MolQuest: Certified Four-Query Molecular Identification for Evaluating LLM Query Planning

Anonymous ACL submission

Abstract

Question-asking benchmarks are hard to interpret when the permitted questions do not uniquely distinguish the candidate answers. MolQuest removes this ambiguity. Each episode contains 16 anonymized ChEBI molecules and 53 deterministic RDKit predicates. Before release, every candidate pool is certified by exact search: one adaptive policy, chosen without access to the target's identity, identifies all 16 targets in four binary questions, the information-theoretic optimum.

The benchmark contains 213 certified decision problems and 3,408 target episodes. On a fixed 100-episode witness-balanced stress test, an exact policy isolates 100 targets within four questions and a greedy maximum-information-gain policy isolates 93. GPT-5.4-mini isolates only 9 when given enough information to reconstruct the complete candidate-predicate table. Across a common fallback-free subset, GPT, Llama, and Qwen runs acquire 2.22–2.42 of the four bits needed for identification after four questions, compared with 3.92 bits for the greedy policy. Across three runs of a descriptor-plus-SMILES ("hybrid") interface, GPT isolates 34/300 when episodes involving a harness substitution are counted as failures. The gap therefore persists after ruling out benchmark infeasibility and removing the need to infer structural predicate outcomes from SMILES. MolQuest provides a controlled test of query planning rather than another molecular property-prediction task.

1 Introduction

An agent that lacks a crucial fact should ask for it. This simple requirement appears in diagnosis, debugging, search, and scientific reasoning, but it is difficult to evaluate cleanly. A failed interaction may reflect poor question selection, an inconsistent answerer, or a candidate set that the available questions could never distinguish.

Twenty-questions games offer a compact test of active information gathering. Existing work has used them to study multi-turn language-model planning (Zhang et al., 2024) and visually grounded dialogue (de Vries et al., 2017); related clinical benchmarks evaluate whether models request missing evidence before answering (Li et al., 2024). These settings are useful, but their free-form question languages and lack of an exact instance-level feasibility certificate make failures difficult to localize.

MolQuest uses molecular structures to construct a finite, auditable alternative. A player sees 16 anonymous candidates, selects binary questions from a public inventory, observes exact answers and eliminations, and tries to isolate a hidden target. The molecular domain supplies reproducible predicates with clear semantics, while the resulting problem is a binary decision tree over a frozen answer table. Here **frozen** means that all descriptor and predicate values are computed once with the pinned RDKit pipeline and reused unchanged across targets, models, and runs.

MolQuest contributes 213 certified molecular decision problems, a strict-identification metric that does not reward guesses among survivors, and exact and greedy algorithmic baselines. Against these controls, a language model remains far from optimal even when all predicate outcomes are directly available or exactly recoverable from visible fields.

The central claim is deliberately narrow: on these certified instances, low strict-identification rates cannot be attributed to an infeasible question language.

2 Task and Certification

A **pool**, or decision problem, comprises a 16-candidate set C, its legal binary predicates Q, and their frozen answer matrix. An **episode**, also called a game, pairs a pool with one hidden target $t \in C$.

At turn k, the player chooses an unasked predicate q, observes $q(t)$, and removes candidates whose predicate value disagrees with the answer. The primary budget is four questions.

Consider a late episode state with survivors `M03` and `M11` and four unused predicates, all legal in the original 16-candidate pool:

Predicate	M03	M11
Contains N	yes	yes
Has amide	yes	no
At least 2 rings	no	no
At least 4 H-bond acceptors	yes	no

If the target is `M03`, "contains N?" returns *yes* but eliminates neither candidate; "has amide?" returns *yes* and isolates `M03`. The other questions need not be asked. This is a two-survivor state, not a separate two-candidate pool.

We report two outcomes. A **correct final guess** names the target, possibly with several candidates still viable. **Strict identification** requires the transcript to leave exactly one candidate and for that candidate to be the target. Strict identification is the primary metric; it does not reward guesses among survivors.

2.1 Exact certificate

For each pool, exact search constructs a single adaptive policy tree without access to the target. **Target-blind** means that question selection may use the candidate table and previous answers, but not the hidden target's identity or any unobserved answer. Let $F(S, d)$ indicate whether survivor set S can be resolved in at most d further questions. The base case is $|S| = 1$. Otherwise, the search considers each predicate that splits S and accepts it only if both answer branches satisfy $F(\cdot, d-1)$. States with $|S| > 2^d$ are pruned. The released tree is replayed against every target and checked to terminate at the correct leaf.

Four questions are not an arbitrary threshold. A binary tree of depth three has at most eight leaves, so identifying one of 16 candidates requires at least four questions in the worst case. Every MolQuest pool has a depth-four tree and therefore attains this lower bound. The certification problem is a small exact instance of optimal binary decision-tree construction, whose general form is computationally hard (Hyafil and Rivest, 1976).

2.2 Greedy baseline

We also evaluate **greedy information gain**, the one-step generalized binary search rule (Nowak, 2011). At each turn it chooses the unasked predicate with maximum binary entropy over the current survivors, with deterministic lexicographic tie-breaking. This policy is target-blind and uses the same frozen table as the exact search, but it has no lookahead. It provides a more informative baseline than the exact certificate alone: it tests whether local balanced splitting is already sufficient on the released pools.

3 Benchmark Construction

MolQuest uses the fully curated three-star subset from ChEBI database snapshot 253 (Hastings et al., 2016). ChEBI is a curated database of biologically relevant chemical entities. RDKit 2026.03.4, an open-source cheminformatics toolkit (RDKit Contributors, 2026), parses and sanitizes the source records, removes explicit hydrogens, and produces canonical isomeric SMILES strings, a line notation for molecular graphs. Of 52,932 SDF records, 52,835 pass this pipeline. We retain source records by ChEBI identifier; we do not impose neutral-only, single-component, organic-only, or drug-like filters.

The question inventory contains 53 RDKit predicates: elemental presence; common substructures expressed in SMARTS, a query language for molecular subgraph matching; aromatic, heterocyclic, and fused-ring indicators; and thresholds over molecular weight, logP (a lipophilicity estimate), ring count, heteroatom count, hydrogen-bond donors and acceptors, rotatable bonds, and carbon sp^3 fraction (the fraction of carbon atoms with sp^3 hybridization). Exact SMARTS patterns and descriptor calls are part of the release. A predicate is legal for a pool when it is nonconstant and both initial answer bins contain at least two candidates.

Candidate pools are proposed within stereochemistry-free Bemis–Murcko scaffold groups: core ring systems and their connecting linkers after terminal side chains are removed (Bemis and Murcko, 1996). A seeded search evaluates 500,000 same-scaffold proposals and greedily retains disjoint pools that admit a depth-four target-blind policy. This produces 213 pools containing 3,408 distinct source records. Expanding each of the 16 candidates as the hidden target yields 3,408 target episodes. These episodes share 213 underlying decision problems and should not be treated as 3,408 independent

candidate matrices.

As a separate integrity check, every pool has 16 distinct complete answer profiles over the 53 predicates; no pair of candidates is indistinguishable even before the four-query certificate is considered.

For stratified sampling, we compute a **target-aware witness**: the shortest legal-predicate sequence that isolates a specified target when the sequence may be chosen using that target's complete answer profile. A depth-2 witness, for example, means that two predicates suffice once the target is known offline. It is a diagnostic certificate, never shown to the player, and not a target-blind policy. The release contains 2,883 depth-2, 473 depth-3, and 52 depth-4 witnesses.

4 Evaluation

We use a fixed 100-episode witness-balanced stress test with 34 depth-2, 33 depth-3, and 33 depth-4 witnesses. It covers 76 distinct pools. This deliberate balancing overrepresents the rare depth-4 stratum and is not an estimate of release-wide accuracy. All conditions replay the same list of pool, target, and anonymous-ID assignments.

The primary LLM interface is **predicate-visible**. Candidate cards show exact numeric descriptors and frozen structural-predicate Booleans, so every legal outcome is recoverable without interpreting SMILES. A secondary **hybrid** interface shows the descriptors and canonical SMILES but withholds the structural Booleans. The earlier *raw* and *explained* runs use this hybrid interface; they are not prompt variants of the primary run.

Exact search and greedy information gain receive the Boolean table directly; predicate-visible GPT receives the same facts in candidate-card form. Neither representation reveals the target. All policies must choose questions adaptively.

We evaluate GPT-5.4-mini-2026-03-17 with reasoning disabled. Supporting hybrid runs use Qwen3.5-9B and an FP8 Llama 3.3 70B Instruct checkpoint with thinking disabled and temperature 1.0. Exact prompts, model identifiers, revisions, sampling settings, and response limits are recorded in the manifests. Because serving and decoding differ, cross-model values are descriptive rather than a ranking.

The evaluation harness replaces an invalid or repeated question with the first unused legal predicate in a fixed order, and a guess outside the surviving set with the first surviving candidate. An episode is **fallback-affected** if either substitution occurs. In protocol-compliant results, every fallback-affected episode is counted as a failure; when fallback rates differ, mean question counts use only fallback-free episodes.

5 Results

5.1 Predicate-visible planning

Table 1 gives the main comparison. Exact search solves the full slice. Greedy information gain misses seven targets but still isolates 93/100, showing that the instances are usually easy for a simple table-aware policy. GPT-5.4-mini isolates 9/100 despite receiving the same predicate information. It guesses 37 targets correctly, but 28 of those guesses leave multiple candidates viable.

On the three matched GPT hybrid runs, strict scores are 16, 10, and 9; under protocol-compliant scoring they become 15, 10, and 9 (34/300). Correct-guess scores are 49, 44, and 37. This spread cautions against treating 16/100 as a stable estimate. The predicate-visible score of 9/100 shows that the planning failure remains after structural inference from SMILES is removed, but one run cannot establish an interface effect. The Llama and Qwen hybrid rows are descriptive comparisons; Qwen's fallback rate further limits interpretation.

5.2 What goes wrong

On the 79 episodes that are fallback-free in every model run, we compute the expected information gain of the selected question and the best available question over the current survivor set. We also track realized cumulative information, $G_t = \log_2 16 - \log_2 |S_t|$, where S_t is the survivor set after turn t; identification requires four bits. This common subset keeps the instances identical across curves.

Figure 1a shows the rate at which the policies acquire those four bits. Exact search gains one bit per question and greedy reaches 3.92 bits after four. The four model/interface curves reach only 2.22–2.42 bits. Their mean selected-question gains are 0.64–0.71 bits, while the best available question in the same states remains near 0.99 bits (Figure 1b). On the full 100-game predicate-visible run, GPT chooses a maximum-information question on 68 of 392 turns (17.3%) and a question that does not split the current survivors on 62 turns (15.8%).

If a policy is allowed to ask every legal predicate, the endpoint is uninformative: certification guarantees that it eventually reaches four bits.

Policy	Interface	Runs	Strict ID	Mean q.	Fallback-affected
Exact target-blind search	Frozen predicate table	1	100/100	4.00	0
Greedy information gain	Frozen predicate table	1	93/100	4.00	0
GPT-5.4-mini	Predicate-visible	1	9/100	3.92	0
GPT-5.4-mini	Hybrid	3	34/300	3.95	6/300
Llama 3.3 70B Instruct	Hybrid	1	13/100	3.95	0
Qwen3.5-9B	Hybrid	1	7/100	3.98	18/100

Table 1: Four-question performance on the witness-balanced 100-episode slice.

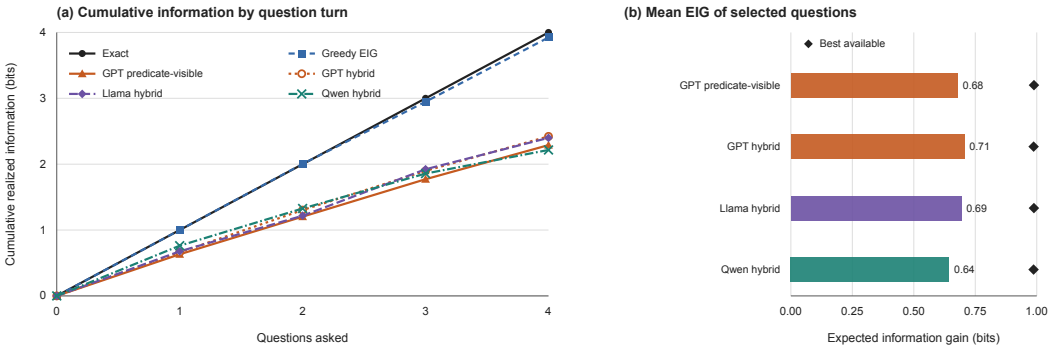

Figure 1: Information acquisition on the common 79-episode fallback-free subset: (a) mean realized cumulative information; (b) mean expected information gain of selected questions. Diamonds show the best legal question in the same state.

The relevant signal is therefore the slope of the information-acquisition curve, not success under a near-exhaustive question allowance. The language model often begins with an imbalanced split or later asks a redundant predicate, leaving too much uncertainty for the remaining budget.

5.3 Near-exhaustive diagnostics

Twenty-five is a turn cap, not a universal inventory size. The 100 games have a median of 21 legal predicates (range 7–36), and 79 have at most 25. GPT, Llama, and Qwen each reach a harness-recorded 100/100 with this hybrid setup. Under protocol-compliant scoring, however, they isolate 96, 97, and 40 targets; mean questions on fallback-free episodes are 9.21, 8.45, and 8.65. Long transcripts can exhaust uncertainty, but these runs are not like-for-like estimates of a four-question planning gap.

For Qwen, 60 episodes require fallback: 97 proposals are invalid (65 repeats and 32 other unavailable questions), and seven final responses name an eliminated candidate after only one survivor remains. Its nominal 100/100 and 9.89-question mean therefore mix model behavior with harness-assisted convergence and expose validity and state-tracking errors.

5.4 Transcript coaching pilot

In a matched 30-game hybrid pilot, GPT-5.1 read completed transcripts and gave GPT-5.4-mini a non-answer hint after games 3, 6, and 9 of each ten-game sequence. Across the 21 post-hint games, coached and control results were identical for correct guesses (9/21 each) and nearly identical for strict identification (0/21 vs. 1/21) and mean selected information gain (0.662 vs. 0.665 bits). This underpowered pilot, whose coach lacked candidate tables, found no detectable effect from hints about balanced splits and redundant questions.

6 Conclusion

All 213 MolQuest problems are optimally solvable in four questions. On the same 100 episodes, exact search isolates 100 targets, greedy information gain 93, and predicate-visible GPT 9. Neither indistinguishable candidates nor structural inference from SMILES explains the gap; the chosen questions leave too much uncertainty.

7 Data and Reproducibility

Upon acceptance, we will archive the benchmark data, policies, prompts, logs, and code under an explicit license. ChEBI identifiers and attribution remain under CC BY 4.0; RDKit is BSD-licensed (EMBL-EBI, 2026; RDKit Contributors, 2026).

Limitations

MolQuest certifies solvability only relative to its frozen predicates and RDKit pipeline. The witness-balanced 100-episode slice is small, clustered by pool, and intentionally unlike the release distribution. The predicate-visible condition currently has one GPT run; repeated runs and predicate-visible evaluations of other models are needed before making comparative claims. Hybrid cross-model runs also differ in serving and decoding. Finally, the benchmark tests text-mediated action selection, not a tool-using agent that can search the table programmatically.

References

Guy W. Bemis and Mark A. Murcko. 1996. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

EMBL-EBI. 2026. Chemical entities of biological interest: About ChEBI. Accessed 18 July 2026.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1):D1214–D1219.

Laurent Hyafil and Ronald L. Rivest. 1976. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems 37*.

Robert D. Nowak. 2011. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906.

RDKit Contributors. 2026. RDKit 2026.03.4.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. 2024. Probing the multi-turn planning capabilities of LLMs via 20 question games. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1495–1516, Bangkok, Thailand. Association for Computational Linguistics.