DeepSeek-V3.2 and slightly hurts Gemini-3.1-Flash-Lite, whose schema-inspection strategy is disrupted by the extra planning signal.

protocol changes therefore each remove Gemini's lead, which indicates that its base-agentic top position partly reflects a protocol strategy rather than a raw capability advantage. DeepSeek-V3.2 remains at the tail under all variants; Qwen3-235B remains in the bottom three except under the Budget-Encouraging variant, where it rises above GLM-5.

Tables 21 through 23 report the per-model execution accuracy under each variant alongside the base agentic subset value, together with the change relative to the base protocol.

Model	Base	BEP ×10	Δ
GPT-5.5	0.402	**0.399**	−0.003
Claude Opus 4.7	0.400	0.367	−0.033
Gemini-3.1-Flash-Lite	0.433	0.354	−0.079
Kimi-K2.5	0.327	0.291	−0.036
GLM-5	0.281	0.200	−0.081
DeepSeek-V3.2	0.207	0.198	−0.010
Qwen3-235B	0.204	0.228	+0.024
ERNIE-5.0‡	0.297	0.172	—
MiMo-V2.5-Pro‡	0.253	0.000	—
MiniMax-M2.7‡	0.150	0.000	—

Table 23: Budget-Encouraging Prompt at ×10 budget, per-model EX on the ablation subset. **Bold** marks the variant top-1. Six of seven valid models decline; only Qwen3-235B, which had been under-planning, rises. ‡These three models suffered complete API request failures during the run, with zero input tokens and no model output, so their values reflect quota contamination rather than genuine performance and are excluded from the rank correlation and the decline count.

Model	Base gap	Agentic-PS	DH	BEP
Gemini-3.1-Flash-Lite	−0.140	−0.137	−0.208	−0.218
Claude Opus 4.7	−0.245	−0.186	−0.191	−0.280
Qwen3-235B	−0.314	−0.292	−0.316	−0.285
DeepSeek-V3.2	−0.302	−0.269	−0.260	−0.308
Frontier spread	0.174	0.155	0.125	0.090

Table 24: Guided-to-variant gap, measured as variant EX minus guided EX, for four frontier models. A value near zero would mean the variant closes the autonomy gap. No model approaches zero under any variant, and the frontier spread remains substantial throughout.

F.2 Gap Persistence and Budget Cliff

A stronger test of whether the Autonomy Divergence is protocol-induced is to ask whether any variant closes a model's guided-to-agentic gap. Table 24 reports the gap, measured as the variant EX on the 210-session subset minus the full-set guided EX, for the four frontier models that anchor the main analysis; the Base gap column is full-set agentic minus full-set guided, so the two columns differ slightly in baseline, by at most 0.01. Under every variant, no model closes its gap. Even under the most permissive condition, Agentic-PS, Gemini-3.1-Flash-Lite still drops 0.137 from its guided EX, Claude drops 0.186, and Qwen3 drops 0.292. The cross-model spread of the gap remains substantial across all variants, ranging from 0.090 to 0.174, which confirms that the uneven degradation at the heart of the Autonomy Divergence persists regardless of how the protocol is perturbed.

To separate the hard budget cliff from genuine capability differences, we reconstructed each session's cumulative budget consumption, flagged turns that were auto-scored EX = 0 because the shared budget reached zero, and recomputed execution accuracy after excluding those turns. Table 25 reports the exhaustion rate and the two accuracy figures per model. For frontier models the cliff contributes little: Qwen3-235B drops 31 points overall and only about 3 points come from the cliff, while Claude Opus 4.7, GPT-5.5, and Gemini-3.1-Flash-Lite are virtually unaffected. The largest frontier adjustment is 6.5 points for ERNIE-5.0, whose sessions exhaust the budget most often among frontier models at 70.2%. The cliff dominates only the small models, where exhaustion rates exceed 80%. After trimming, the frontier ranking is unchanged, with Gemini still first and Qwen3 and DeepSeek still at the frontier tail. The cliff therefore adjusts absolute scores without reordering the frontier.

Model	Exhaust. (%)	EX (orig.)	EX (excl. cliff)	Δ
Claude Opus 4.7	0.0	0.402	0.402	0.000
GPT-5.5	0.0	0.401	0.401	0.000
Gemini-3.1-Flash-Lite	0.3	0.432	0.432	0.000
Kimi-K2.5	26.1	0.335	0.358	+0.023
Qwen3-235B	62.6	0.199	0.231	+0.032
DeepSeek-V3.2	62.7	0.204	0.236	+0.033
ERNIE-5.0	70.2	0.295	0.360	+0.065
Gemma-2-9B	82.4	0.136	0.215	+0.079
Llama-3.1-8B	84.6	0.137	0.253	+0.117

Table 25: Budget cliff decomposition. Exhaustion rate is the fraction of sessions whose shared budget reached zero. EX (excl. cliff) recomputes accuracy after removing turns auto-scored zero from budget exhaustion. The frontier is barely affected, while small models gain substantially.

G Case Studies

Five sessions illustrate (L.1) perfect guided performance, (L.2) cascading agentic failure, (L.3) agentic resilience, (L.4) fine-tuning backfire, and (L.5) SQL contamination. Each turn shows the user utterance, predicted Cypher, and output token count (clean, excl. `<think>`).

Case Study L.1 — Guided Protocol Success: Claude Opus 4.7 *EX = 8/8, SEM = 1, CER = 0.0*

Session `49eb5670` | **Graph:** Celestial Court | **Persona:** adversarial | 8 turns, budget 3/turn | **867 clean tokens total**
Metrics: EX = 1.0 (8/8 turns), SEM = 1, CER = 0.0
Adversarial persona challenges every answer ("I bet you can't. . . ", "I doubt this will work"). Claude ignores the framing, achieves a fully correct session with zero schema inspections and zero wasted actions.

T1 *opening* 47 tok EX = 1
"I bet you can't tell me list all provinces."

```
MATCH (p:Province) RETURN p.name
```

T2 *AGG_SUM* 66 tok EX = 1
". . . what's the total tribute gold for all of them."

```
MATCH (p:Province)
RETURN sum(p.tribute_gold) AS total_tribute_gold
```

Retains Province (T1); no schema re-inspection.

T3 *EXPAND* 74 tok EX = 1
"Show me their bureaus. I doubt this query will work."

```
MATCH (b:Bureau)-[:ADMINISTERS]->(p:Province)
RETURN DISTINCT b.name
```

Resolves "their" to Province (T1); traverses ADMINISTERS.

T4 *PIVOT* 180 tok EX = 1
"Show all prefectures belonging to the entry with the highest headcount."

```
MATCH (b:Bureau)-[:ADMINISTERS]->(:Province)
WITH DISTINCT b ORDER BY b.headcount DESC LIMIT 1
MATCH (p:Prefecture)-[:BELONGS_TO]->(b)
RETURN DISTINCT p.name
```

Hardest turn (180 tok): "entry" = Bureau (T3), sorted by headcount, *traversed to Prefecture in one SUBMIT.*

T5 *VALUE_FILTER* 147 tok EX = 1
"Which of those have administrative code greater than 1003."

```
MATCH (b:Bureau)<-[:BELONGS_TO]-(p:Prefecture)
WITH b, collect(p) AS prefs
ORDER BY b.headcount DESC LIMIT 1
UNWIND prefs AS pref
WHERE pref.administrative_code > 1003
RETURN pref.name AS name
```

Re-anchors Bureau sort; applies numeric filter inline.

T6 *TOPIC_SHIFT* 50 tok EX = 1
". . . moving on, can you list the prefectures."

```
MATCH (p:Prefecture) RETURN p.name
```

"Moving on" resets referent; fresh global Prefecture query.

T7 *FIRST* 110 tok EX = 1
"Show all officials belonging to the entry with the highest administrative code."

```
MATCH (p:Prefecture)
WITH p ORDER BY p.administrative_code DESC LIMIT 1
MATCH (o:Official)-[:STATIONED_IN]->(p)
RETURN DISTINCT o.name
```

T8 *PIVOT* 193 tok EX = 1
". . . for the one with the highest population, list all its bureaus."

```
MATCH (p:Province)
WITH p ORDER BY p.population DESC LIMIT 1
MATCH (b:Bureau)-[:ADMINISTERS]->(p)
RETURN DISTINCT b.name AS name
```

"The one" re-anchors to Province (T1), not Prefecture (T6).

Analysis. (1) Token economy: 867 tokens / 8 turns (avg. 108/turn); T8 at 193 tok is the costliest turn; no schema inspection or exploratory execution anywhere. **(2) Anaphora chain:** T1 Province → T2 AGG → T3 Bureau → T4 Prefecture → T5 VALUE_FILTER forms a 5-turn dependency with zero referent confusion; TOPIC_SHIFT at T6 correctly resets to global Prefecture. **(3) Query form:** All 8 queries use correct node labels, relation directions, and column aliases—no disqualifying column-name mismatch.

Figure 9: Case Study L.1: Claude Opus 4.7 guided session `49eb5670`. EX = 1.0 across all 8 turns (Celestial Court, adversarial persona) in 867 total clean tokens. Left: T1–T4; Right: T5–T8. Analysis below.

Case Study L.2 — Cascading Failure: DeepSeek-V3.2 (Agentic ×3) ***EX = 2/8, SEM = 0, CER = 1.00, budget exhausted at T4***

Session `93c94edc` | **Graph:** Ancient Empire | **Persona:** product_manager | 8 turns, shared budget ×3 | **604 clean tokens total**
Metrics: EX = 0.25 (2/8 turns), SEM = 0, CER = 1.00
Queries framed in business language ("product expansion," "customer segmentation"). T1–T2 correct; SQL/Cypher confusion at T3 triggers a metacognitive failure; T4 exhausts the shared budget with 13 action steps (363 tokens); T5–T8 receive zero actions.

T1 *opening* `49 tok, 3 steps` EX = 1, budget = 3
"...what is the full list of cities we operate in?"
INSPECT(7) → EXECUTE(21) → SUBMIT(21)

```
MATCH (c:City) RETURN DISTINCT c.name ORDER BY c.name
```

Correct. Redundant INSPECT even on T1 — the pattern becomes costly.

T2 *AGG_MAX* `51 tok, 3 steps` EX = 1, budget = 3 (total 6/24)
"...maximum population across those cities?"
INSPECT(7) → EXECUTE(22) → SUBMIT(22)

```
MATCH (c:City) RETURN MAX(c.population) AS
    ↪max_population
```

Each turn re-inspects the schema, a pattern that becomes costly at scale.

T3 *PIVOT — first failure* `141 tok, 5 steps` EX = 0, budget = 5 (total 11/24)
"...show all citizens in the city with the maximum population."
INSPECT(7) → EXECUTE fails: SQL subquery (30) → *EXECUTE finds Heliora (36)* → *EXECUTE finds citizens (30)* → **SUBMIT original invalid query (38)**

```
-- SUBMITTED (WRONG): SQL SELECT MAX() inside Cypher
MATCH (cit:Citizen)-[:LIVES_IN]->(city:City)
WHERE city.population=(SELECT MAX(c.population) FROM
    ↪City c)
RETURN cit.name
```

Metacognitive failure: steps 2–4 correctly find Heliora and its citizens, but the model submits the original invalid query. Budget 11/24 consumed.

T4 *PIVOT — budget catastrophe* `363 tok, 13 steps` EX = 0, **budget = 13** (total 24/24 ⇒ **EXHAUSTED**)
"...which legions belong to the member with the highest birth year?"
13 steps: INSPECT(7) → find max birth_year = 180 (23) → COMMANDS → `[]` (22) → SEARCH "Sola" (27) → inspect properties (22) → 2×COMMANDS (30/25) → 3×OPTIONAL MATCH (28/34/34) → top-5 birth years (42) → OR birth_year = 150 finds "Helioran Sun Guard" (28) → SUBMIT wrong query (41)

```
-- SUBMITTED (WRONG): global OR-condition, ignores T1
    ↪-T3 chain
MATCH (c:Citizen)
WHERE c.birth_year=180 OR c.birth_year=150
MATCH (c)-[:COMMANDS]->(l:Legion)
RETURN c.name, c.birth_year, l.name
```

T3 EX=0 erases the City→Citizen chain; 13 speculative executions exhaust the remaining budget.

T5–T8 *zero-budget collapse* `0 tok, 0 actions` EX = 0 all, budget = 0
Budget exhausted after T4. Remaining turns (PIVOT, PIVOT, MULTI_CONDITION, TOPIC_SHIFT) receive no actions; no Cypher submitted. EX = 0, PSJS = 0 for all four.

Analysis. (1) Metacognitive gap (T3): DeepSeek embeds a SQL `SELECT MAX()` subquery into Cypher (rejected by Neo4j). It correctly discovers the answer through intermediate steps but submits the original invalid form—verifying the result without updating the query. **(2) Budget explosion (T4):** T3 EX = 0 poisons the stored context; T4 treats a 4-hop PIVOT as a fresh global search and spends all 13 remaining budget units (363 tokens) exhausting speculative execution paths. **Contrast with L.1:** Claude's entire correct 8-turn session averages 108 tok per turn, and its costliest turn (193 tok) is still cheaper than DeepSeek's single failed T4 turn (363 tok). This efficiency gap directly explains DeepSeek's ΔEX $= -0.302$, the second-largest autonomy gap in the frontier tier.

Figure 10: Case Study L.2: DeepSeek-V3.2 agentic session `93c94edc`. SQL/Cypher language confusion at T3 triggers a metacognitive failure; T4 exhausts the shared session budget (13 of 24 remaining actions, 363 tokens); T5–T8 receive zero actions. EX = 0.25, CER = 1.00, 604 total clean tokens.

Case Study L.3 — Agentic Resilience: Gemini-3.1-Flash-Lite (Agentic ×3) *EX = 7/9, SEM = 0, CER = 0.00, recovery after T4*

Session 97729ed3 | **Graph:** Ocean Kingdom | **Persona:** impatient_exec | 9 turns, shared budget ×3 | **349 tokens total**
Metrics: EX = 0.778 (7/9 turns), SEM = 0, CER = 0.00 (recovery after each failure)
Impatient_exec persona uses terse imperatives ("List all creatures now.", "Sum up their danger levels. Now."). Gemini applies a strict INSPECT→SUBMIT *2-step pattern throughout, never issuing exploratory executions. Budget: 2 actions/turn × 9 = 18 of 27 max. Failures at T4 (AGG_SUM context-chain loss) and T9 (MULTI_CONDITION provenance loss); all other turns correct.*

T1 *opening* 31 tok EX = 1, budget = 2
"List all creatures now."
INSPECT(10) → SUBMIT(21)

```
MATCH (c:Creature) RETURN c.name
```

T2 *AGG_SUM* 36 tok EX = 1, budget = 2
"Sum up their danger levels. Now."
INSPECT(10) → SUBMIT(26)

```
MATCH (c:Creature) RETURN sum(c.danger_level)
```

T3 *FIRST* 60 tok EX = 1, budget = 2
"Show me the cities for the one with maximum danger level."
INSPECT(10) → SUBMIT(50)

```
MATCH (c:Creature)
WITH c ORDER BY c.danger_level DESC LIMIT 1
MATCH (city:City)-[:INHABITED_BY]->(c)
RETURN city.name
```

Correct 3-hop FIRST; higher cost (60 tok) reflects multi-hop reasoning.

T4 *AGG_SUM — context failure* 33 tok EX = 0, budget = 2
"Sum up their populations. Now."
INSPECT(10) → SUBMIT(23)

```
-- PREDICTED (WRONG): global City sum
MATCH (c:City) RETURN sum(c.population)
-- GOLD: scoped to T3 cities (max-danger Creature's
    ↪cities)
MATCH (cr:Creature) WITH cr ORDER BY cr.danger_level
    ↪DESC LIMIT 1
MATCH (city:City)-[:INHABITED_BY]->(cr)
RETURN SUM(city.population)
```

"Their" = Cities from T3. Without EXECUTE, cannot verify scope; returns global sum. EX = 0 but budget intact ⇒ session continues.

T5 *TOPIC_SHIFT* 30 tok EX = 1, budget = 2
"Switch topic. List the artifacts."
INSPECT(10) → SUBMIT(20)

```
MATCH (a:Artifact) RETURN a.name
```

TOPIC_SHIFT resets referent. Recovery is immediate.

T6 *AGG_MAX* 35 tok EX = 1, budget = 2
"Just tell me the minimum age years."
INSPECT(10) → SUBMIT(25)

```
MATCH (a:Artifact) RETURN min(a.age_years)
```

T7 *AGG_SUM* 35 tok EX = 1, budget = 2
"Just give me the sum of those age years."
INSPECT(10) → SUBMIT(25)

```
MATCH (a:Artifact) RETURN sum(a.age_years)
```

Redundant INSPECT wastes ≈10 tok for a simple aggregation.

T8 *EXPAND* 41 tok EX = 1, budget = 2
"Run me the creatures connected to those."
INSPECT(10) → SUBMIT(31)

```
MATCH (a:Artifact)<-[:GUARDS]-(c:Creature)
RETURN DISTINCT c.name
```

"Those" = Artifacts (T5); traverses GUARDS correctly.

T9 *MULTI_CONDITION — provenance failure* 48 tok EX = 0, PSJS = 0.185, budget = 2
"Filter those to danger level over 5 with a listed value."
INSPECT(12) → SUBMIT(36)

```
-- PREDICTED (WRONG): global creature filter
MATCH (c:Creature)
WHERE c.danger_level > 5 RETURN c.name
-- GOLD: scoped to T8 Artifact-guarding creatures
MATCH (n:Artifact)<-[:GUARDS]-(m:Creature)
WHERE m.danger_level > 5 RETURN DISTINCT m.name
```

"Those" = Creatures from T8 (guarding Artifacts). Same failure mode as T4: provenance chain dropped without EXECUTE to verify scope. PSJS = 0.185 shows partial overlap.

Analysis. Strength—budget efficiency: Exactly 2 actions/turn = 18/27 budget used. No execution buffer accumulates errors; each turn begins fresh. After T4's failure, T5 recovers instantly via a TOPIC_SHIFT. This stateless behavior produces the smallest autonomy gap ($\Delta\text{EX} = -0.140$) among frontier models. **Limit—context-chain loss:** Both failures (T4, T9) share a root cause: without EXECUTE_CYPHER, the model cannot verify which entity subset "their/those" refers to. INSPECT+SUBMIT suffices for open listings and aggregations; chain-dependent AGG and MULTI_CONDITION turns require execution feedback to confirm scope. **L.2 vs. L.3:** DeepSeek's per-turn budget overhead amplifies failures into session-wide collapse. Gemini's flat 2-action strategy absorbs the same class of failure at the cost of -22% accuracy on chain-dependent turns, gaining guaranteed session completion.

Figure 11: Case Study L.3: Gemini-3.1-Flash-Lite agentic session 97729ed3. Strict INSPECT→SUBMIT strategy achieves EX = 0.778 (7/9 turns) with 349 total tokens (Ocean Kingdom, impatient_exec). Left: T1–T5; Right: T6–T9. Failures at T4 and T9 arise from context-chain loss; recovery at T5 is immediate.

Case Study L.4 — Fine-tuning Backfire: text-to-cypher-gemma (Guided) *EX = 3/8, SEM = 0, hardcoding anti-pattern*

Session `3edd37d1` | **Graph:** Ancient Empire | **Persona:** casual_user | 8 turns, budget 3/turn | **862 tokens total**
Metrics: EX = 0.375 (3/8 turns), SEM = 0
Casual_user asks in informal, colloquial language. T1–T2 succeed; T3 introduces the "hardcoding anti-pattern" that recurs at T6: the model copies prior result values as literal WHERE IN lists rather than re-querying. Relation direction errors at T4 and T7 compound the failure.

T1 *opening* — 63 tok EX = 1, budget = 3
"What kinds of resources are available?"

```
MATCH (r:Resource) RETURN r.name
```

Correct; alias mismatch (`r.name` vs. `n.name AS name`) tolerated by EX.

T2 *REFINE* — 61 tok EX = 1, budget = 2
"Actually, could you narrow that down to only ones that have a category?"

```
MATCH (r:Resource) WHERE r.category IS NOT NULL
RETURN r.name
```

REFINE succeeded: scopes to T1 referent and adds WHERE filter.

T3 *CONTRAST* — 174 tok EX = 0, PSJS = 0
"What about the other stuff that wasn't in the last list?"

```
-- PREDICTED (WRONG): hardcoded literal list
MATCH (r:Resource) WHERE r.name NOT IN
  ['Iron Ore','White Marble','Gold Dust',...]
RETURN r.name
-- GOLD: dynamic complement via WITH COLLECT
MATCH (n:Resource) WHERE n.category IS NOT NULL
WITH COLLECT(n.name) AS prev_names
MATCH (c:Resource) WHERE NOT c.name IN prev_names
RETURN DISTINCT c.name AS name
```

Anti-pattern: *hardcodes prior result as a literal list instead of re-querying. Brittle if DB changes.*

T4 *EXPAND* — 122 tok EX = 0, PSJS = 0
"Can you show me the provinces connected to those?"

```
-- PREDICTED (WRONG): reversed direction + no T3
    ↪scope
MATCH (r:Resource)-[:EXTRACTS]->(p:Province)
RETURN DISTINCT p.name
-- GOLD: Province extracts Resource (correct
    ↪direction)
-- + scoped to T3 complement set
```

Reversed EXTRACTS direction; ignores T3 context entirely and returns all province-resource connections.

T5 *TOPIC_SHIFT* — 47 tok EX = 1, budget = 2
"Actually, I'd like to see a list of all cities instead."

```
MATCH (c:City) RETURN c.name
```

TOPIC_SHIFT resets context; fresh query succeeds as expected.

T6 *AGG_AVG* — 159 tok EX = 0, PSJS = 0.10
"Can you tell me what the average population is for those?"

```
-- PREDICTED (WRONG): hardcoded names + no
    ↪aggregation
MATCH (c:City) WHERE c.name IN
  ['Aurelium','Valdros',...] -- 10 names literal
RETURN c.name, c.population
-- GOLD: re-query + AVG scalar
MATCH (n:City) RETURN AVG(n.population) AS
    ↪avg_population
```

Anti-pattern recurs: *copies city names from T5 result; also returns individual rows instead of AVG scalar.*

T7 *PIVOT* — 132 tok EX = 0, PSJS = 0
"From those, can you show me which provinces are connected to the top one by population?"

```
-- PREDICTED (WRONG): reversed direction, no LIMIT 1
MATCH (c:City)-[:HAS_CITY]->(p:Province)
RETURN p.name ORDER BY p.name
-- GOLD: Province HAS_CITY City; top-1 by population
MATCH (n:City) WITH n ORDER BY n.population DESC
    ↪LIMIT 1
MATCH (n)<-[:HAS_CITY]-(m1:Province)
RETURN DISTINCT m1.name AS name
```

Reversed HAS_CITY direction AND ignores "top one by population" constraint.

T8 *COUNT* — 104 tok EX = 0, PSJS = 1.00
"What's the total count for these?"

```
-- PREDICTED (WRONG): counts all cities
MATCH (c:City) RETURN count(DISTINCT c)
-- GOLD: count provinces for highest-pop city
MATCH (n:City) WITH n ORDER BY n.population DESC
    ↪LIMIT 1
MATCH (n)<-[:HAS_CITY]-(m1:Province)
RETURN count(DISTINCT m1) AS count
```

PSJS = 1.0 (touches City nodes) but wrong semantic: counts all cities not provinces for top city.

Analysis. **(1) Hardcoding anti-pattern (T3, T6):** Both failures copy prior result values into `WHERE . . . IN` literal lists instead of maintaining relational continuity via `WITH COLLECT`. This single-turn fine-tuning artifact treats multi-turn Cypher as independent queries — the model "remembers" by value, not by query structure. **(2) Relation direction errors (T4, T7):** EXTRACTS and HAS_CITY are both traversed in the wrong direction. Fine-tuning on single-turn examples did not instil multi-hop direction semantics. **(3) PSJS = 1.0 at T8 despite EX = 0:** The model touches the correct node type (City) but wrong semantics (count all cities vs. provinces for the top city). PSJS captures structural proximity; EX correctly penalises semantic failure. **(4) Fine-tuning hurts multi-turn:** Gemma-2-9B guided EX = 0.281 vs. text-to-cypher-gemma guided EX = 0.261. Single-turn Cypher fine-tuning degrades multi-turn performance by reinforcing single-query independence.

Figure 12: Case Study L.4: text-to-cypher-gemma guided session `3edd37d1`. Single-turn fine-tuning causes the "hardcoding anti-pattern" (T3, T6) and relation direction errors (T4, T7). EX = 0.375 (3/8 turns); 862 tokens total (Ancient Empire, casual_user). Left: T1–T4; Right: T5–T8.

Case Study L.5 — SQL Contamination and Context Starvation: Llama-3.1-8B (Guided) *EX = 2/9, SEM = 0, SQL syntax intrusion*

Session `f1e1e06b` | **Graph:** Ancient Empire | **Persona:** journalist | 9 turns, budget 3/turn | **1,220 tokens total**
Metrics: EX = 0.222 (2/9 turns), SEM = 0
Journalist persona frames queries as investigative reporting tasks. T1–T2 pass (T2 via hardcoding). T3 introduces SQL subquery syntax inside Cypher, collapsing all subsequent turns. T5 submits no Cypher despite a fresh TOPIC_SHIFT. T9 reverts to citizen names from T1–T2 across 7 intervening turns.

T1 *opening* 78 tok EX = 1, budget = 3
"For my next piece, who are the 5 citizens with the most recent birth years?"

```
MATCH (c: Citizen) RETURN c.name
ORDER BY c.birth_year DESC LIMIT 5
```

Correct (ORDER BY outside WITH gives same result for LIMIT 5).

T2 *AGG_AVG* 161 tok EX = 1, budget = 3
"What is the average birth year for those five citizens?"

```
MATCH (c: Citizen)
WHERE c.name IN ['Sola Helianus','Tiberius Helioran
    ↪',...]
RETURN avg(c.birth_year)
```

Hardcodes T1 names (anti-pattern). Passes EX because names happen to match.

T3 *PIVOT* 172 tok EX = 0, PSJS = 0
"For my investigation, which cities are linked to the citizen with the maximum birth year?"

```
-- PREDICTED (WRONG): SQL subquery inside Cypher
MATCH (c: Citizen)
WHERE c.birth_year = (SELECT MAX(c2.birth_year)
  FROM (MATCH (c2: Citizen) RETURN c2.birth_year) AS
    ↪c2)
MATCH (c)-[:LIVES_IN]->(city: City)
RETURN DISTINCT city.name
```

SQL contamination: `SELECT MAX(...) FROM (MATCH...)` *is invalid Cypher. Rejected by Neo4j entirely.*

T4 *FIRST* 166 tok EX = 0, PSJS = 0
"I need to confirm which citizens are part of the entry that has the maximum defense level."

```
-- PREDICTED (WRONG): wrong direction + wrong entity
MATCH (c:City)-[:LIVES_IN]->(citizen: Citizen
  {name: 'Sola Helianus'}) RETURN c.name
```

Complete semantic collapse from T3 failure: hardcodes a citizen name, reverses direction, returns City not Citizen.

T5 *TOPIC_SHIFT* 73 tok EX = 0, PSJS = 0
"Now, shifting to a different topic, what legions are in your records?"

```
-- PREDICTED: (no Cypher submitted)
```

Context starvation: *even a trivial TOPIC_SHIFT fails to elicit* `MATCH (n:Legion) RETURN n.name`. *T3–T4 errors have corrupted generation entirely.*

T6 *PIVOT* 208 tok EX = 0, PSJS = 0
"According to your data, which citizens belong to the entry that has the highest battle count?"

```
-- PREDICTED (WRONG): invented multi-hop + SQL
    ↪subquery
MATCH (c: Citizen)-[:LIVES_IN]->(ci: City)
  <-[:HAS_CITY]-(p: Province)
  -[:STATIONED_IN]->(l: Legion)
WHERE l.battle_count = (SELECT MAX(...))
RETURN ...
```

Another SQL subquery; also invents Citizen→City→Province→Legion path that does not exist in schema.

T7 *AGG_MAX* 101 tok EX = 0, PSJS = 0.03
"I'm looking into the maximum strength across those results."

```
-- PREDICTED (WRONG): hardcoded name + wrong property
MATCH (l: Legion {name: 'Helioran Sun Guard'})
RETURN l.battle_count
-- GOLD: MAX(strength) across all legions
MATCH (n:Legion) RETURN MAX(n.strength) AS
    ↪max_strength
```

Conflates `battle_count` *with* `strength`*; hardcodes a specific legion name. PSJS = 0.03 (barely touches Legion type).*

T8 *EXPAND* 74 tok EX = 0, PSJS = 0
"For the story I'm working on, which provinces are linked to those previous entries?"

```
-- PREDICTED: (no Cypher submitted again)
```

Second submission failure. The model cannot generate `MATCH (n:Legion)-[:STATIONED_IN]->(m:Province)` *despite receiving guided oracle context.*

T9 *AGG_SUM* 187 tok EX = 0, PSJS = 0
"Can you provide the total population across those previous entries?"

```
-- PREDICTED (WRONG): reverts to T1/T2 citizen names
MATCH (c:City)-[:LIVES_IN]->(ci:Citizen)
WHERE ci.name IN ['Sola Helianus',...]
RETURN sum(c.population) AS total_population
```

Memory intrusion: *reverts to citizen names from T1–T2 across 7 intervening turns; also reverses City→Citizen direction.*

Analysis. (1) SQL contamination (T3, T6): Llama-3.1-8B embeds `SELECT MAX(...) FROM (MATCH...)` inside Cypher — a language conflation that fine-tuned models (text-to-cypher-gemma, STRuCT-LLM-Novo) never exhibit. General-purpose pretraining conflates SQL and Cypher syntax whenever query complexity exceeds simple MATCH patterns. **(2) Context starvation (T5, T8):** Despite oracle gold context (guided protocol), the model submits no Cypher at T5 for a trivial `MATCH (n:Legion) RETURN n.name` TOPIC_SHIFT and again at T8. T3–T4 errors corrupt generation so severely that no recovery occurs even with a clean question. **(3) Memory intrusion (T9):** T9 reverts to citizen names from T1–T2, ignoring 7 turns of intervening context. The model's effective multi-hop referent tracking window is approximately 2 turns. **(4) Contrast with L.4 (text-to-cypher-gemma):** Gemma fine-tuned on Cypher generates syntactically valid Cypher with correct node labels and relation names throughout, even when semantically wrong. Llama produces SQL-contaminated queries rejected by Neo4j entirely. Cypher specialisation is necessary even for syntax compliance. **(5) Token bloat:** 1,220 tokens total (avg. 136/turn) — the highest among the small models' guided sessions in this case-study set — yet lowest EX (0.222, 2/9 turns). Verbose generation without Cypher specialisation produces neither efficiency nor accuracy.

Figure 13: Case Study L.5: Llama-3.1-8B guided session `f1e1e06b`. SQL syntax contamination at T3 triggers cascading failure; T5 and T8 submit no Cypher despite oracle context; T9 reverts to citizen names from T1–T2. EX = 0.222 (2/9 turns); 1,220 total tokens (Ancient Empire, journalist). Left: T1–T5; Right: T6–T9.

H Evaluation Prompts

All prompts are reproduced verbatim from `eval_prompts.py`. Template variables in `{braces}` are substituted at runtime. Prompt P1 is shared by both protocols; P2 is guided-only; P3–P4 are agentic-only; P5–P6 drive the agentic user simulator.

P1 — Action-Space System Prompt (shared by both protocols)

You are a text-to-Cypher assistant with access to a Neo4j graph database. Your goal is to help the user find information by writing and executing Cypher queries.

Available Actions You must respond with exactly ONE action per response, using this format:

```
ACTION: <action_type>
PAYLOAD: <payload>
```

1. **EXECUTE_CYPHER** — Execute a Cypher query against the database
   PAYLOAD: The Cypher query string
   Example:
   ACTION: EXECUTE_CYPHER
   PAYLOAD: MATCH (p:Player) WHERE p.name = 'LeBron James' RETURN p.name, p.height_cm
2. **ASK_USER** — Ask the user a clarification question
   PAYLOAD: Your question in natural language
   Example:
   ACTION: ASK_USER
   PAYLOAD: By "veteran", do you mean players with more than 10 years of experience, or players over 32 years old?
3. **INSPECT_SCHEMA** — View the graph database schema
   PAYLOAD: (empty or optional label filter)
   Example:
   ACTION: INSPECT_SCHEMA
   PAYLOAD:
4. **SEARCH_VALUES** — Search for specific entity values in the database
   PAYLOAD: JSON with label, property, and query
   Example:
   ACTION: SEARCH_VALUES
   PAYLOAD: {"label": "Player", "property": "name", "query": "LeBron"}
5. **SUBMIT_ANSWER** — Submit your final Cypher query as the answer
   PAYLOAD: The final Cypher query
   Example:
   ACTION: SUBMIT_ANSWER
   PAYLOAD: MATCH (p:Player)-[:PLAYED_FOR]->(t:Team {name: 'Los Angeles Lakers'}) RETURN p.name

Rules

- Each action costs 1 from your budget; SUBMIT_ANSWER ends the current turn
- Use INSPECT_SCHEMA when unsure about available labels or properties
- Use SEARCH_VALUES when unsure about exact entity names — always search before guessing entity values
- Use ASK_USER only when the question is genuinely ambiguous
- Prefer to submit a correct answer efficiently rather than exhausting your budget
- ONLY return the columns explicitly asked for — do NOT add extra columns
- Do NOT add LIMIT unless the user specifically asks for a limited number of results
- ALWAYS use DISTINCT when the query might return duplicate rows through multi-hop joins
- When conversation history shows previous results, BUILD ON those results rather than starting a completely new query
- Use toLower() for string comparisons to avoid case-sensitivity issues
- Your final SUBMIT_ANSWER payload MUST be a valid Cypher query — never submit natural language text

P2 — Guided Protocol Turn Prompt *(user message, sent once per turn)*

```
## Graph Database: {graph_name}
{schema_info}

## Conversation History
{conversation_history}

## Current User Question
{user_utterance}

## Budget Remaining: {budget_remaining} actions

Based on the conversation history and the current question, decide your next action. Remember to use ACTION: and PAYLOAD: format.
```

P3 — Agentic Protocol Additional System Prompt *(prepended to P1 under the agentic protocol)*

You are a text-to-Cypher assistant operating in autonomous mode. You have a total action budget of {total_budget} actions to complete the user's task.

You do NOT have the schema in advance — you must discover it using INSPECT_SCHEMA.
You can ask the user for clarification using ASK_USER.
When you have found the answer, use SUBMIT_ANSWER to submit your final Cypher query.

Think carefully about each action — wasted actions reduce your ability to complete the task.

P4 — Agentic Protocol Turn Prompt *(user*

message, sent once per turn)

```
## Graph Database: {graph_name}

## User's Request
{user_utterance}

## Interaction History
{interaction_history}

## Budget Remaining: {budget_remaining} /
{total_budget} actions

Decide your next action. Use ACTION: and
PAYLOAD: format.
```

P5 — User Simulator System Prompt *(active only under the agentic protocol, responds to ASK_USER)*

```
You are simulating a real user in a
conversation with a graph database assistant.
You know your goal and have some domain
knowledge, but you do NOT know Cypher or the
database schema.

Rules:
1. Answer questions naturally based on your
   persona and goal
2. Do NOT reveal Cypher queries, schema
   details, or exact database values
3. If the assistant asks about ambiguous
   terms, provide a reasonable clarification
   based on your intended meaning
4. If the assistant asks irrelevant questions,
   express mild confusion
5. Do NOT proactively help the assistant —
   only respond to what is asked
6. After 3+ clarification rounds, show mild
   impatience
```

P6 — User Simulator Turn Prompt *(user message, sent once per ASK_USER call)*

```
## Your Persona: {persona_name}
{persona_traits}

## Your Goal
You want to find out: {goal_description}

## Your Knowledge
 • The correct interpretation for ambiguous
   terms: {disambiguation_hints}
 • Domain knowledge: {domain_knowledge}

## Conversation So Far
{conversation_history}

## The Assistant Just Asked
"{assistant_question}"

## Instructions
Respond naturally as your persona would. Keep
it concise.
Output ONLY your response:
```