SOLVE: ADAPTIVE EVIDENCE ACQUISITION FOR VERIFIABLE MULTI-AGENT DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents increasingly integrate tools and multi-step reasoning, but consequential decisions may often require selective evidence acquisition, non-compensatory constraints, and auditable termination. We formulate this setting as *evidence-constrained decision making* and introduce **MarsSite-Bench**, a 600-task benchmark spanning six task types with explicit evidence budgets, structured outputs, and method-independent constraint-gated evaluation. We further propose ***S**equential **O**rchestration with **L**ogical **V**erification and **E**vidence Acquisition* (**SOLVE**), which coordinates adaptive evidence acquisition and role-specialized reasoning through a shared decision state. On 120 hidden-test tasks, SOLVE achieves a Primary CGDQ of **0.6488**, compared to **0.1795** for ReAct. Matching ReAct's evidence access raises its score to **0.3976** but leaves a substantial gap. A real-Mars case study further demonstrates regional ranking followed by deterministic footprint-level verification and abstention under heterogeneous geospatial evidence. These results support adaptive acquisition and constraint-aware orchestration for verifiable LLM-agent decisions.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as agents that interleave reasoning with actions and external tools, enabling information retrieval, environment interaction, and feedback-based revision (Yao et al., 2023; Schick et al., 2023; Gou et al., 2024). However, consequential decisions often require more than plausible answers: an agent must decide what evidence to acquire, distinguish between observed and unavailable information, satisfy non-compensatory constraints, and leave an auditable basis for its decision. We study this setting as *evidence-constrained decision making*, using Mars landing-site selection as a concrete scientific domain.

Unlike ordinary tool use, evidence acquisition changes the decision state itself: each request consumes a budget, may resolve or leave constraints unresolved, and determines whether the system should continue, select, or abstain. The challenge, therefore, is not only to reason over retrieved information but also to control which information enters the decision process and when acquisition should stop. Consequently, evidence selection, stopping behavior, and formal validity become coupled components of the decision policy rather than independent post-processing steps.

Existing frameworks address important aspects of this problem. ReAct and Toolformer support reasoning through external actions and tools; Self-Refine, Reflexion, and CRITIC incorporate iterative or externally grounded correction; and multi-agent methods coordinate complementary reasoning processes (Yao et al., 2023; Schick et al., 2023; Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2024; Du et al., 2024). However, these capabilities do not inherently provide a unified framework in which evidence acquisition is budgeted, decision-supporting claims are linked to acquired evidence, and hard constraints are verified independently of the language model. This motivates a central question: *how can an LLM-based system acquire evidence and coordinate reasoning while ensuring that its final decision remains formally valid?*

To investigate this question independently of any particular architecture, we introduce MarsSite-Bench, a formal benchmark containing 600 evidence-constrained tasks across six task types, including 120 held-out hidden-test tasks. The tasks present a common evidence interface and a structured decision contract, while evaluation remains independent of internal agent roles. Within this setting, we develop ***S**equential **O**rchestration with **L**ogical **V**erification and **E**vidence Acquisition* (**SOLVE**),

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

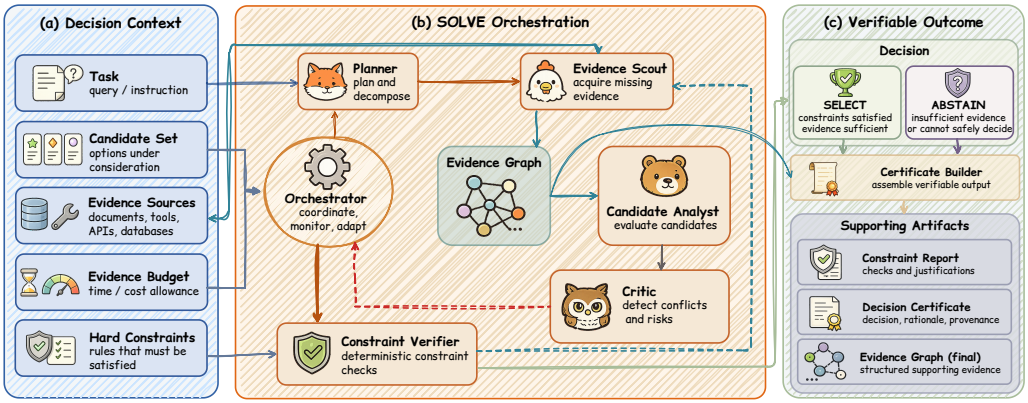

Figure 1: Overview of SOLVE. Specialized agents coordinate evidence acquisition and candidate assessment under explicit budgets, while deterministic verification checks protocol and hard constraints before returning an evidence-linked decision.

a system that coordinates planning, evidence acquisition, candidate analysis, critique, and deterministic verification within an evolving decision state. Figure 1 summarizes this framework.

Our contributions are threefold. First, we formulate the problem of evidence-constrained decision-making and introduce MarsSite-Bench, a method-independent testbed. Then, we propose SOLVE, a framework for adaptive evidence acquisition and constraint-aware orchestration. Third, we evaluate the proposed approach through a 12-method hidden-test comparison, matched-access controls, and a real-Mars case study. SOLVE achieves a Primary CGDQ of 0.6488, significantly outperforming ReAct (0.1795), while matched-access Tool-ReAct reaches 0.3976. The real-Mars case study further demonstrates how deterministic footprint-level verification can trigger abstention when regional admissibility does not guarantee local feasibility.

2 Related Work

Tool use, refinement, and multi-agent reasoning. LLM agents increasingly combine reasoning with external actions and feedback. For instance, ReAct interleaves reasoning and actions, while Toolformer, ToolLLM, and Gorilla investigate tool or application programming interface (API) usage at different scales (Yao et al., 2023; Schick et al., 2023; Qin et al., 2024; Patil et al., 2024). In contrast, approaches such as Self-Refine, Reflexion, CRITIC, and LATS focus on iterative correction, external feedback, or search over reasoning trajectories (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2024; Zhou et al., 2024). Multiagent Debate, AutoGen, and MetaGPT further explore reasoning through interaction or role-specialized agents (Du et al., 2024; Wu et al., 2024; Hong et al., 2024). SOLVE lies at the intersection of these directions: evidence acquisition is a budgeted decision, specialized roles share an evolving evidence state, and formal checks remain outside open-ended LLM generation.

Agent benchmarks and scientific decision making. AgentBench, GAIA, τ-bench, AppWorld, and AgentDojo extend agent evaluation to encompass interactive environments, tool use, policy following, state-based evaluation, and reliable external interaction (Liu et al., 2024; Mialon et al., 2024; Yao et al., 2024; Trivedi et al., 2024; Debenedetti et al., 2024). MarsSite-Bench complements these benchmarks by isolating selective evidence acquisition under an explicit information budget, evidence-linked decisions, and non-compensatory formal constraints. Mars landing-site selection provides a natural scientific setting as actual mission studies combine heterogeneous orbital observations with scientific and engineering criteria during progressive candidate down-selection (Golombek et al., 2012; Grant et al., 2018). MARS-MAS further formulates Martian landing-site selection as a multi-agent decision process involving distinct decision roles (Wang et al., 2025). We use Mars as a controlled scientific domain to study reproducible evidence-constrained decisions, rather than as a proxy for operational mission certification.

3 PROBLEM FORMULATION AND MARSSITE-BENCH

3.1 EVIDENCE-CONSTRAINED DECISION MAKING

We study decision problems in which only partial information is initially available, and supplementary evidence can be obtained prior to reaching a final decision. A task is represented as

$$\mathcal{T} = (x, \mathcal{C}, \mathcal{E}, \mathcal{H}, B, \tau), \qquad s_t = (x, \mathcal{C}, \mathcal{A}_t, \mathcal{H}, b_t), \tag{1}$$

where x denotes the task instruction, $\mathcal{C}$ the candidate set, $\mathcal{E}$ the registered evidence space, $\mathcal{H}$ the formal constraints, B the evidence-acquisition budget, and τ the task type. The state s_t contains the evidence acquired by step t, $\mathcal{A}_t \subseteq \mathcal{E}$, together with the remaining budget b_t.

When evidence acquisition is permitted, an item e with cost $\kappa(e)$ may be acquired only while

$$\sum_{e \in \mathcal{A}_t} \kappa(e) \leq B, \qquad b_t = B - \sum_{e \in \mathcal{A}_t} \kappa(e). \tag{2}$$

A decision system must therefore decide whether to allocate its finite information budget to acquire additional evidence or to terminate the process based on the currently available record. At any admissible state, the system may instead choose to terminate and output the structured decision required by the task.

Hard constraints are treated as non-compensatory rather than as weighted score components. Let $z_j(c \mid \mathcal{A}_t)$ denote the status of constraint j for candidate c, taking values in $\{\text{SATISFIED}, \text{VIOLATED}, \text{UNRESOLVED}\}$. The corresponding feasibility gate is

$$G(c \mid \mathcal{A}_t) = \prod_{j=1}^{|\mathcal{H}|} \mathbb{I}\left[z_j(c \mid \mathcal{A}_t) = \text{SATISFIED}\right]. \tag{3}$$

Thus, a violated or unresolved hard constraint cannot be offset by strong performance on other criteria. Depending on τ, a task may require evidence acquisition or termination, feasibility assessment, candidate selection or ranking, evidence-linked claims, abstention, or revision under an explicitly specified counterfactual condition.

3.2 MARSSITE-BENCH

MarsSite-Bench implements this framework by employing Mars landing-site selection as a controlled domain for scientific reasoning. Mars landing-site selection combines heterogeneous observations, scientific objectives, and engineering constraints, making it suitable for evidence-constrained evaluation (Golombek et al., 2012; Grant et al., 2018). The benchmark abstracts these characteristics into formal decision tasks; it is not intended to replicate operational mission certification or to encode a specific agent architecture.

MarsSite-Bench contains **600 balanced tasks** across six task types: T1 (acquisition/termination), T2 (feasibility), T3 (preference selection), T4 (constrained ranking), T5 (evidence-linked claims), and T6 (counterfactual revision). Each type contributes 40/20/20/20 instances to the Train, Development, Public Test, and Hidden Test, respectively, yielding 480 public input-only tasks and 120 held-out tasks. Public instances provide only task-visible information and task-specific output contracts, while hidden evaluation references remain withheld (Figure 2).

3.3 METHOD-INDEPENDENT EVALUATION

All systems receive the same task-visible state and output schemas. Raw execution validity is recorded before normalization by checking whether the finalized output satisfies the public minimal-output contract; this diagnostic is not used as a scoring gate. A deterministic normalizer may then apply only representation-equivalent transformations such as documented aliases or unambiguous namespace mappings, but cannot alter decision semantics, invent evidence, or access hidden evaluator state.

Let $\mathcal{N}$ denote this interface normalization, $\hat{y}_i$ the model output for task i, and $\mathcal{R}_i$ its held-out evaluation reference, and N the number of evaluated tasks. For task type τ_i, the normalized task score

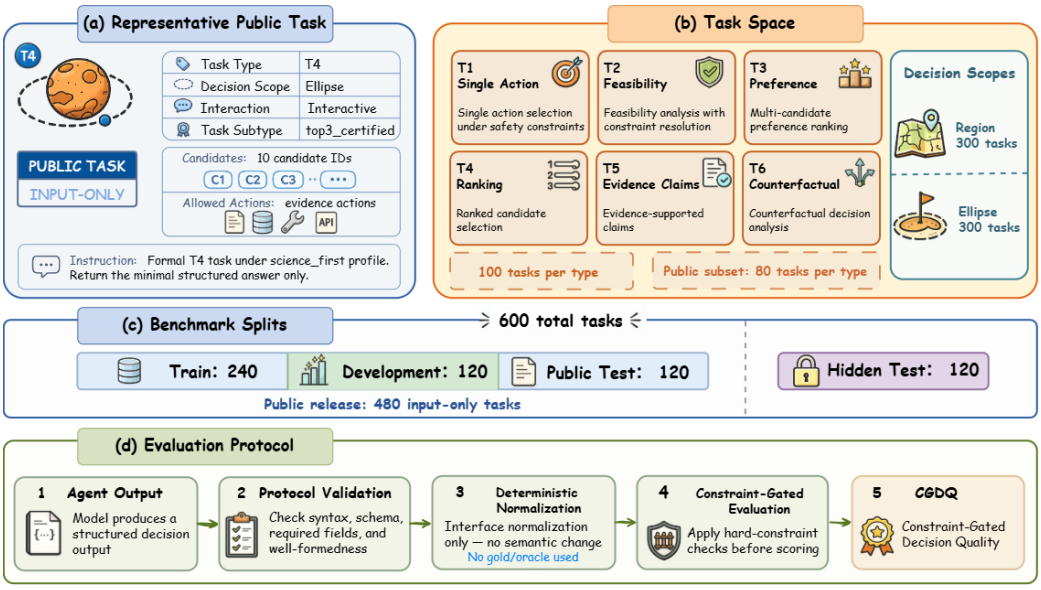

Figure 2: Overview of MarsSite-Bench. Six complementary task types instantiate evidence-constrained decision making across balanced Train, Development, Public Test, and Hidden Test splits, with all methods evaluated through task-specific structured output contracts.

is

$$q_i = \mathcal{E}_{\tau_i}\left(\mathcal{N}(\hat{y}_i); \mathcal{R}_i\right) \in [0,1], \qquad \text{Primary CGDQ} = \frac{1}{N}\sum_{i=1}^{N} q_i, \tag{4}$$

where CGDQ denotes *Constraint-Gated Decision Quality*. The evaluator $\mathcal{E}_\tau$ is task-type specific: Primary CGDQ is therefore not defined by a single additive formula shared across T1–T6. Instead, each task type activates the decision, evidence, and constraint checks appropriate to its output contract.

Formal protocol validity and constraint satisfaction remain distinct diagnostics. Consequently, a syntactically valid answer may still fail a hard constraint, whereas an unresolved constraint does not automatically satisfy the feasibility requirement. Crucially, neither the task interface nor the evaluator contains SOLVE-specific roles, communication patterns, or orchestration logic. Single-agent, self-refining, and multi-agent systems are consequently evaluated under the same observable state and external decision contract; normalization and protocol details are provided in Appendix A.2.

4 SOLVE

4.1 State-Aware Multi-Agent Orchestration

SOLVE uses role-specialized reasoning over a shared decision state rather than aggregating answers from independent agents. An orchestrator coordinates a Planner, Evidence Scout, Candidate Analyst, Critic, and Constraint Verifier over the same evolving evidence context. The working state of SOLVE is represented as

$$\mathcal{M}_t = (\mathcal{A}_t, \Gamma_t, \mathcal{R}_t, \mathcal{Z}_t, b_t)\,, \tag{5}$$

where $\mathcal{A}_t$ is the acquired evidence, Γ_t denotes the structured evidence-link state connecting observations to candidate-level claims, $\mathcal{R}_t$ stores current candidate assessments, $\mathcal{Z}_t$ records formal constraint states, and b_t is the remaining acquisition budget. The Planner identifies unresolved decision requirements from $\mathcal{M}_t$, while the orchestrator dispatches the next role and writes its structured result back to the shared state. This design makes the intermediate reasoning stateful: subsequent roles operate on existing evidence and assessments rather than restarting from the original prompt or relying on unconstrained conversation history.

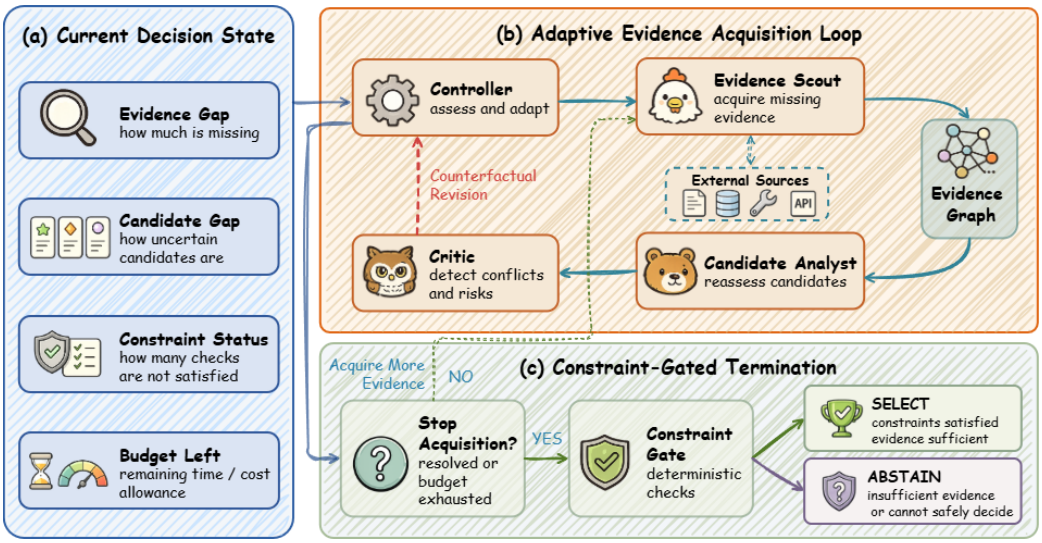

Figure 3: Adaptive evidence acquisition and constraint-gated termination in SOLVE. The controller uses the states of unresolved evidence, candidates, and constraint states to coordinate subsequent acquisition and analysis, and terminates only after the resulting decision is passed through the deterministic constraint gate.

4.2 Adaptive Evidence Acquisition

The central control problem in SOLVE (Figure 3) involves determining whether the available evidence is sufficient for a reliable decision or if an additional observation is required. The Planner initially identifies gaps that still affect candidate discrimination or constraint resolution. Subsequently, the Evidence Scout translates a selected gap into an admissible request from the registered evidence space. For an acquisition action $a_t = \mathrm{ACQUIRE}(e_t)$, the state transition is

$$\mathcal{A}_{t+1} = \mathcal{A}_t \cup \{e_t\}, \qquad b_{t+1} = b_t - \kappa(e_t), \qquad \kappa(e_t) \leq b_t. \tag{6}$$

Importantly, SOLVE does not follow a fixed evidence schedule. The next request is conditioned on the current state of candidates, evidence, and constraints. The controller may terminate acquisition when the evidence is judged sufficient to fulfill the task contract or when the remaining budget prevents further admissible requests. Consequently, evidence acquisition is integrated into the reasoning policy itself, rather than being treated as a pre-processing stage.

4.3 Evidence-Grounded Analysis and Revision

Once incorporating new evidence, the Candidate Analyst reassesses the affected candidates based on the current evidence state and must associate supporting evidence references with candidate-level claims. These candidate-level conclusions, together with their supporting evidence links in Γ_t, provide a provenance path connecting the final decision to the observations that motivated it. Then the Critic evaluates the current assessment for unsupported claims, conflicting evidence, unresolved risks, or inconsistencies between candidate conclusions and the accumulated record. If such issues are identified, the Critic provides structured feedback to the orchestrator, which may prompt reassessment or, if the remaining budget permits, an additional evidence request.

Roles are not interchangeable votes: while the Critic can identify unsupported claims, conflicts, or unresolved risks, it cannot overwrite measurements, generate missing evidence, or determine hard-constraint satisfaction. This approach preserves flexible semantic revision while ensuring the evidence interface remains auditable.

4.4 Constraint-Gated Verification and Termination

The final stage separates semantic assessment from formal validity. The Constraint Verifier deterministically evaluates the structured output, candidate references, and task-specific hard constraints

against the observable state defined in Section 3.1. Let $\tilde{y} = \mathrm{SELECT}(c^\star)$ denote the proposed selection output. For selection-style tasks, the terminal decision can be summarized as

$$y = \begin{cases} \tilde{y}, & \text{if } P(\tilde{y}) = 1 \ \wedge \ G(c^\star \mid \mathcal{A}_t) = 1, \\ \mathrm{ABSTAIN}, & \text{if no candidate can be supported as admissible.} \end{cases} \tag{7}$$

where $P(\tilde{y})$ denotes protocol validity and $G(\cdot)$ is the non-compensatory gate in Eq. 3. Task types with ranking, evidence reporting, or counterfactual outputs utilize the corresponding structured termination contracts, but the same principle applies: semantic confidence cannot override formal invalidity.

The verifier is not a learned scoring agent; it establishes an external decision boundary for protocol and constraint violations. SOLVE therefore employs LLM reasoning where interpretation is required and deterministic checks where validity can be formalized.

5 Experiments

5.1 Experimental Setup

We evaluate SOLVE on the 120-task hidden split of MarsSite-Bench after finalizing method configurations using the Development split. No hidden-test information is used for method selection. All methods use GPT-5.4-mini (OpenAI, 2026) and are evaluated with identical task contracts and the Primary CGDQ metric. We compare SOLVE with eleven standard alternatives, including direct prompting and chain-of-thought (CoT) reasoning (Brown et al., 2020; Wei et al., 2022), tool-augmented and self-refining agents (Yao et al., 2023; Gou et al., 2024; Madaan et al., 2023), planning and voting strategies (Wang et al., 2023a;b), and multi-agent reasoning through debate, role specialization, and shared context (Du et al., 2024; Hong et al., 2024; Wu et al., 2024).

The standard comparison preserves each method's native reasoning and action structure, which results in imperfectly matched evidence access; we control this confound in Section 5.4. Since cross-method Evidence Grounding diagnostics are not operationally identical, we report the Primary CGDQ, execution validity, nonzero-score rate, and mean LLM calls. Here, the nonzero-score rate is the fraction of normalized evaluations for which the primary score is strictly greater than zero. Confidence intervals (CIs) are calculated using 10,000 bootstrap resamples; paired tests employ 10,000 sign-flip permutations with Holm correction across the eleven comparisons.

5.2 Main Hidden-Test Results

Table 1 shows a large separation between SOLVE and the standard baselines. SOLVE achieves a Primary CGDQ of **0.6488** (95% CI: [0.5722, 0.7205]), versus **0.1795** ([0.1245, 0.2376]) for the strongest standard baseline, ReAct, an absolute margin of **0.4693**. All other standard baselines remain below 0.18.

Figure 4 visualizes both the aggregate performance gap and its task-level consistency. Compared to ReAct, SOLVE achieves superior performance on 84 of 120 hidden tasks, ties on 31, and underperforms on only 5. This qualitative pattern is consistent across the other ten baselines. All eleven paired comparisons remain significant after Holm correction, with adjusted $p = 1.10\times10^{-3}$, indicating that the results are not attributable to a small subset of unusually favorable tasks. Nevertheless, performance varies by task type. SOLVE approaches ceiling performance on T5 evidence-linked claims (0.9950), while T2 feasibility assessment (0.5250), T4 constrained ranking (0.5642), and T6 counterfactual revision (0.5000) remain substantially more challenging. Thus, the aggregate improvement does not eliminate difficult decision modes. Complete task-wise results are reported in Appendix A.4.

5.3 Validity and Inference Cost

SOLVE achieves 99.2% execution validity and a 78.3% nonzero-score rate, versus 85.8% and 34.2% for ReAct, respectively. We use observed LLM completion calls per task as a descriptive cost proxy; deterministic tool and verifier operations are excluded. SOLVE averages 4.17 calls per task;

Table 1: MarsSite-Bench Hidden-Test results ($N = 120$). Best results are bold; the strongest standard baseline by Primary CGDQ is underlined.

Method	Primary ↑	95% CI	Exec. Valid. (%) ↑	Nonzero Score (%) ↑	Calls / Task ↓
SOLVE	**0.6488**	[0.5722, 0.7205]	**99.2**	**78.3**	4.17
ReAct	<u>0.1795</u>	[0.1245, 0.2376]	85.8	34.2	2.42
Multi-Agent Debate	0.1700	[0.1194, 0.2251]	42.5	32.5	3.16
CRITIC	0.1671	[0.1144, 0.2258]	10.8	31.7	0.55
CoT	0.1446	[0.0944, 0.2006]	84.2	29.2	1.26
Shared-memory	0.1311	[0.0863, 0.1801]	40.8	28.3	3.08
Majority Voting	0.1233	[0.0789, 0.1719]	26.7	28.3	2.02
Planner-only	0.1040	[0.0651, 0.1491]	60.8	25.0	1.93
Role-specialized	0.0945	[0.0584, 0.1347]	5.0	24.2	1.06
Direct Prompting	0.0926	[0.0542, 0.1347]	89.2	20.8	1.35
Self-Refine	0.0894	[0.0565, 0.1269]	13.3	24.2	0.92
Planner-Executor	0.0855	[0.0576, 0.1160]	49.2	26.7	2.13

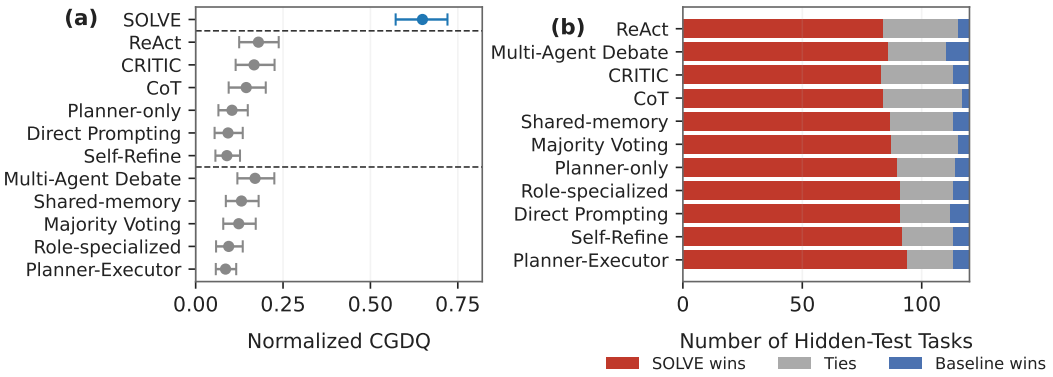

Figure 4: Hidden-test comparison on MarsSite-Bench. (a) Primary CGDQ with 95% bootstrap confidence intervals; dashed separators distinguish SOLVE, single-agent baselines, and multi-agent baselines. (b) Paired task-level win, tie, and loss counts for SOLVE against each standard baseline.

however, higher call counts alone do not explain performance, as Multi-Agent Debate and Shared-memory use comparable calls but achieve much lower CGDQ. Figure 5 summarizes these validity and inference-cost results.

5.4 Matched-Access Controlled Analysis

The controlled variants were fixed after Development-stage configuration freeze and evaluated once on the Hidden Test without Hidden-driven tuning. To isolate the effect of evidence access, *Tool-ReAct* retains ReAct-style reasoning but adopts SOLVE's evidence-access boundary and tool interface. *SOLVE w/o Verification* retains planning, acquisition, and candidate analysis while removing Critic-driven revision and verifier feedback.

Matching evidence access substantially improves ReAct from 0.1795 to **0.3976**. The paired increase is +0.2181 (95% CI: [0.1327, 0.3060], Holm-adjusted $p = 3.00 \times 10^{-4}$). Nevertheless, full SOLVE remains superior by **+0.2512** over Tool-ReAct ([0.1597, 0.3424], $p = 3.00 \times 10^{-4}$). Thus, matched evidence access explains an important part of the baseline gap but does not close it. Because the controlled methods also differ in execution validity, the residual difference should not be attributed to orchestration alone.

Removing verification yields a Primary CGDQ of 0.6577, compared with 0.6488 for full SOLVE. The paired difference is -0.0088 ([-0.0551, 0.0373]) and is not statistically significant (Holm-adjusted $p = 0.719$). Therefore, we do not interpret the verification loop as the primary source of higher Primary CGDQ. Instead, the verification loop is designed to provide explicit constraint

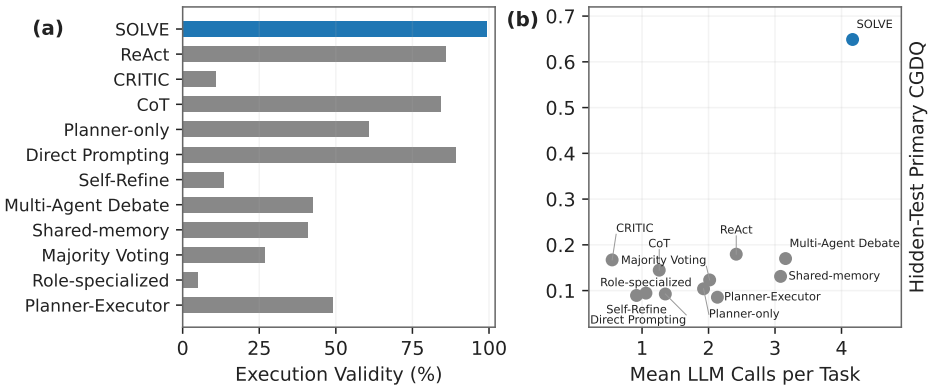

Figure 5: Validity and inference cost on the hidden test. (a) Execution validity. (b) Primary CGDQ versus mean LLM calls per task, with SOLVE highlighted.

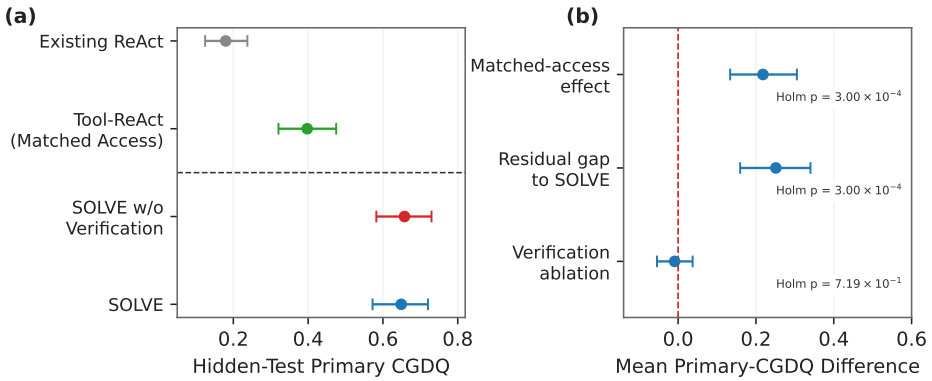

Figure 6: Matched-access controlled analysis. (a) Hidden-test Primary CGDQ with 95% bootstrap confidence intervals for Existing ReAct, matched-access Tool-ReAct, SOLVE without verification, and full SOLVE. (b) Paired effect estimates with 95% confidence intervals; the dashed vertical line denotes zero effect and annotations report Holm-adjusted permutation-test p-values.

checking, auditability, and verifiable termination, properties that are not fully captured by Primary CGDQ alone. Figure 6 summarizes these controlled effects; complete access conditions and paired statistics are detailed in Appendix C.1.

6 Real-Mars Application

We apply SOLVE to ten candidate regions on Mars, utilizing registered topographic, thermophysical, geomorphic, and mineralogical evidence, including products derived from the Mars Orbiter Laser Altimeter (MOLA), the Thermal Emission Imaging System (THEMIS), Compact Reconnaissance Imaging Spectrometer for Mars (CRISM), and Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité (OMEGA) instruments (Smith et al., 2001; Christensen et al., 2004; Murchie et al., 2007; Bibring et al., 2004). Terrain (T), environment (E), and local-surface (L) components are deterministic, whereas Science Utility (S) is assigned under a fixed evidence-grounded LLM rubric. Regional admissibility requires elevation ≤ -2800 m, slope $\leq 9°$, albedo ≤ 0.25, thermal inertia ≥ 100 thermal inertia units (TIU), and dust-cover index ≥ 0.95; unresolved gates do not pass. For admissible candidates,

$$F(c) = \text{round}\left(0.20T + 0.15E + 0.15L + 0.50S,\ 4\right). \tag{8}$$

Evidence acquisition is limited to a maximum of 24 abstract units per candidate over 20 registered evidence types; unavailable registered evidence is not treated as an observed zero. The com-

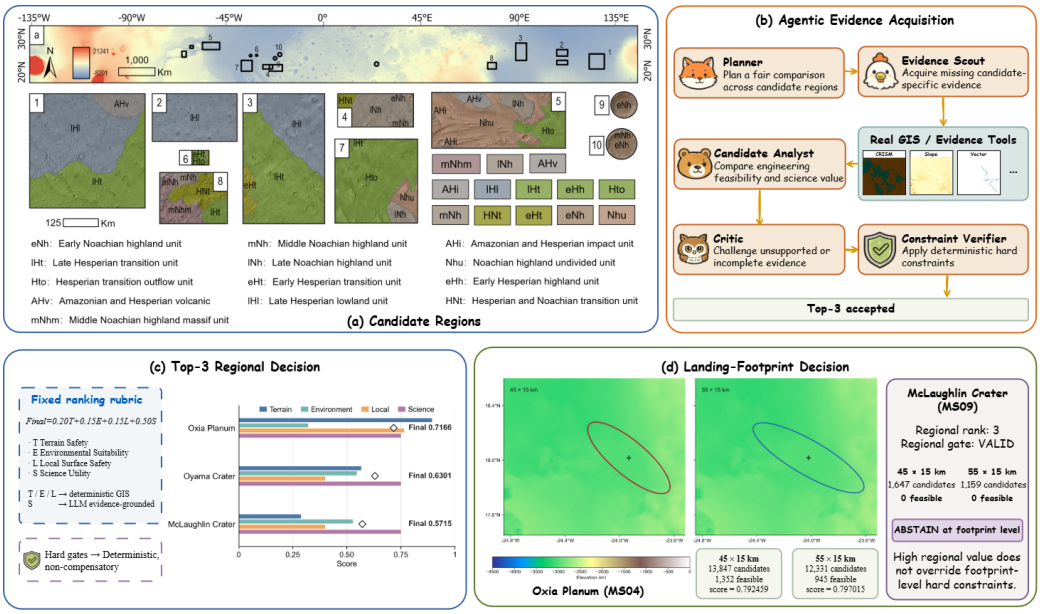

Figure 7: Representative real-Mars outcomes illustrate the distinction between regional admissibility and footprint-level feasibility. Oxia Planum remains selectable after deterministic footprint verification, whereas McLaughlin Crater, while regionally admissible, yields no feasible 45×15 or 55×15 footprint and therefore triggers abstention.

plete evidence universe, abstract cost schedule, and missing-evidence semantics are detailed in Appendix D.3.

Seven of the ten candidate regions pass the complete regional gate. **Oxia Planum (MS04)** ranks first with $F = 0.7166$, followed by **Oyama Crater (MS10)** at 0.6301 and **McLaughlin Crater (MS09)** at 0.5715; Zhurong, Isidis, and Nili are excluded before ranking. Thus, Oxia is the highest-ranked candidate *within the specified ten-region universe under the stated evidence, constraints, budget, and scoring protocol*, rather than a claim of global optimality over Mars.

Regional admissibility does not guarantee a feasible landing footprint. At Oxia, 1,352 of 13,847 evaluated 45×15 footprints and 945 of 12,331 55×15 footprints satisfy all local constraints, enabling selection. McLaughlin provides the complementary case: despite regional validity, none of 1,647 45×15 or 1,159 55×15 footprints are feasible, so the verified outcome is ABSTAIN. Across three repeated runs and evidence budgets of 17.5, 24, and 32.5 abstract units, Oxia remains Top-1 and the Top-3 membership remains $\{\text{MS04}, \text{MS09}, \text{MS10}\}$, although the ordering of MS09 and MS10 can change (Figure 7).

7 Conclusion

We introduce evidence-constrained decision making, MarsSite-Bench, and SOLVE, which integrates adaptive evidence acquisition with constraint-aware multi-agent orchestration. On 120 hidden tasks, SOLVE achieves 0.6488 Primary CGDQ versus 0.1795 for ReAct; matched access raises ReAct to 0.3976 but leaves a substantial gap. The real-Mars study further demonstrates regional ranking followed by deterministic footprint verification, including abstention when no locally feasible footprint exists. However, our evidence is currently limited to MarsSite-Bench, a single LLM backbone, and a fixed ten-region Mars case, leaving the transferability across benchmarks, domains, and model families an open question. Although verification does not significantly improve the Primary CGDQ, it provides explicit constraint checking, auditability, and verifiable termination.

AI USE STATEMENT

Generative AI tools, including general-purpose LLM assistants and LLM-based coding assistants, were used in this work for language editing and software-engineering support. For the manuscript, these tools were used to enhance author-written text for clarity, grammar, and readability. For code, they assisted in organizing and refining implementations, as well as in debugging and code review. When AI-assisted code contributed to method implementation, the authors inspected and tested the resulting code before its use.

Beyond these applications, generative AI tools were not utilized to formulate or refine research hypotheses; design or provide feedback on the research methodology or experiments; develop theoretical or mathematical claims or proofs; or generate or clean datasets. The LLMs that constitute components of the systems evaluated in this paper are part of the reported experimental methodology rather than serving as authoring aids, and their roles are detailed separately in the main text and appendix.

AI-assisted code was verified through deterministic checks, regression tests, and manual review. Benchmark scoring, statistical tests, GIS-derived engineering quantities, and formal constraint checks were generated by frozen programmatic procedures, while LLM outputs used as explicit components of the evaluated systems were managed according to the experimental protocol. The authors reviewed all AI-assisted manuscript edits and code and take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human participants, personal data, or sensitive user information. MarsSite-Bench is a research benchmark for studying formal evidence-constrained decision making, and the real-Mars case study uses planetary geospatial and scientific data rather than human-derived data. SOLVE is intended as a research decision-support and evaluation framework and should not be interpreted as a substitute for operational mission certification, engineering safety review, or expert scientific judgment. To preserve benchmark integrity, held-out evaluation references are not released, while public task definitions, schemas, and protocol descriptions are provided to support transparent research use.

REPRODUCIBILITY STATEMENT

The formal task definition, benchmark evaluation protocol, SOLVE mechanisms, and experimental setup are detailed in Sections 3–5. Appendices A–E provide evaluator and normalization details, task-wise and paired results, controlled-study statistics, real-Mars engineering equations and evidence semantics, robustness and footprint analyses, algorithmic pseudo-code, and runtime and configuration details. Anonymous supplementary material includes the SOLVE implementation, benchmark schemas, protocol-validation components, and sanitized examples for inspecting the public interface. The accompanying MarsSite-Bench release contains 480 public input-only tasks. The 120-task Hidden Test and its private evaluation references are withheld to maintain the integrity of the held-out evaluation; this release boundary is explicitly documented to enable reproduction of the public benchmark interface and deterministic evaluation components without exposing the held-out evaluation set.

REFERENCES

J-P Bibring, Alain Soufflot, Michel Berthé, Yves Langevin, Brigitte Gondet, Pierre Drossart, Marc Bouyé, Michel Combes, Pascal Puget, Alain Sémery, et al. OMEGA: Observatoire pour la minéralogie, l'eau, les glaces et l'activité. *In: Mars Express: the scientific payload. Ed. by Andrew Wilson, scientific coordination: Agustin Chicarro. ESA SP-1240, Noordwijk, Netherlands: ESA Publications Division, ISBN 92-9092-556-6, 2004, p. 37-49*, 1240:37–49, 2004. URL https://ui.adsabs.harvard.edu/scan/manifest/2004ESASP1240...37B.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel

Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Philip R Christensen, Bruce M Jakosky, Hugh H Kieffer, Michael C Malin, Harry Y McSween Jr, Kenneth Nealson, Greg L Mehall, Steven H Silverman, Steven Ferry, Michael Caplinger, et al. The thermal emission imaging system (THEMIS) for the Mars 2001 odyssey mission. *Space Science Reviews*, 110(1):85–130, 2004. URL https://doi.org/10.1023/B:SPAC.0000021008.16305.94.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82895–82920. Curran Associates, Inc., 2024. doi: 10.52202/079017-2636. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/97091a5177d8dc64b1da8bf3e1f6fb54-Paper-Datasets_and_Benchmarks_Track.pdf.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/du24e.html.

M Golombek, John Grant, Devin Kipp, A Vasavada, Rhona Kirk, R Fergason, P Bellutta, Fred Calef, K Larsen, Y Katayama, et al. Selection of the Mars Science Laboratory landing site. *Space Science Reviews*, 170(1):641–737, 2012. URL https://doi.org/10.1007/s11214-012-9916-y.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf.

John A Grant, Matthew P Golombek, Sharon A Wilson, Kenneth A Farley, Ken H Williford, and Al Chen. The science process for selecting the landing site for the 2020 Mars rover. *Planetary and Space Science*, 164:106–126, 2018. URL https://doi.org/10.1016/j.pss.2018.07.001.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6507b115562bb0a305f1958ccc87355a-Paper-Conference.pdf.

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, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e9df36b21ff4ee211a8b71ee8b7e9f57-Paper-Conference.pdf.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine:

Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25ae35b5b1738d80f1f03a8713e405ec-Paper-Conference.pdf.

Scott Murchie, Raymond Arvidson, Peter Bedini, Kirk Beisser, J-P Bibring, Janice Bishop, J Boldt, P Cavender, T Choo, RT Clancy, et al. Compact reconnaissance imaging spectrometer for Mars (CRISM) on Mars reconnaissance orbiter (MRO). *Journal of Geophysical Research: Planets*, 112(E5), 2007. URL https://doi.org/10.1029/2006JE002682.

OpenAI. Introducing GPT-5.4 mini and nano, March 2026. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126544–126565. Curran Associates, Inc., 2024. doi: 10.52202/079017-4020. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4c61f578ff07830f5c37378dd3ecb0d-Paper-Conference.pdf.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/28e50ee5b72e90b50e7196fde8ea260e-Paper-Conference.pdf.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551. Curran Associates, Inc., 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

David E Smith, Maria T Zuber, Herbert V Frey, James B Garvin, James W Head, Duane O Muhleman, Gordon H Pettengill, Roger J Phillips, Sean C Solomon, H Jay Zwally, et al. Mars Orbiter Laser Altimeter: Experiment summary after the first year of global mapping of Mars. *Journal of Geophysical Research: Planets*, 106(E10):23689–23722, 2001. URL https://doi.org/10.1029/2000JE001364.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 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)*, pp. 16022–16076, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2634, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.147. URL https://aclanthology.org/2023.acl-long.147/.

Rong Wang, Yongjiu Feng, Panli Tang, Pengshuo Li, Yiyan Dong, Yusheng Xu, Chao Wang, Sicong Liu, Yanmin Jin, Shijie Liu, et al. A multi-agent decision-making model for martian landing site selection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. URL https://doi.org/10.1109/JSTARS.2025.3589288.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62138–62160. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhou24r.html.

A MarsSite-Bench Details

A.1 Task Types and Benchmark Construction

MarsSite-Bench contains 600 balanced tasks across six formal evidence-constrained decision types. Each type contributes 100 instances, split as 40/20/20/20 across the Train, Development, Public Test, and Hidden Test. The public release contains the first three splits as 480 input-only tasks; Hidden-Test tasks and all private evaluation references are withheld.

Benchmark construction is deterministic rather than LLM-generated. A frozen blueprint and deterministic seeds define task slots, candidate truths, evidence states, output contracts, and internal evaluation references. T1–T5 are constructed using a formal task generator and canonicalized into the released benchmark format. T6 employs a separate clean-causal pathway: a base state and an outcome-independent counterfactual specification are first constructed, the counterfactual transformation is applied deterministically, and the base and counterfactual decisions are subsequently derived from the corresponding states. Benchmark generation and formal evaluation do not require an LLM or external API calls.

The benchmark candidates are formal decision objects, not 600 independent real-GIS landing sites. Region-level entities are selected from the registered MS01–MS10 entity pool, while formal ellipse candidates are task-level objects generated from deterministic pools, slots, and stable identifiers. The separate real-Mars application in Appendix D utilizes actual registered geospatial evidence and should not be conflated with these formal benchmark objects.

The six task types assess complementary decision capabilities. T1 implements task-subtype-specific ACTION/STOP/ABSTAIN decisions; for an ACTION instance, the target is selected from the allowed action options, whereas STOP and ABSTAIN are specified by the task subtype rather than determined through a runtime value-of-information search. T2 evaluates feasibility classification, T3 selects a preferred candidate, T4 performs an ordered ranking, T5 generates evidence-linked support and limitation claims, and T6 assesses consistency under an explicit counterfactual modification. Across task types, the benchmark varies candidate sets, observable evidence states, action catalogues, budgets, formal constraints, and counterfactual conditions while retaining a common structured interface.

For T1–T5, the formal generator uses a mission-profile utility

$$U_p(c) = w_s \,\mathrm{science}(c) - w_r \,\mathrm{risk}(c) - w_u \,\mathrm{uncertainty}(c), \tag{A1}$$

with profile weights (w_s, w_r, w_u) equal to $(10, 7, 5)$ for `balanced`, $(8, 11, 5)$ for `safety_first`, $(12, 6, 4)$ for `science_first`, and $(9, 7, 9)$ for `uncertainty_averse`. For the T1 ACTION subtype, an action tied to candidate c receives

$$V(a) = \mathrm{round}(U_p(c)/20, 3)\,, \tag{A2}$$

and the Gold ACTION is the highest-valued allowed action. This formal benchmark utility is distinct from the real-Mars ranking formula in Eq. A21.

A.2 Evaluation, Normalization, and Constraint Gating

MarsSite-Bench distinguishes between interface validity and semantic decision quality. Each raw model output is first validated against the task-specific protocol, including required fields, namespaces, and well-formed structured values. A deterministic normalization layer is then applied to repair only representation-equivalent interface differences. Examples include documented field aliases, scalar-to-singleton wrapping where the schema permits a unique interpretation, and unambiguous one-to-one mapping for candidate or action namespace.

Normalization is deliberately restricted to the interface layer. It cannot change the selected candidate or action, rewrite a claim, introduce unobserved evidence, resolve an ambiguous reference, or access hidden evaluation state. Consequently, normalization may make an already interpretable answer protocol-compatible, but it cannot convert a semantically incorrect decision into a correct one.

After protocol handling, the evaluation process employs the task-specific scoring branch described in Section 3. Formal protocol validity and hard-constraint satisfaction remain separate: a well-formed

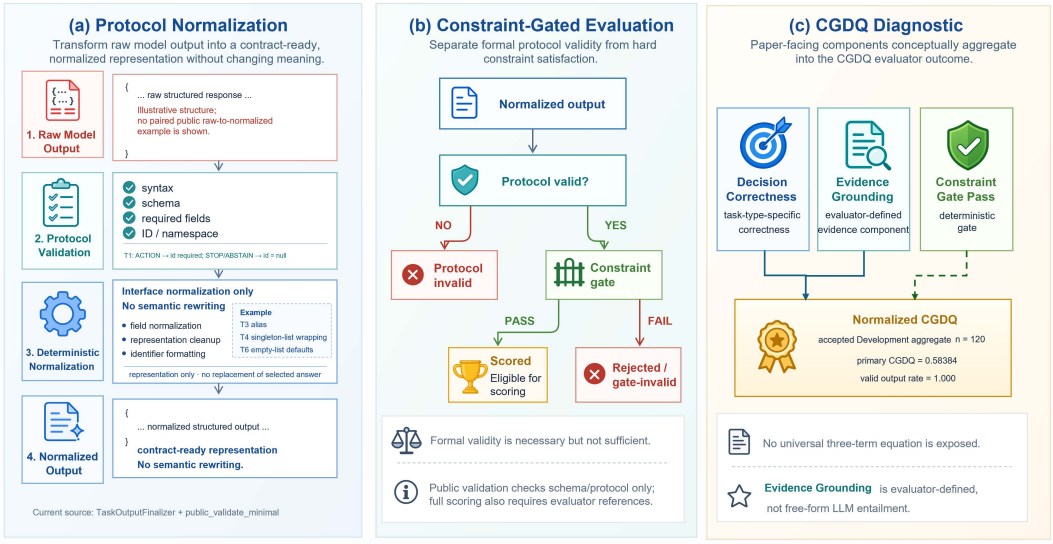

Figure A1: Protocol normalization and constraint-gated evaluation in MarsSite-Bench. Deterministic normalization is restricted to interface-equivalent transformations and does not rewrite decision semantics. Formal protocol validity and hard-constraint satisfaction are evaluated separately, while CGDQ uses task-specific evaluation branches rather than a universal additive formula. Protocol validity is reported as a separate diagnostic and does not itself zero the task-specific primary score.

output may still violate a constraint, while an unresolved required constraint does not automatically result in a passing evaluation. Similarly, the Primary CGDQ is derived from task-specific normalized scores, rather than from a single additive decomposition shared across tasks T1–T6. Figure A1 summarizes this distinction.

A.3 Task-Specific Primary CGDQ

The formal evaluator presents the task-specific primary score as Constraint-Gated Decision Quality. It first validates the task-specific schema and evidence references. A core-equivalent output that satisfies its evidence requirements receives a score of one; otherwise, the partial-credit branch is activated based on the task type. The following summary aligns with the implemented evaluator and is provided to clarify that CGDQ is not a universal weighted sum.

For T1, if both the Gold and submitted outputs are ACTION decisions but the action is not exact, the score is the normalized action-value ratio

$$q_{\mathrm{T1}} = \mathrm{clip}\left(\frac{V_{\mathrm{pred}}}{V_{\mathrm{gold}}}, 0, 0.99\right), \tag{A3}$$

where exact output receives 1 and a task-subtype mismatch receives 0. Here the three-argument $\mathrm{clip}(x, l, u)$ denotes clipping to $[l, u]$ and is separate from the one-argument engineering shorthand introduced later in Appendix D.1.

For T2, let c indicate status correctness and let g indicate the presence of a valid acquired-evidence reference. Except for the false-safe branch, the score is

$$q_{\mathrm{T2}} = c\,(0.5 + 0.5g). \tag{A4}$$

A prediction of feasible when the Gold status is infeasible or unresolved is explicitly assigned zero.

For T3, define $b = 1$ for a correct abstention or exact preferred set, $b = 0.5$ for a non-empty set overlap, and $b = 0$ otherwise. With evidence factor $e = 1$ when evidence references are present and $e = 0.5$ otherwise,

$$q_{\mathrm{T3}} = b\,e. \tag{A5}$$

Table A1: Primary CGDQ by MarsSite-Bench Hidden-Test task type. Each task type contains 20 instances. The final column reports the balanced aggregate over all 120 tasks.

Method	T1	T2	T3	T4	T5	T6	Primary
SOLVE	**0.6864**	**0.5250**	**0.6225**	**0.5642**	**0.9950**	**0.5000**	**0.6488**
ReAct	0.7023	0.0000	0.2250	0.0250	0.0000	0.1250	0.1795
Multi-Agent Debate	0.5577	0.0000	0.2125	0.1583	0.0000	0.0917	0.1700
CRITIC	0.6275	0.0000	0.1000	0.1500	0.0000	0.1250	0.1671
CoT	0.6137	0.0000	0.1125	0.0083	0.0000	0.1333	0.1446
Shared-memory	0.3696	0.0000	0.1750	0.1250	0.0000	0.1167	0.1311
Majority Voting	0.4105	0.0000	0.1125	0.0917	0.0000	0.1250	0.1233
Planner-only	0.2864	0.0000	0.1125	0.1000	0.0000	0.1250	0.1040
Role-specialized	0.1921	0.0000	0.1000	0.1500	0.0000	0.1250	0.0945
Direct Prompting	0.2806	0.0000	0.1000	0.0750	0.0000	0.1000	0.0926
Self-Refine	0.1362	0.0000	0.1250	0.1500	0.0000	0.1250	0.0894
Planner-Executor	0.0295	0.0000	0.2500	0.1083	0.0000	0.1250	0.0855

For T4, any predicted identifier outside the task-visible semantic candidate set receives zero. Otherwise, the ranking term is assigned a value of one for an exact match in the Top-k (or a correct abstention). For a non-exact ranking, partial credit measures candidate-set recovery through overlap with the Gold Top-k set, capped at one; it does not separately score pairwise ordering among partially correct candidates. Exact credit still requires an exact Top-k match. This ranking term is then multiplied by the evidence factor e as defined above. It is important to note that the candidate-membership gate in the current evaluator should not be interpreted as the physical GIS hard-gate verification used in the real-Mars application.

For T5, let s and l indicate the presence of support and limitation/uncertainty claims. The partial-credit branch is

$$q_{\mathrm{T5}} = (0.5s + 0.5l)\, e. \tag{A6}$$

The evidence factor checks whether the response cites acquired evidence; it is not an LLM entailment judge for whether a cited item semantically proves a claim.

For T6, let b_{ok}, c_{ok}, and z_{ok} denote exact agreement for the base decision, counterfactual decision, and changed flag. Then

$$q_{\mathrm{T6}} = \frac{b_{\mathrm{ok}} + c_{\mathrm{ok}} + z_{\mathrm{ok}}}{3}, \tag{A7}$$

so the T6 primary score takes values in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. These task-specific rules are the reason Primary CGDQ is averaged across tasks only after each output has been evaluated under its own contract.

A.4 Task-Wise and Paired Hidden-Test Results

Table A1 decomposes Primary CGDQ by task type. Because MarsSite-Bench is balanced with 20 Hidden-Test instances for each of T1–T6, the aggregate score reflects all six decision modes equally. SOLVE performs strongly across the task space but does not saturate every task type: SOLVE reaches 0.9950 on T5 evidence-linked claims, whereas feasibility assessment (T2), constrained ranking (T4), and counterfactual revision (T6) remain substantially more challenging.

Table A2 reports the corresponding paired task-level comparisons. Statistical tests operate on the same 120 Hidden-Test tasks for SOLVE and each baseline. We use 10,000 Monte-Carlo sign-flip permutations and apply Holm correction across the eleven SOLVE–baseline comparisons.

All eleven standard comparisons remain statistically significant after correction. However, these paired results should be interpreted together with the matched-access control in Section 5.4: the standard suite does not perfectly align with evidence-access behavior across methods. Furthermore, the controlled study demonstrates that providing ReAct with matched access substantially reduces, but does not eliminate, the performance gap relative to SOLVE.

Table A2: Paired Hidden-Test comparisons between SOLVE and the eleven standard baselines. W/T/L denotes the numbers of tasks on which SOLVE wins, ties, or loses. Raw p-values use 10,000 Monte-Carlo sign-flip permutations; Holm correction is applied across all eleven comparisons.

Baseline	Δ Primary	W/T/L	Raw p	Holm p
ReAct	+0.4693	84/31/5	$< 10^{-4}$	1.10×10^{-3}
Multi-Agent Debate	+0.4788	86/24/10	$< 10^{-4}$	1.10×10^{-3}
CRITIC	+0.4817	83/30/7	$< 10^{-4}$	1.10×10^{-3}
CoT	+0.5042	84/33/3	$< 10^{-4}$	1.10×10^{-3}
Shared-memory	+0.5178	87/26/7	$< 10^{-4}$	1.10×10^{-3}
Majority Voting	+0.5256	87/28/5	$< 10^{-4}$	1.10×10^{-3}
Planner-only	+0.5449	90/24/6	$< 10^{-4}$	1.10×10^{-3}
Role-specialized	+0.5543	91/22/7	$< 10^{-4}$	1.10×10^{-3}
Direct Prompting	+0.5562	91/21/8	$< 10^{-4}$	1.10×10^{-3}
Self-Refine	+0.5595	92/21/7	$< 10^{-4}$	1.10×10^{-3}
Planner-Executor	+0.5634	94/19/7	$< 10^{-4}$	1.10×10^{-3}

Δ Primary is SOLVE minus the corresponding baseline. The Monte-Carlo test uses the corrected estimator $p = (k + 1)/(B + 1)$ with $B = 10{,}000$ permutations.

B SOLVE Execution and Verifiability Details

B.1 Orchestration and Shared Decision State

SOLVE maintains a shared decision state, rather than allowing role-specialized agents to produce independent answers that are aggregated only at the end. This state includes the current candidate set, acquired evidence, unresolved evidence gaps, constraint status, remaining acquisition budget, and the structured decision record. Consequently, each role operates with a common view of what has been observed and what remains unresolved.

The Planner decomposes the decision problem and identifies comparison requirements; the Evidence Scout translates unresolved information needs into admissible evidence requests; the Candidate Analyst compares candidates using the available evidence; and the Critic identifies unsupported conclusions, conflicts, or decision-relevant evidence gaps. The Constraint Verifier operates deterministically, evaluating formal constraints independently of open-ended LLM reasoning. The orchestrator coordinates these components and updates the shared state after every admissible action.

This separation is intentional. Semantic interpretation and decision-relevant comparison are delegated to LLM-based roles where flexible reasoning is beneficial, whereas evidence availability, budget accounting, protocol checks, and formal constraint satisfaction are managed through explicit interfaces or deterministic procedures.

B.2 Adaptive Evidence Acquisition and Revision

Evidence acquisition proceeds conditionally based on the evolving decision state. A requested item is executed only if it is registered, available, not already acquired, and affordable within the remaining budget. Successful acquisition updates the shared evidence state before candidate analysis is repeated. Consequently, subsequent requests may differ from the previous ones as candidate separation, uncertainty, and constraint status change.

The Critic provides a secondary mechanism for adaptive acquisition. Instead of directly changing measurements or constraint outcomes, it may identify a decision-relevant information gap and request additional support. This request is validated through the evidence interface before execution. If the evidence is available, the corresponding item can be acquired and the affected assessment revised; if it is registered but unavailable, the unavailable status is explicitly returned and must be interpreted as a limitation rather than as negative observed evidence.

Acquisition terminates when the task contract can be resolved with the current evidence, when no admissible decision-relevant requests remain, or when the remaining budget prevents further

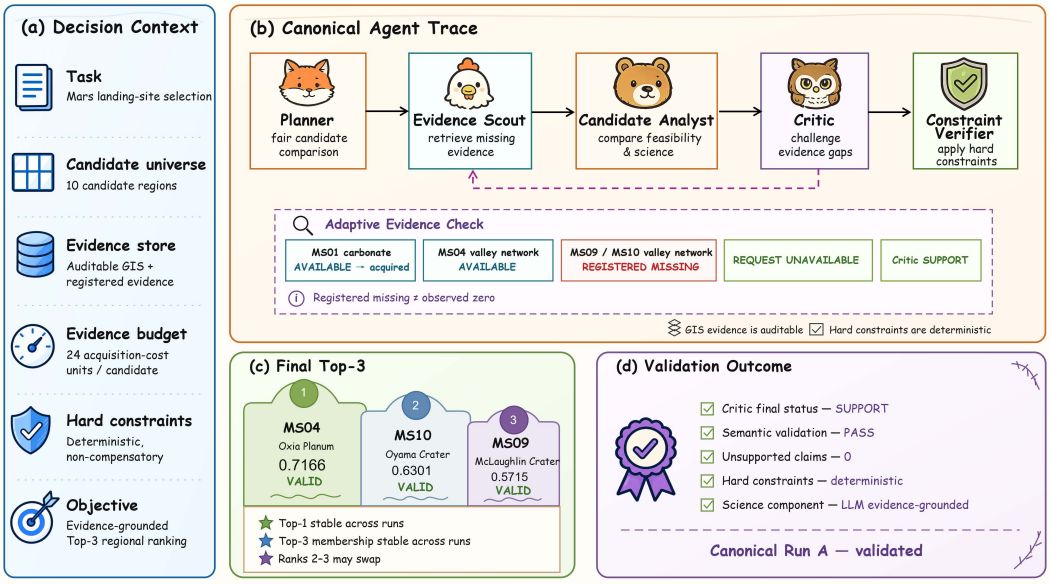

Figure A2: Representative SOLVE execution trace. Role-specialized agents operate over a shared decision context, while evidence requests update an explicit evidence state. The example also illustrates the distinction between available evidence and registered-but-unavailable evidence; unavailable evidence is retained as a limitation rather than interpreted as an observed zero.

admissible requests. Termination then transfers the current decision state to deterministic protocol and constraint checking. Figure A2 illustrates a representative execution trace.

B.3 Verifiable Decision Artifacts

SOLVE retains structured links between the final decision and the supporting evidence. Evidence is referenced via identifiers rather than reconstructed from free-form rationales, enabling downstream verification of whether a cited item was both available and accessed. Constraint outcomes are recorded separately from semantic candidate assessments, preventing a high soft score from implicitly compensating for a failed hard constraint.

The resulting artifacts capture different forms of provenance. Structured decision fields record the selected action or candidate; evidence references identify supporting observations; deterministic constraint records indicate the status of formal checks; and validation records summarize whether the output satisfies the applicable decision contract. These artifacts are designed to make the final decision inspectable without requiring access to private chain-of-thought reasoning.

Not every execution necessarily emits every conceptual artifact as a separate serialized file. In particular, a visualization derived from evidence references should be distinguished from a directly emitted artifact, and the absence of a standalone decision-certificate object must not be compensated for by inferred fields. Figure A3 explicitly illustrates this distinction for the canonical real-Mars execution.

C Additional Experimental Results

C.1 Matched-Access Controlled Study

The standard baseline comparison preserves each method's native reasoning and evidence-access behavior. This design reflects the natural operation of the methods but introduces evidence access as a potential confound. Therefore, we report a separate controlled study that modifies access conditions while preserving the corresponding reasoning structure.

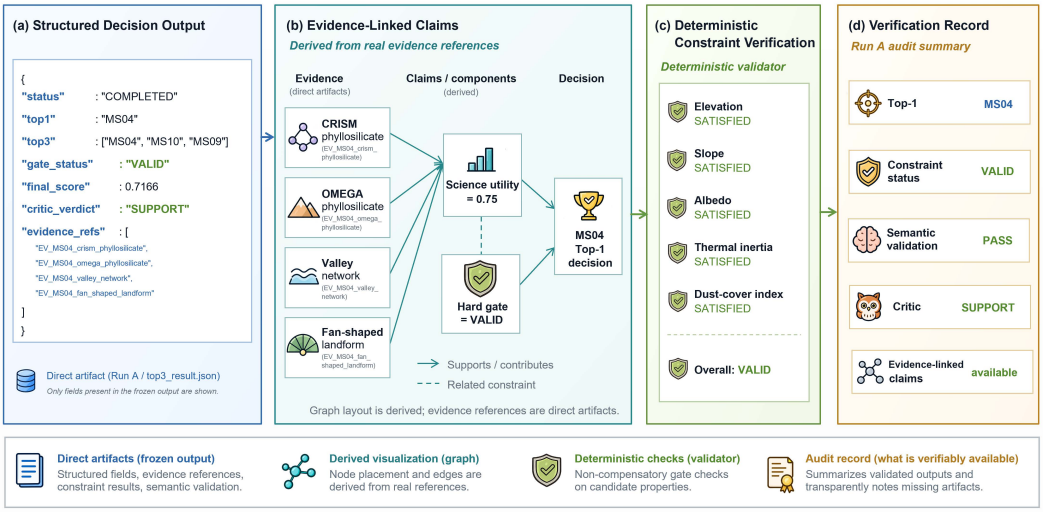

Figure A3: Evidence-linked decision artifacts for the canonical real-Mars execution. Structured output fields and evidence references are direct execution artifacts, whereas the evidence-link graph is a derived visualization of those references. Deterministic constraint checks and the available validation record are shown separately. No unavailable certificate fields are inferred or fabricated.

Tool-ReAct retains ReAct-style reasoning but adopts the evidence interface and access boundary used by SOLVE. This modification increases Primary CGDQ from 0.1795 to 0.3976, representing a paired improvement of 0.2181. Full SOLVE remains higher by 0.2512, indicating that matched evidence access explains a substantial fraction of the standard baseline gap but does not eliminate it. The residual difference should be interpreted together with the corresponding execution-validity differences and therefore cannot be attributed to orchestration alone.

The second control removes Critic-driven revision and verifier feedback while retaining SOLVE's planning, acquisition, and candidate-analysis pathway. SOLVE without verification reaches a Primary CGDQ of 0.6577, compared with 0.6488 for full SOLVE. The paired difference is small and not statistically significant ($\Delta = -0.0088$, Holm-adjusted $p = 0.719$). Thus, we do not attribute SOLVE's Primary CGDQ advantage primarily to the verification loop. Instead, the verifier provides an explicit mechanism for constraint checking, auditability, and controlled termination, properties that are only partially reflected in Primary CGDQ.

Table A3 reports the complete controlled results and paired statistics.

D Real-Mars Protocol and Reproducibility

D.1 Deterministic Engineering Scoring and Hard Gates

The real-Mars application separates deterministic engineering quantities from LLM-based scientific assessment. Terrain Safety (T), Environmental Suitability (E), and Local Surface Safety (L) are computed from registered GIS-derived measurements using fixed numerical transformations. The language model does not assign these three engineering components, modify their formulas, or override the regional hard gates. Science Utility (S) is handled separately through evidence-grounded LLM assessment, as described in Appendix D.3.

For compactness, define

$$\mathrm{clip}(x) = \min\{1, \max\{0, x\}\}. \tag{A8}$$

The following equations reproduce the engineering components used by the reported real-Mars execution.

Table A3: Matched-access controlled study on the MarsSite-Bench Hidden Test. Controlled variants are evaluated separately from the standard baseline suite.

(a) Controlled-method performance

Method	Access condition	Primary CGDQ	95% CI	Exec. valid.	Calls / task
Existing ReAct	Native ReAct access	0.1795	[0.1245, 0.2376]	85.8%	2.42
Tool-ReAct	Matched evidence access	0.3976	[0.3207, 0.4750]	57.5%	2.33
SOLVE w/o Verification	SOLVE acquisition; verification removed	0.6577	[0.5824, 0.7299]	96.7%	3.17
SOLVE	Full SOLVE	**0.6488**	[0.5722, 0.7205]	**99.2%**	4.17

(b) Paired controlled effects

Comparison	Mean Δ	95% CI	Raw p	Holm p
Tool-ReAct − Existing ReAct	+0.2181	[0.1327, 0.3060]	$< 10^{-4}$	3.00×10^{-4}
SOLVE − Tool-ReAct	+0.2512	[0.1597, 0.3424]	$< 10^{-4}$	3.00×10^{-4}
SOLVE − SOLVE w/o Verification	-0.0088	[-0.0551, 0.0373]	7.19×10^{-1}	7.19×10^{-1}

Note. Matched evidence access substantially improves ReAct but does not close the gap to SOLVE. Removing Critic-driven revision and verifier feedback does not produce a statistically significant change in Primary CGDQ. The latter result should not be interpreted as showing that verification has no functional role: explicit constraint checking and auditable termination are not fully captured by the scalar Primary CGDQ metric.

Terrain Safety. Let $\bar{s}$ denote mean slope in degrees and $\bar{h}$ mean elevation in meters. We compute

$$s_{\text{slope}} = \max\left(0, 1 - \frac{\bar{s}}{9}\right), \tag{A9}$$

$$n_{\text{elev}} = \text{clip}\left(\frac{\bar{h} - (-2800)}{-4500 - (-2800)}\right), \tag{A10}$$

$$s_{\text{elev}} = 1 - n_{\text{elev}}, \tag{A11}$$

$$T = \text{round}\left[0.5\, s_{\text{slope}} + 0.5\, s_{\text{elev}}, 4\right]. \tag{A12}$$

In the as-executed numerical scoring branch, a missing or false slope mean falls back to 0.0, whereas the corresponding elevation fallback is -2800 m. These are implementation-level numerical fallbacks and should not be interpreted as converting a registered-unavailable evidence state into an observed measurement.

Environmental Suitability. For mean surface albedo $\bar{a}$, mean thermal inertia $\overline{\text{TI}}$, and mean dust-cover index $\overline{\text{DCI}}$, the component scores are

$$A = \max\left(0, 1 - \frac{\bar{a}}{0.25}\right), \tag{A13}$$

$$I = \text{clip}\left(\frac{\overline{\text{TI}} - 100}{400 - 100}\right), \tag{A14}$$

$$D = \text{clip}\left(\frac{\overline{\text{DCI}} - 0.95}{1 - 0.95}\right), \tag{A15}$$

$$E = \text{round}\left[0.4A + 0.3I + 0.3D, 4\right]. \tag{A16}$$

The corresponding numerical fallbacks are 0.25 for albedo, 100 for thermal inertia, and 0.95 for the dust-cover index. Thus, the deterministic Environment component assigns weights $0.4/0.3/0.3$ to albedo, thermal inertia, and dust-cover index, respectively.

Local Surface Safety. Let $\bar{r}$ denote mean roughness and $\bar{c}$ the crater metric read by the regional scoring branch. We compute

$$R = \max\left(0, 1 - \frac{\bar{r}}{20}\right), \tag{A17}$$

$$C = \max\left(0, 1 - \frac{\bar{c}}{0.5}\right), \tag{A18}$$

$$L = \text{round}\left[0.6R + 0.4C, 4\right]. \tag{A19}$$

Table A4: Worked deterministic scoring example for Oxia Planum (MS04) in the canonical real-Mars execution.

Input	Value
Mean elevation	-2942.1523 m
Mean slope	1.157756°
Mean albedo	0.213357
Mean thermal inertia	336.1968 TIU
Mean dust-cover index	0.954314
Mean roughness	7.846353
Science Utility	0.7500
Terrain Safety T	0.8939
Environmental Suitability E	0.3207
Local Surface Safety L	0.7646
Science Utility S	0.7500
Final score F	$\mathbf{0.7166}$

An important runtime detail is necessary for interpreting the reported regional Local scores. The registered crater evidence contains intersection- and feature-count statistics, whereas the regional scoring branch reads a mean-valued crater field. For all ten candidates in the canonical run, this mean-valued field is unavailable to the branch, and the numerical fallback $\bar{c} = 0.0$ is used, giving $C = 1.0$ for every candidate. Consequently, candidate-to-candidate variation in the reported regional L values is driven by roughness rather than by candidate-specific crater statistics.

This regional fallback is distinct from footprint-level crater verification. Landing-footprint evaluation employs an independent GIS-derived crater-overlap constraint, which remains active and is not replaced by the regional fallback.

Non-compensatory regional gates. Engineering scores are not themselves sufficient for regional admissibility. Candidate c is marked VALID only if all five raw-evidence conditions are satisfied:

$$\begin{aligned} \overline{h}(c) &\le -2800 \text{ m}, \\ \overline{s}(c) &\le 9^\circ, \\ \overline{a}(c) &\le 0.25, \\ \overline{\text{TI}}(c) &\ge 100 \text{ TIU}, \\ \overline{\text{DCI}}(c) &\ge 0.95. \end{aligned} \tag{A20}$$

All five conditions must be satisfied. An unresolved required gate results in failure, and invalid candidates are excluded before weighted ranking. Accordingly, a high Science Utility cannot compensate for a failed engineering gate.

For a region that passes the complete gate, the final ranking score is

$$F(c) = \text{round}\left[0.20T(c) + 0.15E(c) + 0.15L(c) + 0.50S(c), 4\right], \qquad c \in \mathcal{C}_{\text{VALID}}. \tag{A21}$$

The engineering components T, E, and L are deterministic GIS-derived quantities, whereas S is the evidence-grounded Science Utility. The gate itself is not an additive term in Eq. A21; it determines whether a candidate is eligible to enter the ranking at all.

D.2 WORKED EXAMPLE: OXIA PLANUM

Oxia Planum (MS04), the canonical Top-1 region, serves as a concise case study demonstrating that the reported engineering components can be derived from GIS-based regional statistics, rather than being assigned by the LLM. The canonical run uses the inputs in Table A4.

For example,

$$s_{\text{slope}} = 1 - \frac{1.157756}{9} \approx 0.8714, \tag{A22}$$

$$n_{\text{elev}} = \frac{-2942.1523 + 2800}{-1700} \approx 0.0836, \tag{A23}$$

$$s_{\text{elev}} \approx 0.9164, \tag{A24}$$

which gives

$$T = \text{round}\left[0.5(0.8714) + 0.5(0.9164), 4\right] = 0.8939. \tag{A25}$$

Similarly, the albedo, thermal-inertia, and dust-cover transformations produce $E = 0.3207$. Because the regional crater branch uses $\bar{c} = 0.0$ in this execution, $C = 1.0$; together with $\bar{r} = 7.846353$, Eq. A19 gives $L = 0.7646$.

The final regional score is therefore

$$\begin{aligned} F(\text{MS04}) &= 0.20(0.8939) + 0.15(0.3207) + 0.15(0.7646) + 0.50(0.7500) \\ &= \mathbf{0.7166}. \end{aligned} \tag{A26}$$

D.3 Evidence Universe, Costs, and Missing-Evidence Semantics

The real-Mars evidence interface registers 20 evidence types, spanning engineering, geomorphic, mineralogical, and contextual information. This comprehensive registry should be distinguished from both the evidence actually acquired during a particular SOLVE trajectory and the subset of layers visualized in Figure A4. In particular, the figure presents 12 representative raw or derived geospatial panels, which do not enumerate the full evidence registry.

Each registered type is assigned an abstract protocol acquisition cost. The schedule includes six raster-statistic types at a cost of 1.0, five vector or geomorphic types at a cost of 1.5, eight spectral-mineral types at a cost of 2.0, and one curated-literature type at a cost of 3.0. The nominal schedule sum is therefore

$$C_{\text{nominal}} = 6(1.0) + 5(1.5) + 8(2.0) + 1(3.0) = 32.5. \tag{A27}$$

This value is the *nominal full registered evidence-universe schedule cost*; it does not reflect the cost of evidence necessarily available for every candidate. Some registered items have incomplete coverage. Moreover, these acquisition-cost units are author-defined protocol quantities, rather than monetary cost, GIS runtime, model inference cost, sensor expense, or operational mission resources.

Whereas Section 3.1 presents a general task-level budget B, the real-Mars instantiation uses a fixed non-transferable per-candidate allocation. The default acquisition budget is maintained independently for every candidate,

$$B_c = 24.0, \qquad \forall c \in \mathcal{C}, \tag{A28}$$

and unused budget cannot be transferred between candidates. An acquisition is rejected if its cost would make the corresponding remaining balance negative. The choice $24.0 < 32.5$ prevents indiscriminate acquisition of the entire registered schedule and thereby creates selective-acquisition pressure. We make no claim that 24.0 is an optimal or physically calibrated budget.

For budget-sensitivity analysis, we additionally define

$$B_{\min} = \max_{c} C_{\text{mandatory}}(c) = 17.5. \tag{A29}$$

The mandatory engineering packet contributes 7.5 units (6×1.0 raster-statistic items plus the 1.5-unit crater item), while the largest available mandatory science packet contributes 10.0 units. Thus, $B_{\min}$ is a registry-derived diagnostic minimum for the comparable mandatory packet, not an optimized operating point.

Evidence-state semantics. SOLVE explicitly distinguishes among three categories: available-but-unacquired evidence, acquired observations (positive or zero), and registered-but-unavailable coverage. The corresponding states are

$$\{\text{AVAILABLE-NOT-ACQUIRED}, \text{OBSERVED-POSITIVE}, \text{OBSERVED-ZERO}, \text{REGISTERED-MISSING}\}.$$

Table A5: Registered evidence universe for the real-Mars application. Costs are abstract acquisition-cost units defined by the protocol and are not physical, monetary, computational, or mission costs.

Category	Registered evidence type	Modality	Cost
Engineering	Elevation	Raster statistic	1.0
Engineering	Slope	Raster statistic	1.0
Engineering	Surface roughness	Raster statistic	1.0
Engineering	Surface albedo	Raster statistic	1.0
Engineering	Dust-cover index	Raster statistic	1.0
Engineering	Thermal inertia	Raster statistic	1.0
Geomorphic	Crater	Vector / geomorphic	1.5
Science	Geologic unit	Vector / geomorphic	1.5
Science	Valley network	Vector / geomorphic	1.5
Science	Ancient lake	Vector / geomorphic	1.5
Science	Fan-shaped landform	Vector / geomorphic	1.5
Mineralogy	CRISM phyllosilicate	Spectral raster	2.0
Mineralogy	CRISM carbonate	Spectral raster	2.0
Mineralogy	CRISM sulfate	Spectral raster	2.0
Mineralogy	CRISM hydrated silica	Spectral raster	2.0
Mineralogy	OMEGA phyllosilicate	Spectral raster	2.0
Mineralogy	OMEGA carbonate	Spectral raster	2.0
Mineralogy	OMEGA sulfate	Spectral raster	2.0
Mineralogy	OMEGA hydrated silica	Spectral raster	2.0
Context	Registered curated literature evidence	Text / metadata	3.0
	Registered evidence types	**20**	**32.5 total**

Note. The total 32.5 is the nominal schedule sum over all registered evidence types. It does not imply that every candidate has available coverage for all 20 types. A higher abstract cost also does not imply greater scientific importance.

Only an available item may be executed as a new acquisition. If a requested item is REGISTERED-MISSING, the request returns REQUEST-UNAVAILABLE; no acquisition cost is charged and the missing state is retained as a limitation. In particular, REGISTERED-MISSING is not converted into OBSERVED-ZERO or treated as negative scientific evidence. Implementation-level numerical fallbacks in specific engineering-score branches do not change the registry state and cannot cause an unresolved hard gate to pass.

This distinction is crucial for the Critic-driven revision loop. The Critic requests additional decision-relevant evidence, but the request is validated against the registry before execution. Available evidence can be acquired and incorporated into the revised assessment, whereas unavailable coverage must be acknowledged without repeated acquisition or semantic reinterpretation.

Science Utility. Science Utility is the only one of the four regional ranking components assigned through an evidence-grounded LLM assessment. The Analyst processes acquired science evidence and must link corresponding evidence identifiers to factual scientific claims. The rubric provides reference anchors at

$$0.00,\ 0.25,\ 0.50,\ 0.75,\ 1.00,$$

while intermediate values are permitted only when accompanied by an explicit evidence-based justification. These anchors should therefore not be interpreted as a schema that restricts the runtime to five discrete values.

The LLM does not assign Terrain Safety, Environmental Suitability, or Local Surface Safety; nor does it modify hard-gate thresholds or final ranking weights. This separation maintains flexibility in scientific interpretation while preserving deterministic engineering and constraint evaluation.

The full registered schedule is listed in Table A5.

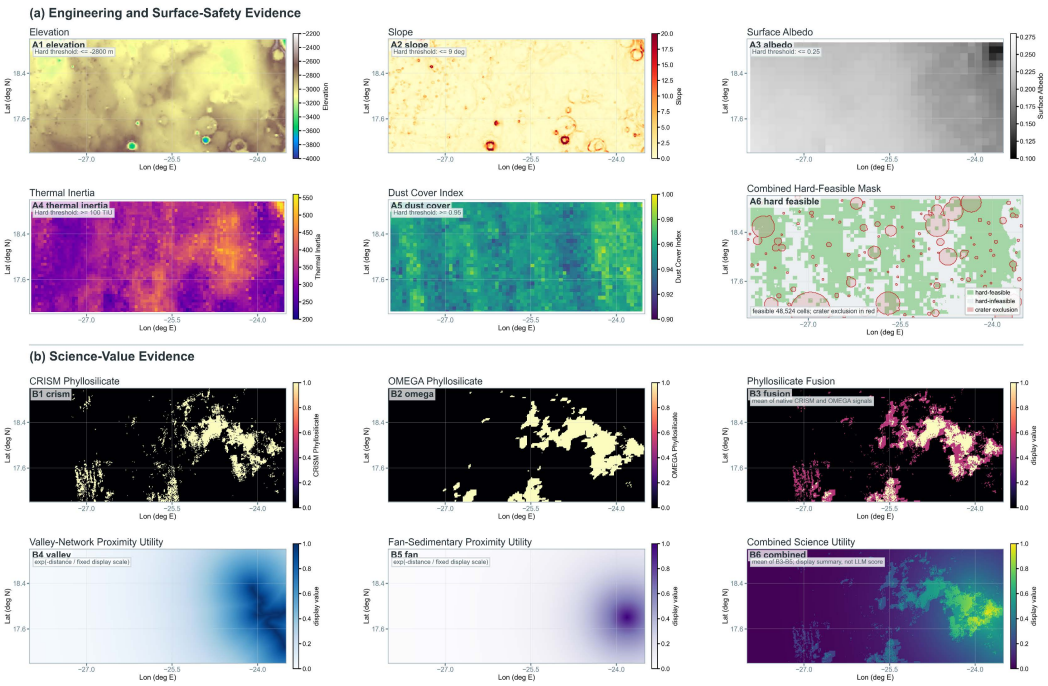

Figure A4: Representative geospatial evidence for the Oxia Planum real-Mars case. The twelve panels visualize selected raw and derived engineering and science layers, including terrain, thermophysical, mineralogical, and geomorphic evidence. They are a visualization subset and should not be interpreted as the complete registered evidence universe of 20 types. Derived displays such as the combined hard-feasible mask, mineral fusion, and combined science utility are visualization products rather than additional registered evidence types.

D.4 Real-Mars Algorithms

Algorithms A1–A5 summarize the reported real-Mars execution at a paper-facing level. They distinguish deterministic, GIS-derived engineering operations from LLM-based scientific assessment, while preserving the evidence-state and budget semantics described above. The pseudocode omits implementation details that do not influence the reported decision logic, yet retains all operations necessary for interpreting the results at both the regional and footprint levels.

For the reported canonical run, no score tie affects the Top-3. The resulting ranking is MS04 Oxia Planum, MS10 Oyama Crater, and MS09 McLaughlin Crater.

Footprint verification is deterministic and does not request a new LLM engineering score. The footprint-level crater-overlap constraint is computed directly from the corresponding GIS geometry and is therefore independent of the regional crater fallback described in Appendix D.1. If no footprint satisfies all hard constraints, the algorithm returns ABSTAIN, regardless of the parent region's regional score.

Together, Algorithms A1–A5 define two distinct decision levels. Regional selection first combines deterministic engineering gates with evidence-grounded scientific assessment to rank admissible regions. Footprint verification then re-evaluates local feasibility over an exhaustive candidate grid and may abstain even when the parent region is regionally valid. This separation prevents regional preference from overriding local engineering infeasibility.

Algorithm A1 Deterministic Engineering Scoring and Regional Hard-Gate Filtering

Require: Candidate region c; registered regional GIS evidence
Ensure: Engineering scores (T, E, L) and regional gate status $G(c)$

```
Read the registered regional engineering measurements and their availability states
Construct scoring values using only the numerical fallbacks in Appendix D.1
Compute Terrain Safety T using Eqs. A9–A12
Compute Environmental Suitability E using Eqs. A13–A16
Compute Local Surface Safety L using Eqs. A17–A19
for all five regional hard-gate variables do
    if the required raw measurement is unavailable or unresolved then
        mark the corresponding gate UNRESOLVED
    else if the raw measurement satisfies its threshold then
        mark the corresponding gate SATISFIED
    else
        mark the corresponding gate VIOLATED
    end if
end for
G(c) ← 1 iff all five gate states are SATISFIED; otherwise G(c) ← 0
status(c) ← VALID if G(c) = 1, else INVALID
return T, E, L, G(c), status(c)
```

Table A6: Complete regional results for the canonical SOLVE real-Mars execution. Weighted scores and ranks are reported only for candidates that satisfy all five non-compensatory regional hard gates.

Rank	Region	ID	Terrain	Environment	Local	Science	Final	Gate	Failed gate
1	**Oxia Planum**	MS04	0.8939	0.3207	0.7646	0.7500	**0.7166**	VALID	–
2	Oyama Crater	MS10	0.5661	0.5461	0.4000	0.7500	0.6301	VALID	–
3	McLaughlin Crater	MS09	0.2864	0.5280	0.4000	0.7500	0.5715	VALID	–
4	Ares Vallis	MS06	0.6503	0.5101	0.7750	0.2500	0.4478	VALID	–
5	Simud Valles	MS07	0.6979	0.3739	0.8365	0.2500	0.4461	VALID	–
6	Kasei Valles paleo-coastline	MS05	0.7352	0.4127	0.6697	0.2500	0.4344	VALID	–
7	Elysium Planitia	MS01	0.6760	0.2063	0.8345	0.0000	0.2913	VALID	–
–	Zhurong landing site	MS02	–	–	–	–	–	INVALID	Dust-cover index
–	Isidis Planitia	MS03	–	–	–	–	–	INVALID	Dust-cover index
–	Nili Fossae	MS08	–	–	–	–	–	INVALID	Elevation

Note. The ranking is conditional on the specified ten-region candidate universe, registered evidence, hard constraints, evidence budget, and scoring protocol. Oxia Planum is not claimed to be globally optimal over Mars. INVALID candidates do not receive compensatory weighted scores.

D.5 Complete Regional and Footprint Results

Tables A6–A8 report the complete numerical results supporting the real-Mars application in Section 6. We distinguish three complementary analyses: canonical regional ranking, stochastic repeatability and evidence-budget sensitivity, and deterministic landing-footprint verification.

The canonical execution yields seven region-level VALID candidates. Oxia Planum (MS04), Oyama Crater (MS10), and McLaughlin Crater (MS09) form the Top-3, while MS02, MS03, and MS08 are removed by non-compensatory engineering gates before score-based ranking. Table A6 reports all ten candidates, rather than only the Top-3 presented in the main paper.

The regional table also illustrates why the hard gate and the weighted ranking must remain separate. INVALID candidates are excluded before the ranking stage, while the seven VALID candidates are ordered using Eq. A21. Thus, the reported Top-3 candidates represent a ranking among admissible candidates, rather than the outcome of assigning a soft penalty to constraint violations.

The default-budget diagnostic in panel (b) should not be confused with the canonical Run A trajectory in panel (a). Run A acquired one additional MS01 CRISM-carbonate record during Critic-driven revision, resulting in 107 acquired records and a total abstract cost of 138.0 units, whereas the matched budget diagnostic contains 106 items at 136.0 units.

Algorithm A2 Per-Candidate Budgeted Evidence Acquisition

Require: Candidate set $\mathcal{C}$, evidence registry $\mathcal{E}$, per-candidate budget $B_c = 24.0$
Ensure: Acquired evidence sets $\{\mathcal{A}_c\}$ and remaining budgets $\{b_c\}$

for all $c \in \mathcal{C}$ **do**
  $\mathcal{A}_c \leftarrow \varnothing$; $b_c \leftarrow B_c$
  for all e in the mandatory engineering packet for c **do**
    $(\mathcal{A}_c, b_c) \leftarrow \textsc{Acquire}(c, e, \mathcal{A}_c, b_c)$
  end for
  Evaluate regional hard gates with Algorithm A1
  if status$(c) =$ VALID **then**
    for all e in the available mandatory science packet for c **do**
      $(\mathcal{A}_c, b_c) \leftarrow \textsc{Acquire}(c, e, \mathcal{A}_c, b_c)$
    end for
  end if
  while the controller or Critic proposes decision-relevant evidence e **do**
    if e is available, not yet acquired, and $\kappa(e) \le b_c$ **then**
      $(\mathcal{A}_c, b_c) \leftarrow \textsc{Acquire}(c, e, \mathcal{A}_c, b_c)$
    else if e is REGISTERED-MISSING **then**
      Record REQUEST-UNAVAILABLE feedback for e ▷ no cost charged
    else
      Reject the inadmissible or duplicate request
    end if
  end while
end for
return $\{\mathcal{A}_c, b_c\}_{c \in \mathcal{C}}$
procedure $\textsc{Acquire}(c, e, \mathcal{A}_c, b_c)$
  if e is available, not already acquired, and $\kappa(e) \le b_c$ **then**
    Execute the registered evidence interface
    $\mathcal{A}_c \leftarrow \mathcal{A}_c \cup \{e\}$
    $b_c \leftarrow b_c - \kappa(e)$
  end if
  return $(\mathcal{A}_c, b_c)$
end procedure

Table A7: Real-Mars repeatability and evidence-budget sensitivity. Runs A–C characterize stochastic repeatability under the same protocol, whereas the budget diagnostic varies the tested acquisition budget without making an optimality claim.

(a) Three SOLVE executions

Run	Budget	Rank 1	Rank 2	Rank 3	Critic	Validation	LLM calls
Run A	24.0	MS04	MS10	MS09	SUPPORT	PASS	5
Run B	24.0	MS04	MS09	MS10	SUPPORT	PASS	3
Run C	24.0	MS04	MS09	MS10	SUPPORT	PASS	2

(b) Evidence-budget diagnostic

Setting	Budget	Top-1	Top-3 membership	Evidence cost	Evidence items	Critic	LLM calls
$B_{\min}$	17.5	MS04	{MS04, MS09, MS10}	136.0	106	SUPPORT	3
Default	24.0	MS04	{MS04, MS09, MS10}	136.0	106	SUPPORT	3
Full schedule	32.5	MS04	{MS04, MS09, MS10}	283.0	182	SUPPORT	3

Note. Top-1 is MS04 in all three repeated runs and the Top-3 membership is unchanged, although ranks 2–3 swap between Run A and Runs B/C. $B_{\min} = 17.5$ is a derived diagnostic minimum, and 24.0 is the default acquisition budget rather than an optimized value. Evidence costs are abstract protocol units.

Across the three tested budget settings, both Top-1 and Top-3 membership remain unchanged. This supports stability over the tested range but does not establish that 24.0 is optimal or that the ranking would remain invariant under arbitrary cost schedules or substantially different budgets.

Algorithm A3 Evidence-Grounded Science Assessment and Critic Revision

Require: Valid candidate c, acquired science evidence $\mathcal{A}_c^S$, evidence registry, remaining budget b_c
Ensure: Science Utility $S(c)$, evidence-linked assessment, and Critic verdict
Candidate Analyst is conditioned on the task-visible acquired evidence record $\mathcal{A}_c^S$
Produce candidate-level scientific claims with supporting acquired-evidence identifiers
Assign provisional Science Utility $S(c) \in [0, 1]$ under the fixed evidence-grounded rubric
Permit intermediate rubric values only with explicit evidence-based justification
Critic checks unsupported claims, conflicting evidence, unresolved risks, and decision-relevant gaps
while Critic requests additional evidence e and revision remains admissible **do**
    if e is available, not yet acquired, and $\kappa(e) \leq b_c$ **then**
        Acquire e using Algorithm A2
        Update the affected claims and, if warranted, $S(c)$
    else if e is REGISTERED-MISSING **then**
        Record REQUEST-UNAVAILABLE feedback
        Preserve the unavailable item as a limitation
    else
        Reject the request
    end if
    Critic re-evaluates the revised assessment
end while
Require final Critic verdict SUPPORT before finalization
return $S(c)$, evidence references, scientific assessment, Critic verdict

Algorithm A4 Verified Regional Candidate Ranking

Require: Candidate set $\mathcal{C}$; deterministic scores $T(c), E(c), L(c)$; Science Utility $S(c)$; regional gate status $G(c)$
Ensure: Ranked admissible candidates and Top-K output
$\mathcal{C}_{\text{valid}} \leftarrow \{c \in \mathcal{C} : G(c) = 1\}$
for all $c \in \mathcal{C}_{\text{valid}}$ **do**
    $F(c) \leftarrow \text{round}[0.20T(c) + 0.15E(c) + 0.15L(c) + 0.50S(c), 4]$
end for
Exclude every candidate with $G(c) = 0$ from score-based ranking
Sort $\mathcal{C}_{\text{valid}}$ by decreasing $F(c)$ using deterministic implementation tie handling if required
Return ranked admissible candidates and requested Top-K

The two decision levels therefore answer different questions. Regional ranking identifies preferred admissible regions, whereas footprint verification assesses whether a specific landing geometry satisfies local deterministic constraints. Oxia succeeds at both levels. McLaughlin illustrates the complementary outcome: it remains region-level VALID and scientifically competitive, yet no tested footprint survives the local hard constraints, leading to the correct footprint-level decision of ABSTAIN.

Taken together, Tables A6–A8 support three bounded conclusions. First, Oxia is the Top-1 region within the specified ten-region candidate set under the stated evidence and protocol. Second, membership in the Top-1 and Top-3 regions is stable across the reported repeated executions and tested evidence budgets, without implying global or budget-independent optimality. Third, regional validity does not guarantee footprint feasibility, motivating the explicit footprint-level abstention mechanism.

E PROMPTS AND IMPLEMENTATION DETAILS

E.1 MODEL AND RUNTIME CONFIGURATION

All LLM-based methods in the standard MarsSite-Bench comparison utilize GPT-5.4-mini as the underlying language model. Method configurations were selected using the Development split before

Algorithm A5 Deterministic Landing-Footprint Verification

Require: Selected region c; requested footprint geometry (45×15 km or 55×15 km); registered footprint candidates and GIS layers
Ensure: Best feasible footprint or ABSTAIN

$\mathcal{Q}_{\text{feas}} \leftarrow \varnothing$
for all registered footprint candidates q for the requested geometry in c **do**
    Compute deterministic GIS footprint metrics
    Check geometric containment inside the region boundary
    Check valid-data ratio $r_{\text{valid}}(q) \geq 0.95$
    Check safety ratio $r_{\text{safety}}(q) \geq 0.90$
    Check crater overlap $r_{\text{crater}}(q) \leq 0.10$
    if all footprint hard constraints are satisfied **then**
        Compute the frozen footprint profile score
        Add q to $\mathcal{Q}_{\text{feas}}$
    end if
end for
if $\mathcal{Q}_{\text{feas}} = \varnothing$ **then**
    return ABSTAIN
else
    Rank feasible footprints by the frozen profile score and deterministic tie handling
    return the first-ranked feasible footprint
end if

Table A8: Landing-footprint verification for Oxia Planum and the McLaughlin Crater contrast case. Candidate counts refer to the complete tested footprint grids for each configuration.

Region	Footprint	Candidates	Feasible	Selected footprint	Azimuth	Composite score	Decision
Oxia Planum (MS04)	45×15 km	13,847	1,352	`Region_04_ellipse_45x15_013366`	130°	0.792459	SELECT
Oxia Planum (MS04)	55×15 km	12,331	945	`Region_04_ellipse_55x15_011787`	120°	0.797015	SELECT
McLaughlin Crater (MS09)	45×15 km	1,647	0	–	–	–	ABSTAIN
McLaughlin Crater (MS09)	55×15 km	1,159	0	–	–	–	ABSTAIN

Note. McLaughlin is region-level VALID but footprint-level ABSTAIN because no tested footprint satisfies all deterministic footprint constraints. The footprint composite score is used only after feasibility filtering and should not be compared directly with the region-level final score.

evaluation on the Hidden Test, and no Hidden-Test task or evaluation reference was used for prompt selection or method revision. The standard comparison preserves each method's native reasoning and interaction structure rather than forcing all baselines into the SOLVE orchestration pattern.

In SOLVE, LLM calls are separated from deterministic operations. The Planner, Evidence Scout, Candidate Analyst, and Critic perform semantic reasoning and return structured outputs, whereas evidence-interface validation, budget accounting, protocol checks, and formal constraint verification are executed deterministically. Consequently, the mean LLM-call counts reported in the main paper exclude deterministic tool and verifier operations.

The real-Mars application follows the same separation of responsibilities. Its LLM-based roles use GPT-5.4-mini with structured JSON outputs and concise task-relevant rationales. The Orchestrator and Constraint Verifier are deterministic, and GIS statistics, hard gates, regional engineering components, footprint generation, and footprint-feasibility checks are not delegated to the language model.

E.2 Role Instructions and Structured Interfaces

SOLVE uses role-specific instructions rather than a single unrestricted multi-agent conversation. Each LLM role receives the task-visible portion of the shared decision state together with a restricted output contract. The prompts define responsibilities and admissible outputs but do not provide hidden evaluator references, Gold decisions, or task-level oracle information. The role-level contracts relevant to method interpretation are summarized here; the evaluator, hard gates, and reported deterministic scores do not depend on undisclosed prompt text.

Planner. The Planner identifies unresolved decision requirements from the current state and proposes the next reasoning or evidence-acquisition need. It does not directly modify measurements, hard-constraint statuses, or evidence records.

Evidence Scout. The Evidence Scout maps decision-relevant information needs to registered evidence requests. The resulting request is checked by the evidence interface for identity, availability, duplicate acquisition, and remaining budget before execution. Consequently, a proposed request does not itself generate evidence.

Candidate Analyst. The Candidate Analyst assesses candidates based on the currently acquired evidence. Candidate-level factual claims must reference supporting evidence identifiers as required by the task contract. In the real-Mars application, this role also assigns Science Utility according to a fixed evidence-grounded rubric, while Terrain, Environment, and Local engineering components remain deterministic.

Critic. The Critic examines the current assessment for unsupported claims, conflicting observations, unresolved risks, and decision-relevant evidence gaps. It may request reassessment or additional available evidence, but it cannot create measurements, reinterpret unavailable coverage as an observed zero, alter a hard-gate result, or directly overwrite deterministic engineering scores.

Deterministic components. The Orchestrator updates the shared state and enforces admissible transitions, while the Constraint Verifier evaluates protocol and formal-constraint conditions independently of the open-ended LLM generation process. This division of labor maintains flexibility in semantic interpretation while preventing the free-form model output from directly changing the factual or constraint-based interface.

E.3 Prompt and Configuration Control

Prompt and configuration control are separated from benchmark evaluation. In Hidden-Test experiments, method configurations are finalized during the Development-stage selection and are not revised using Hidden-Test outcomes. The matched-access controlled variants are predefined before evaluation, rather than being optimized using the resulting scores.

For the reported real-Mars executions, a consistent set of parameters is applied across repeated runs, including candidate registry, ranking weights, evidence-cost schedule, per-candidate budget, hard-gate definitions, science rubric, and role-prompt identities. Prompt and configuration identities are retained in the execution records, enabling traceability to corresponding runtime states.

The real-Mars reproducibility record should be interpreted as an *as-executed final freeze*. Reported equations and results are bound to the final executed runtime, frozen input/output records, independent deterministic recomputation, and retained integrity records. It is important to note that while the final implementation is not cryptographically preregistered before canonical execution, the above measures ensure reproducibility and transparency.

E.4 Code, Data, and Release Boundaries

The publicly released MarsSite-Bench contains 480 input-only tasks: 240 Train, 120 Development, and 120 Public Test instances. The Public task files provide the participant-visible task state and structured interfaces but exclude Gold answers or private evaluation references. The 120-task Hidden Test and its evaluation references are withheld to ensure the integrity of the held-out evaluation.

The anonymous supplementary software package includes the SOLVE implementation, benchmark schemas, protocol-validation components, and de-identified examples necessary for inspecting the public interface. The release has been sanitized to remove author identities, private local paths, credentials, build caches, and Hidden-Test content. Public benchmark validation assesses the released schemas and task-facing protocol, rather than the private evaluator references used for scoring the Hidden-Test.

For the real-Mars application, raw planetary GIS products and internal execution ledgers are treated separately from the benchmark release. The appendix reports the data products, deterministic trans-

formations, computational procedures, evidence semantics, and aggregate outcomes required to audit the results presented in the paper. Independent regeneration from raw GIS products additionally requires the corresponding source datasets. Private provider transcripts, credentials, and chain-of-thought content are not part of the evaluation contract and are not required to inspect the deterministic scoring and verification logic. The artifacts provided for the paper expose structured decisions, evidence identifiers, aggregate results, deterministic equations, and verification outcomes.