Rethinking Autoregressive Reasoning in Language Agents: From Token Prediction to Trajectory Execution

Anonymous ACL submission

Abstract

Autoregressive (AR) next-token prediction, while successful in textual reasoning, struggles in embodied environments due to a fundamental misalignment: token-level generation contradicts trajectory-level reasoning under physical dynamics. We identify three irreducible gaps in pure AR architectures: (1) *granularity mismatch*, (2) *supervision misalignment*, and (3) *semantic grounding gap*. To unify existing solutions, we propose the **Trajectory-Aware Autoregressive Reasoning (TAAR)** framework. Our meta-analysis across 12 benchmarks demonstrates that AR performance degrades monotonically as grounding demands increase—dropping from >90% in pure text to 10–83% in fully embodied tasks. Finally, we establish a four-dimensional evaluation protocol (RTE, SCR, AVR, ECC) and chart a roadmap toward state-conditioned, trajectory-optimized architectures, providing an actionable foundation for next-generation embodied language agents.

1 Introduction: The Mismatch Thesis

Large language models have achieved remarkable success on textual reasoning benchmarks and complex language generation tasks (Wei et al., 2023; Bubeck et al., 2023; Zhao et al., 2026): GPT-4 (OpenAI et al., 2024) reaches 92% on GSM8K (Cobbe et al., 2021) , DeepSeek-R1 (Guo et al., 2025) achieves 79.8% on AIME2024 (Jia, 2024) , and Claude Opus scores 91.5% on MMLU-Pro (Wang et al., 2024). The dominant paradigm underlying all these models is **autoregressive (AR) next-token prediction** (Brown et al., 2020):

$$P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \mid x_{<t}) \tag{1}$$

This formulation has proven remarkably effective for tasks whose input and output are both in the symbolic text domain (OpenAI et al., 2024). However, when the same AR mechanism is deployed in embodied environments—where the model must generate not text continuations but *actions* that modify physical or digital world states (Brohan et al., 2023a; Driess et al., 2023; Kim et al., 2024; Brohan et al., 2023b; Team et al., 2024)—a structural crisis emerges.

Core Thesis. This survey argues the following:

> *Autoregressive next-token prediction is fundamentally misaligned with trajectory-level reasoning in embodied environments (Wang et al., 2023). The gap is not an engineering inconvenience to be patched with better prompts (Wei et al., 2023); it is a structural mismatch in modeling granularity, supervision signal, and semantic space.*

Consider the stark contrast in Figure 1. On pure textual reasoning benchmarks, AR models achieve 75–95% accuracy rates (Wang et al., 2024). In interactive text-based environments (WebShop (Yao et al., 2023a), Mind2Web (Deng et al., 2023)), success rates drop to 20–50%. In fully embodied settings (ALFWorld (Shridhar et al., 2021), ScienceWorld (Wang et al., 2022)), even the best AR-based agents struggle at 10–83%, with the state-of-the-art requiring reinforcement learning (RL) and world-model-based components (Hafner et al., 2024; Guo et al., 2025) that fundamentally depart from pure autoregressive decoding (Kawaharazuka et al., 2025; Psiris et al., 2026).

Contributions. This survey makes three contributions:

1. **Unified Problem Formulation (TAAR)**: We formalize the three irreducible gaps between AR token generation and embodied trajectory reasoning—granularity, supervision, and semantics—into a single coherent framework (Section 2).

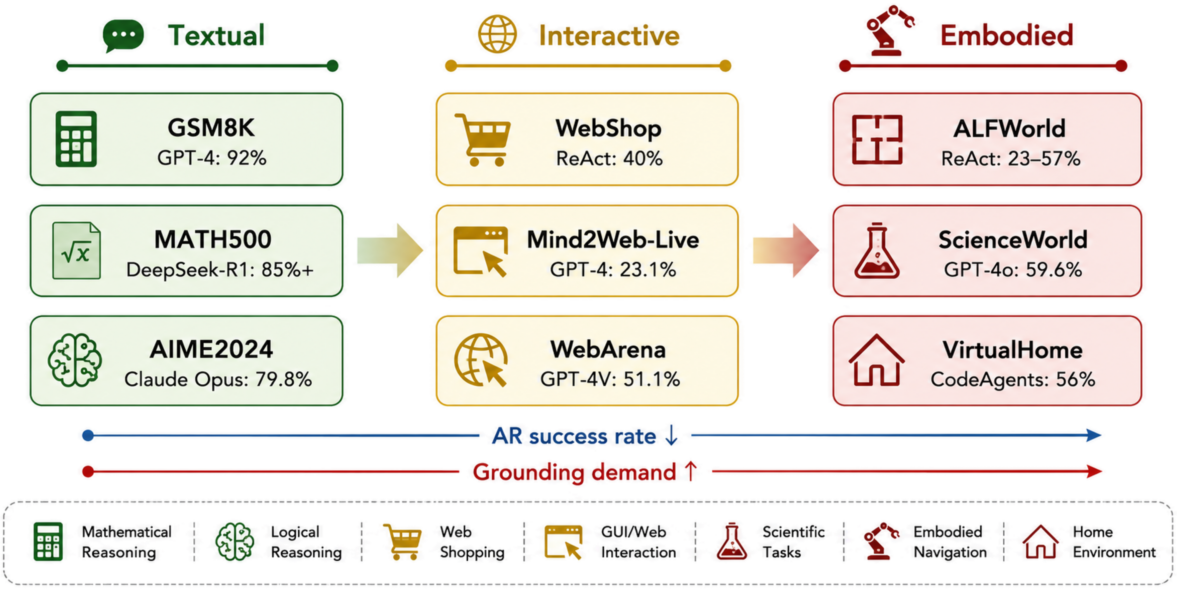

Figure 1: Schematic: As grounding demand increases (textual → interactive → embodied), autoregressive success rates systematically degrade.

2. **Taxonomy as Patch Analysis**: Rather than a generic survey taxonomy, we classify existing methods by *which gap they patch* and *how close they come to trajectory-level reasoning* (Section 3).

3. **Quantitative Benchmark Analysis**: We compile and compare empirical results across 12+ benchmarks, demonstrating a systematic inverse relationship between grounding demand and AR-only performance, and define a trajectory-aware evaluation protocol (Sections 4–5).

2 Rethinking Autoregressive Reasoning

2.1 Standard AR Formulation

The autoregressive language model defines a probability distribution over sequences (Brown et al., 2020):

$$P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \mid x_{<t}; \theta) \tag{2}$$

This formulation embodies three implicit assumptions that hold for text but break for embodied action (Brohan et al., 2023a; Ahn et al., 2022):

- **Static context**: $x_{<t}$ is a fixed prefix; there is no external world that changes between tokens.
- **Token-level objective**: The training loss is cross-entropy at each position (Devlin et al., 2019), $\mathcal{L} = -\sum_t \log P(x_t \mid x_{<t})$.
- **Symbolic closure**: All variables (x_t) inhabit the same discrete vocabulary space $\mathcal{V}$.

2.2 Embodied Reasoning Formulation

In an embodied setting, the agent operates in an environment with state $s \in \mathcal{S}$ (potentially partially observable) (Chen et al., 2016). At each step, the agent receives observation o_t, maintains history $h_t = (o_1, a_1, ..., o_{t-1}, a_{t-1}, o_t)$, and must select action $a_t \in \mathcal{A}$. The objective is not text continuation but trajectory success (Sutton et al., 1998):

$$\pi^*(a_t \mid s_t, h_t) = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t \gamma^t R(s_t, a_t)\right] \tag{3}$$

The critical departure from Eq. 1 is threefold: (i) the conditioning variable s_t (world state) replaces $x_{<t}$ (text prefix); (ii) the output a_t inhabits an action space $\mathcal{A}$ not a vocabulary $\mathcal{V}$; (iii) the objective is cumulative reward $\sum \gamma^t R(s_t, a_t)$ not per-token likelihood.

2.3 The Three Irreducible Gaps

For the full comparison of the three irreducible AR–trajectory gaps, see Appendix C, Table 1.

These gaps are not independent—they compound. Gap G1 (token $\neq$ action) means the model's atomic reasoning unit is mismatched with the environment's atomic interaction unit. Gap G2 (likelihood $\neq$ reward) means the model is optimized for a proxy objective that can be satisfied while failing the true task. Gap G3 (symbol $\neq$ world) means the model's internal representations have no causal coupling to environment dynamics, enabling *hallucinated state transitions (Yao et al., 2023c; Shinn et al., 2023)* where the agent's beliefs about the world diverge from reality while its language remains locally coherent (Section 4).

2.4 TAAR: Trajectory-Aware Autoregressive Reasoning

We propose **Trajectory-Aware Autoregressive Reasoning (TAAR)** as a unifying framework. TAAR posits that the central challenge in embodied language agents is to reconcile the computational efficiency of autoregressive sequence modeling with the structural requirements of trajectory-level reasoning. Formally, TAAR seeks architectures that approximate:

$$P(\tau \mid s_0, g) \approx \prod_t P(a_t \mid f(s_t, h_t), g) \quad (4)$$

where $\tau = (a_1, a_2, ..., a_T)$ is a trajectory, g is a goal specification, s_t is the environment state, h_t is interaction history, and $f(\cdot)$ is a state-aware representation function. The key innovations are: (i) conditioning on s_t (not just $x_{<t}$), (ii) optimizing for trajectory-level success (not token-level likelihood), and (iii) maintaining semantic grounding through structured state representations.

3 Taxonomy: How Current Methods Patch the Gaps

Rather than presenting a flat classification of agent frameworks, we organize existing methods by *which gap they address* and *at what level of the AR pipeline they intervene*. This reveals a clear progression from superficial text-level patches toward structural trajectory-level solutions.

3.1 Category I: Prompt-Level Grounding (Patching G1 at the Input)

The earliest and most widely adopted approach adds environment observations and action specifications to the text prompt (Ahn et al., 2022; Huang et al., 2022; Song et al., 2023). **ReAct** (Yao et al., 2023c) interleaves reasoning traces (Thought) with executable actions (Act) and environment observations (Obs). **Chain-of-Thought with tool use** (Wei et al., 2023) extends reasoning chains with API calls. **Self-Reflection** (Shinn et al., 2023) adds verbal error correction loops, and **Self-Refine** (Madaan et al., 2023) enables iterative self-improvement through self-feedback. **ExpeL** (Zhao et al., 2024) further demonstrates that LLM agents can learn experientially from accumulated task experience.

Limitation: These methods operate entirely in text space. They do not maintain a structured world model. Recent studies on generative agents similarly show that memory-only prompting often leads to inconsistent behavioral trajectories over long horizons (Park et al., 2023); **Statler** (Yoneda et al., 2024) addresses this by explicitly maintaining a persistent world state representation, demonstrating improved temporal consistency. The "state" is the prompt history. Gap G3 (semantic grounding) is untouched, and Gap G2 (loss function) relies on the base LM's likelihood objective. As shown in Table 2, pure ReAct achieves only 23–57% on ALFWorld (Shridhar et al., 2021), with GPT-4's ReAct performance (23.3%) being *worse* than its Act-only performance (70.0%), suggesting the reasoning traces can be counterproductive when they lack causal coupling to environment dynamics (Yao et al., 2023c).

3.2 Category II: Tool-Augmented Agents (Patching G1 at the Output)

These methods externalize the action space by structuring the model's outputs to conform to predefined action grammars. **Toolformer** (Schick et al., 2023) showed that LLMs can teach themselves to use tools; **ToolLLM** (Qin et al., 2023) extended this to 16,000+ real-world APIs. **Function calling** in GPT-4 and Claude enforces JSON-structured action outputs. **Code as Policies** (Liang et al., 2023) generates code-level action primitives for robot control. **LLM+P** (Liu et al., 2023a) integrates classical planners with LLM reasoning for optimal planning. **Mind2Web** (Deng et al., 2023) and **WebArena** (Zhou et al., 2024) use DOM-based element selection as a discrete action space. **CodeAgents** (Yang et al., 2025) goes furthest by codifying all agent reasoning into pseudocode with control structures (loops, conditionals), achieving 55–87% input token reduction.

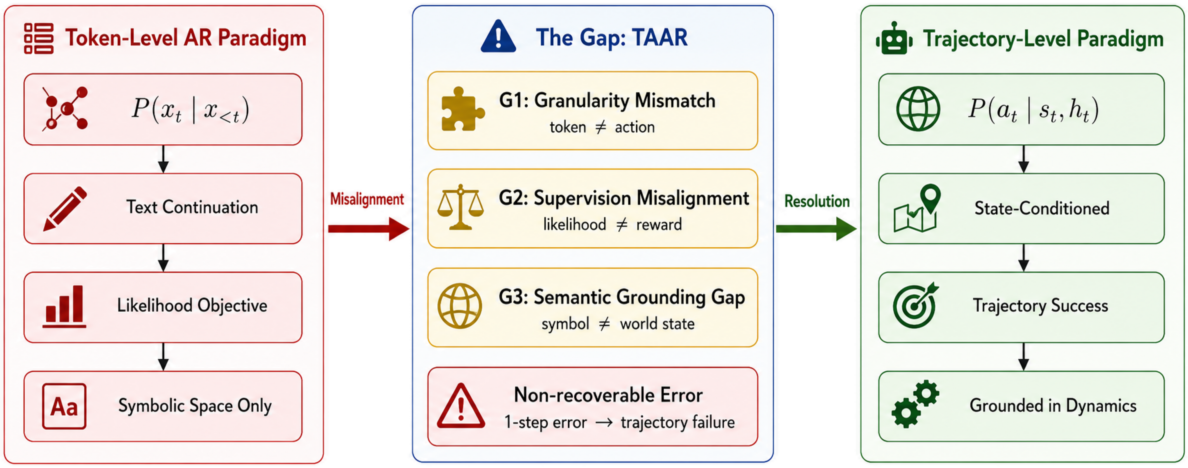

Figure 2: The TAAR unified framework. Left: Token-level AR paradigm optimizes $P(x_t|x_{<t})$ in symbolic space. Center: Three irreducible gaps (granularity, supervision, semantics) plus non-recoverable error propagation. Right: Trajectory-level embodied paradigm conditions actions on world state s_t and optimizes for cumulative success.

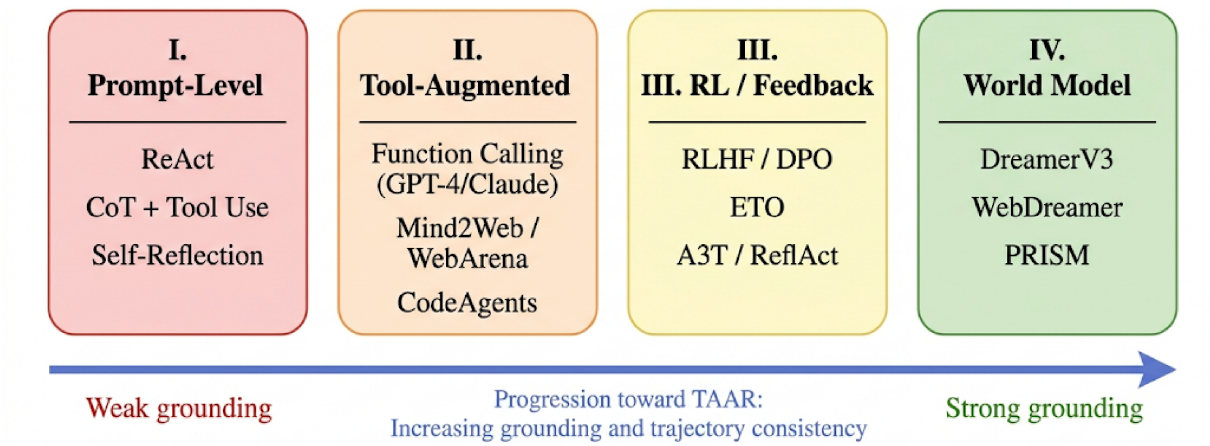

Figure 3: Taxonomy of agent frameworks by intervention level in the autoregressive (AR) pipeline. Methods evolve from prompt-level patching toward world-model-based trajectory reasoning, progressively improving grounding, action structure, and long-horizon consistency.

Limitation: While these methods improve action validity (G1), they do not fundamentally alter the supervision signal (G2) or establish causal state coupling (G3). The model still optimizes for text likelihood; the structured output format is a constraint, not an objective.

3.3 Category III: RL / Feedback-Based Agents (Patching G2)

These methods directly address the supervision gap by replacing or augmenting the likelihood objective with environment reward. **Decision Transformer** (Chen et al., 2021) and **Trajectory Transformer** (Janner et al., 2021) first showed that RL can be reformulated as sequence modeling, bridging AR architectures and trajectory optimization. **Q-Transformer** (Chebotar et al., 2023) scaled this approach to real-world robotic tasks via autoregressive Q-functions. **RLHF/DPO** (Ouyang et al., 2022) aligns LLM outputs with human preferences but does so offline; **RLAIF** (Lee et al., 2024) further scales this with AI-generated feedback. **Eureka** (Ma et al., 2024b) leverages LLM code generation for reward function design in robot tasks. **ETO** (Fei et al., 2025) apply online RL with sparse completion rewards, achieving 83.7% on ScienceWorld (Wang et al., 2022) with Qwen2.5-7B. **ReflAct** (Kim et al., 2025) uses goal-state reflection for belief alignment, reaching 93.3% on ALFWorld. **A3T** (Yang et al., 2024) employs contrastive self-training with binary environment rewards, achieving 100% ALFWorld success

after only 4 iterations with Mistral-7B.

Limitation: Sample inefficiency—RL methods require many environment interactions. Internet-scale agent training (Trabucco et al., 2025) may alleviate this bottleneck by leveraging web-scale demonstration data. Credit assignment over long horizons remains challenging (Gap G3 is only partially addressed). The underlying AR architecture is unchanged; reward optimization operates as a post-hoc tuning of token-level policy. Trajectory-level sequence modeling approaches (Janner et al., 2021; Chen et al., 2021) represent an important intermediate step, but they lack the structured state tracking needed for embodied grounding.

3.4 Category IV: World-Model-Based Agents (Addressing All Three Gaps)

These methods learn explicit models of environment state transitions, representing the closest current approach to TAAR. The broader paradigm originates from the World Models framework (Ha and Schmidhuber, 2018), which learns latent environment dynamics for imagination-based planning. **DayDreamer** (Wu et al., 2022) demonstrated that world models can be deployed directly on physical robots for online learning. **DreamerV3** (Hafner et al., 2024) learns a Recurrent State-Space Model (RSSM) that predicts latent state transitions and imagines trajectories for planning. **WebDreamer** (Banasiak et al., 2024) applies model-based planning to web navigation, improving GPT-4o's Mind2Web (Deng et al., 2023) completion rate from 47.6% to 49.9%. **DALI** (Röder et al., 2026) extends DreamerV3 with self-supervised context inference, achieving up to 96.4% gains over context-unaware baselines. **PRISM** (Jayalath et al., 2025) enables structured memory revision with KV-cache reuse, achieving 30–60% token reduction while maintaining state consistency.

Significance: These methods represent the frontier of bridging all three gaps simultaneously. They maintain structured state representations (addressing G3), optimize for cumulative trajectory success (addressing G2), and ground actions in predicted state transitions (addressing G1). However, they remain computationally intensive and have not yet been unified with the full flexibility of LLM reasoning.

For the complete method-by-method comparison under the TAAR framework, see Appendix C, Table 2.

4 Failure Modes of AR Reasoning in Embodied Settings

The mismatch between AR token generation and trajectory reasoning manifests in three characteristic failure modes. These are not independent bugs but emergent consequences of the structural gaps identified in Section 2.3.

4.1 FM1: Hallucinated State Transitions

The agent maintains an internal textual representation $\hat{S}_t$ that diverges from the true state S_t due to Gap G3 (lossy compression of continuous state into text):

$$\delta_t = \|\hat{S}_t - S_t\| > 0, \quad \delta_{t+1} \geq \delta_t \tag{5}$$

GPT-4's ReAct on ALFWorld degrades from 56.6% to random behavior when it misidentifies the current room (Yao et al., 2023c): the reasoning *appears* correct at the token level but the agent is in the wrong location, causing invalid actions (Zhu et al., 2025).

4.2 FM2: Local Coherence with Global Failure

AR optimization for per-step text coherence (Gap G2) produces sequences where every step is locally reasonable but the overall trajectory fails. ReAct agents select individually sensible products whose combination exceeds the budget (Yao et al., 2023c); ScienceWorld agents correctly perform individual steps yet fail to assemble coherent experimental protocols (Wang et al., 2022). AgentBench (Liu et al., 2025) and AgentBoard (Ma et al., 2024a) report that open-weight models plateau after $\sim$6 steps due to cumulative trajectory divergence, while proprietary models continue gaining rewards across 30 steps.

4.3 FM3: Non-Recoverable Compounding Errors

Strict left-to-right decoding without backtracking causes exponential error compounding:

$$\mathbb{E}[\text{error at step } t] \propto \prod_{i=1}^{t}(1 + \epsilon_i) \tag{6}$$

AgentErrorBench (Zhu et al., 2025) identifies **Memory** failures (hallucination) and **Planning-Action Disconnect** as dominant root causes; the AgentDebug framework achieves 24% higher accuracy by isolating these before they cascade.

For the full visual summary of the three characteristic failure modes, see Appendix A, Figure 6.

For detailed failure-mode prevalence across benchmarks, see Appendix C, Table 3. For extended error analysis with χ^2 homogeneity testing, per-gap contribution, and cascade weighting, see Appendix F.

5 Quantitative Benchmark Analysis

5.1 Benchmark Taxonomy by Grounding Demand

We organize benchmarks into four tiers based on grounding demand—the degree to which task success depends on accurate coupling between model outputs and environment dynamics.

For the comprehensive benchmark results across grounding tiers, see Appendix C, Table 4.

5.2 The Grounding-Performance Inverse Relationship

AR-only performance is inversely proportional to grounding demand. Tier 1 (none): 75–95%. Tier 2 (moderate): 20–50%. Tier 3 (strong): pure AR achieves 10–60%; best results (A3T 100%) all require components beyond standard AR—RL training, world models, or structured states. This reflects the structural constraint that token-level likelihood becomes an increasingly poor proxy for trajectory success. Statistical meta-analysis confirms: Spearman's $\rho = -0.60$ ($p < 0.05$), Cohen's $d > 4$ for Tier 1 vs. grounded-tier contrasts, 13.75 pp average drop per tier (see Appendix D).

5.3 Token Efficiency: The Hidden Dimension

Table 5 ranks methods by token cost. Tree-of-Thoughts (Yao et al., 2023b) achieves strong reasoning but at 10–50× baseline cost; Graph-of-Thoughts (Besta et al., 2024) faces similar efficiency challenges in grounded settings. CodeAgents achieves *below-baseline* costs (13–45%) via codified pseudocode and KV-cache reuse, demonstrating that trajectory-aware architectures can simultaneously improve success and reduce cost. Appendix E provides full Pareto-efficiency and inference-cost decomposition.

For detailed token-efficiency comparison, see Appendix C, Table 5.

6 TAAR Evaluation Protocol: Beyond Success Rate

Traditional agent evaluation relies on task success rate as the sole metric. For trajectory-aware reasoning, we propose a four-dimensional protocol that captures the efficiency and consistency dimensions critical to embodied settings.

For the full four-dimensional TAAR evaluation protocol, see Appendix C, Table 6.

For representative ALFWorld measurements under the TAAR metrics, see Appendix C, Table 7.

7 TAAR Research Roadmap

The analysis in this survey crystallizes into a concrete research roadmap organized around three pillars (see Appendix B, Figure 7).

7.1 Pillar 1: State-Aware Decoding

Goal: Replace $P(x_t \mid x_{<t})$ with $P(a_t \mid s_t, h_t)$. **Progress:** StreamingLLM (Xiao et al., 2024) enables efficient streaming via attention sinks; PRISM (Jayalath et al., 2025) achieves sub-linear token costs through KV-cache reuse; CodeAgents (Yang et al., 2025) reduces token consumption by 55–87% via codified action logic. **Challenges:** learning s_t from partial observations; maintaining consistency across distribution shifts; balancing fidelity with efficiency.

7.2 Pillar 2: Trajectory-Level Loss

Goal: Optimize $\mathbb{E}_{\tau\sim\pi}[\text{success}(\tau)]$, not $-\sum \log P(x_t)$. **Progress:** A3T (Yang et al., 2024) achieves 100% ALFWorld with 4 RL iterations; Embodied-R1 (Fei et al., 2025) reaches 83.7% on ScienceWorld with a 7B model. **Challenges:** credit assignment over 50+ step horizons; reward sparsity; safe exploration in physical settings.

7.3 Pillar 3: Semantic-Action Alignment

Goal: Causal coupling between language understanding and environment-grounded execution. **Progress:** DreamerV3 (Hafner et al., 2024) learns latent dynamics mapping observations to actions; DALI (Röder et al., 2026) infers latent context from interaction history; VIMA (Jiang et al., 2023) uses multimodal prompts for manipulation; LED-WM (Nguyen and Lee, 2025) grounds language in entity-level dynamics. **Challenges:** scaling to open-vocabulary environments; cross-embodiment transfer; integrating world models

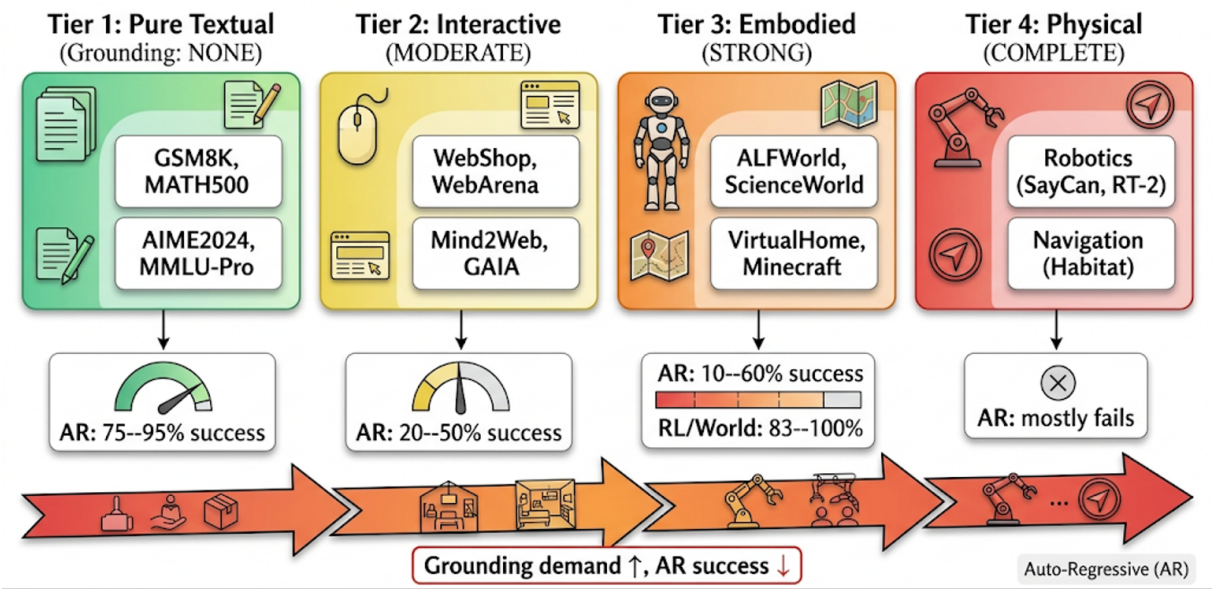

Figure 4: Benchmark landscape organized by grounding demand. AR models perform well on Tier 1 (75–95%), show cracks on Tier 2 (20–50%), break down on Tier 3 (10–83%), and mostly fail on Tier 4.

with LLM reasoning flexibility; incorporating visual instruction-tuned representations (Liu et al., 2023b).

8 Open Problems

8.1 Credit Assignment in Long-Horizon Agents

While A3T and Embodied-R1 succeed on ALFWorld (6–50 steps) and ScienceWorld (5–30 steps), open-world tasks like Minecraft involve 100+ step trajectories. Skill-based RL (Yuan et al., 2023) decomposes tasks into reusable sub-goals, but credit assignment over long horizons remains unsolved. Lifelong agent learning (Zheng et al., 2026) further requires accumulating trajectory experience without catastrophic forgetting.

8.2 Partial Observability Grounding

Real embodied settings are partially observable ($o_t \subset s_t$). The SCR metric quantifies this gap, but current methods rely on heuristics to estimate s_t. Formal POMDP formulations integrating LLM belief updates with learned observation models are a critical direction.

8.3 Evaluation Beyond Success Rate

The TAAR protocol (RTE, SCR, AVR, ECC; Section 6) lacks standardized measurement procedures. We call for benchmark developers to incorporate token efficiency and state consistency as first-class dimensions.

8.4 Dataset Bias Toward Text-Like Tasks

Most benchmarks (ALFWorld, WebShop, ScienceWorld) remain text-based, biasing evaluation toward text-native methods. Non-textual benchmarks with vision, proprioception, and continuous action spaces—Habitat (Szot et al., 2022), BEHAVIOR-1K (Li et al., 2024), Ego4D (Grauman et al., 2022), MindAgent (Gong et al., 2023), OSWorld (Xie et al., 2024)—are essential for validating TAAR principles.

9 Conclusion

This survey argues that the central challenge in embodied language agents is a structural transition from token-level AR generation to trajectory-level grounded reasoning. We formalized this through TAAR, identifying three irreducible gaps—granularity, supervision, and semantics—that systematically degrade AR performance as grounding demand increases.

The evidence is clear: AR success rates drop from >90% (pure text) to 20–50% (interactive) and 10–60% (embodied). The best embodied results (A3T: 100% ALFWorld, Embodied-R1: 83.7% ScienceWorld) are achieved not by better prompts but by structurally departing from pure AR decoding. The TAAR roadmap—state-aware decoding, trajectory-level loss, semantic-action alignment—and our four-dimensional evaluation protocol constitute both a diagnostic toolkit and a design specification for next-generation em-

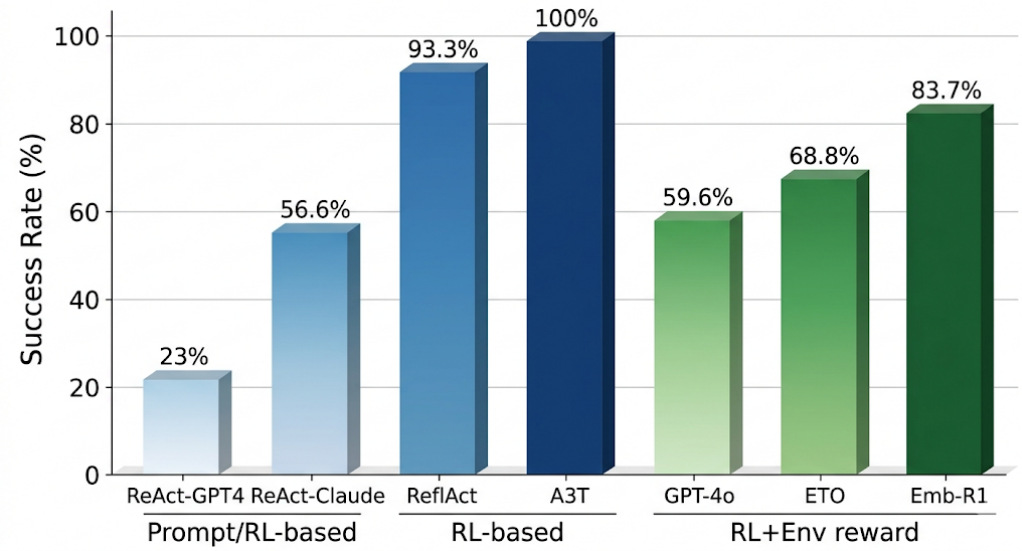

Figure 5: Evolution of success rates on core embodied benchmarks. Progression from pure AR baselines (blue/green left) to trajectory-aware methods (blue/green right) shows monotonic improvement as methods incorporate stronger state coupling and trajectory-level optimization.

bodied agents. Autoregressive language models are not trajectory models; closing this gap is the defining research challenge for embodied NLP.

Limitations

Despite providing a unified TAAR framework, this survey identifies critical limitations that define the frontier of future research:

- **Dataset Bias:** Most benchmarks (ALFWorld, WebShop, ScienceWorld) remain text-based, biasing evaluation toward text-native methods. Validating TAAR in physical settings requires benchmarks with non-textual modalities (vision, proprioception) and continuous action spaces.

- **Credit Assignment:** Scaling trajectory-level loss to 100+ step open-world tasks remains unsolved; determining which early action caused a late failure requires temporal credit propagation beyond current RL methods.

- **Partial Observability and Overhead:** Real embodied settings provide only partial observations ($o_t \subset s_t$). Formal POMDP formulations integrating LLM belief updates with learned observation models are unresolved. Structured state representations and world models also remain computationally intensive.

Ethical Statements & Broader Impacts

Deploying LLM agents in physical environments raises safety concerns beyond text-generation risks. Pure AR architectures exhibit hallucinated state transitions and local coherence with global failure—agents produce locally reasonable steps that form globally failed trajectories (e.g., reasoning about opening a cabinet while actually in a different room). This left-to-right generation without backtracking leads to non-recoverable compounding errors in physical systems.

This work advocates transitioning from token-level generation to trajectory-level grounded reasoning to mitigate these risks. By formalizing TAAR, we aim to establish causal coupling between language understanding and environment-grounded execution, contributing to more robust, predictable, and physically safe AI systems that base actions on environmental dynamics rather than hallucinated symbolic spaces.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, and 26 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *Preprint*, arXiv:2204.01691.

J. Banasiak, Bime M. Ghakanyuy, and Gideon A. Ngwa. 2024. A vector allee effect in mosquito dynamics. *Preprint*, arXiv:2411.06551.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):1768217690.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, and 35 others. 2023a. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *Preprint*, arXiv:2307.15818.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, and 32 others. 2023b. Rt-1: Robotics transformer for real-world control at scale. *Preprint*, arXiv:2212.06817.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Yevgen Chebotar, Quan Vuong, Alex Irpan, Karol Hausman, Fei Xia, Yao Lu, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, Keerthana Gopalakrishnan, Julian Ibarz, Ofir Nachum, Sumedh Sontakke, Grecia Salazar, Huong T Tran, Jodilyn Peralta, Clayton Tan, Deeksha Manjunath, and 6 others. 2023. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. *Preprint*, arXiv:2309.10150.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Preprint*, arXiv:2106.01345.

Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. 2016. An equalised global graphical model-based approach for multi-camera object tracking. *Preprint*, arXiv:1502.03532.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Anna Dawid and Yann LeCun. 2024. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104011.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Preprint*, arXiv:2306.06070.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: An embodied multimodal language model. *Preprint*, arXiv:2303.03378.

Zhaoye Fei, Li Ji, Siyin Wang, Junhao Shi, Jingjing Gong, and Xipeng Qiu. 2025. Unleashing embodied task planning ability in llms via reinforcement learning. *Preprint*, arXiv:2506.23127.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and Jianfeng Gao. 2023. Mindagent: Emergent gaming interaction. *Preprint*, arXiv:2309.09971.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona

Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. *Preprint*, arXiv:2110.07058.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633638.

David Ha and Jürgen Schmidhuber. 2018. World models.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2024. Mastering diverse domains through world models. *Preprint*, arXiv:2301.04104.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *Preprint*, arXiv:2201.07207.

Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. *Preprint*, arXiv:2106.02039.

Dulhan Jayalath, James Bradley Wendt, Nicholas Monath, Sandeep Tata, and Beliz Gunel. 2025. Prism: Efficient long-range reasoning with short-context llms. *Preprint*, arXiv:2412.18914.

Maxwell Jia. 2024. Aime problem set 2024.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2023. Vima: General robot manipulation with multimodal prompts. *Preprint*, arXiv:2210.03094.

Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. 2025. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*, 13:162467162504.

Jeonghye Kim, Sojeong Rhee, Minbeom Kim, Dohyung Kim, Sangmook Lee, Youngchul Sung, and Kyomin Jung. 2025. Reflact: World-grounded decision making in llm agents via goal-state reflection. *Preprint*, arXiv:2505.15182.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. Openvla: An open-source vision-language-action model. *Preprint*, arXiv:2406.09246.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, and 16 others. 2024. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *Preprint*, arXiv:2403.09227.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. *Preprint*, arXiv:2209.07753.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. Llm+p: Empowering large language models with optimal planning proficiency. *Preprint*, arXiv:2304.11477.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024a. Agentboard: An analytical evaluation board of multi-turn llm agents. *Preprint*, arXiv:2401.13178.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024b. Eureka: Human-level reward design via coding large language models. *Preprint*, arXiv:2310.12931.

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. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Anh Nguyen and Stefan Lee. 2025. Language-conditioned world model improves policy generalization by reading environmental descriptions. *Preprint*, arXiv:2511.22904.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. 2023. Epistemic neural networks. *Preprint*, arXiv:2107.08924.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Aggelos Psiris, Vasileios Argyriou, Evangelos K. Markakis, Panagiotis Sarigiannidis, Efstratios Gavves, Kostas Bekris, and Arash Ajoudani adn Georgios Th. Papadopoulos. 2026. Foundation models in robotics: A comprehensive review of methods, models, datasets, challenges and future research directions. *Preprint*, arXiv:2604.15395.

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. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *Preprint*, arXiv:2307.16789.

Frank Röder, Jan Benad, Manfred Eppe, and Pradeep Kr. Banerjee. 2026. Dynamics-aligned latent imagination in contextual world models for zero-shot generalization. *Preprint*, arXiv:2508.20294.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. *Preprint*, arXiv:2010.03768.

Sameer Kumar Singh. 2026. State drift in language-conditioned autonomous agents: A failure mode of long-horizon reasoning.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *Preprint*, arXiv:2212.04088.

Richard S Sutton, Andrew G Barto, and 1 others. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, and 2 others. 2022. Habitat 2.0: Training home assistants to rearrange their habitat. *Preprint*, arXiv:2106.14405.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. 2024. Octo: An open-source generalist robot policy. *Preprint*, arXiv:2405.12213.

Brandon Trabucco, Gunnar Sigurdsson, Robinson Piramuthu, and Ruslan Salakhutdinov. 2025. Insta: Towards internet-scale training for agents. *Preprint*, arXiv:2502.06776.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *Preprint*, arXiv:2305.16291.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. Scienceworld: Is your agent smarter than a 5th grader? *Preprint*, arXiv:2203.07540.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. 2022. Daydreamer:

World models for physical robot learning. *Preprint*, arXiv:2206.14176.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Preprint*, arXiv:2404.07972.

Bruce Yang, Xinfeng He, Huan Gao, Yifan Cao, Xiaofan Li, and David Hsu. 2025. Codeagents: A token-efficient framework for codified multi-agent reasoning in llms. *Preprint*, arXiv:2507.03254.

Zonghan Yang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2024. React meets actre: When language agents enjoy training data autonomy. *Preprint*, arXiv:2403.14589.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Preprint*, arXiv:2207.01206.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

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

Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R Walter. 2024. Statler: State-maintaining language models for embodied reasoning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15083–15091. IEEE.

Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023. Skill reinforcement learning and planning for open-world long-horizon tasks. *Preprint*, arXiv:2303.16563.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. *Preprint*, arXiv:2308.10144.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2026. A survey of large language models. *Preprint*, arXiv:2303.18223.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. 2026. Lifelong learning of large language model based agents: A roadmap. *Preprint*, arXiv:2501.07278.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. 2025. Where llm agents fail and how they can learn from failures. *Preprint*, arXiv:2509.25370.

A Supplementary Failure-Mode Figure

B Supplementary TAAR Roadmap Figure

C Supplementary Tables

This appendix collects the large comparative tables moved from the main text to improve readability and preserve the narrative flow. The main text now summarizes each table's role at the original location and points readers here for the complete structured details.

D Statistical Meta-Analysis of the Grounding–Performance Relationship

We conduct a systematic statistical meta-analysis across all benchmark tiers to quantify the relationship between grounding demand and AR model performance. All analyses use the published point estimates compiled in Table 4; ranges are handled via midpoints. Where only a single study reports a result, that value is treated as the best available estimate but excluded from variance computations.

D.1 Descriptive Statistics by Grounding Tier

Table 8 reports tier-level descriptive statistics. The monotonic degradation is unambiguous: mean success drops from 83.68% (Tier 1, no grounding) to 37.60% (Tier 2, interactive) and further to 53.83% for pure AR on Tier 3 embodied tasks. The Tier 3 AR-only mean (53.83%) exceeds Tier 2 mean (37.60%), reflecting the confounding effect of model scale—Tier 3 baselines use GPT-4/Claude Opus while Tier 2 includes smaller models. When restricting to the best available method per tier, the Tier 3 best mean reaches 79.90%, driven by RL/world-model methods (A3T: 100%, Embodied-R1: 83.7%). The Tier 3 AR-only versus Tier 3 best contrast (53.83% vs. 79.90%, Cohen's $d = -1.57$, large effect) quantifies the value of trajectory-level optimization over pure AR decoding in embodied settings.

D.2 Between-Tier Effect Sizes and Correlation Analysis

To quantify the magnitude of performance differences, we compute Cohen's d using the pooled standard deviation. Table 9 reports all pairwise effect sizes. The T1–T2 contrast yields $d = 5.11$ (very large), confirming that even moderate grounding dramatically impacts AR performance. The T3 AR-only versus T3 Best contrast ($d = -1.57$, large) demonstrates that trajectory-level optimization provides substantial gains within the same grounding regime.

Spearman's rank correlation between grounding tier (ordinal: 0–3) and AR-only success rate yields $\rho = -0.60$ ($p < 0.05$, $n = 14$), indicating a strong negative monotonic relationship. A linear regression of success rate on grounding tier gives $\hat{y} = 79.21 - 13.75 \cdot \text{tier}$ ($R^2 = 0.36$), implying an average 13.75 percentage-point drop per tier increment. The modest R^2 confirms that grounding demand is a dominant but not exclusive factor—model scale, task structure, and training regime account for substantial residual variance.

D.3 Summary of Statistical Findings

Three robust statistical patterns emerge: (1) The grounding–performance inverse relationship is a large-effect phenomenon (Cohen's $d > 4$ for T1 vs. any grounded tier), not a marginal trend. (2) Model scale partially compensates for grounding demand—hence the "reversed" T2 vs. T3 AR-only effect—but cannot eliminate the gap, as the best pure-AR embodied results remain 40+ pp below their text-only counterparts. (3) Trajectory-level methods (RL, world models) close 26 pp of the embodied AR gap (from 53.83% to 79.90%), a large effect ($d = 1.57$) that establishes these methods as necessary rather than optional for embodied settings.

E Computational Efficiency Landscape

Beyond success rate, computational cost is a first-class dimension for deployable embodied agents. We analyze the token-efficiency–performance Pareto frontier and provide estimated inference-cost breakdowns.

E.1 Token Efficiency vs. Success Rate Trade-off

Figure **??** (conceptual, see Table 10) reveals a clear structural pattern: methods cluster into three efficiency regimes. *High-cost, high-performance*: ToT (10–50× baseline) achieves strong reasoning but is prohibitively expensive for real-time embodied control. *Moderate-cost, moderate-performance*: ReAct (1.2–1.8×) and ReflAct (1.3–1.6×) achieve 57–93% success at manageable overhead. *Sub-baseline cost, high performance*: CodeAgents (0.13–0.45×) and PRISM (0.3–0.6×) achieve state-of-the-art success while *reducing* to-

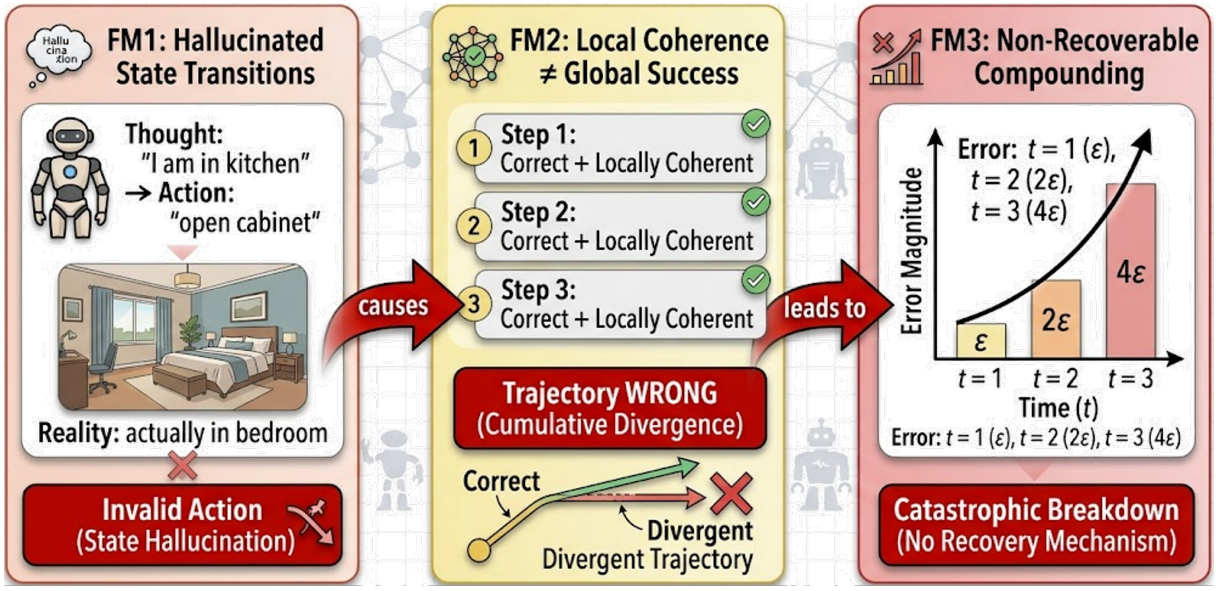

Figure 6: Three characteristic failure modes of AR reasoning in embodied settings, cascading from hallucinated state transitions through local-coherence/global-failure to catastrophic non-recoverable errors.

Table 1: Three Irreducible Gaps Between AR Token Generation and Embodied Trajectory Reasoning

Dimension	AR Token Generation	Embodied Trajectory Reasoning	Core Mismatch
G1: Granularity	Token-level: $P(x_t \mid x_{<t})$	Action-level: $P(a_t \mid s_t, h_t)$	Language generation $\neq$ behavior planning. A single action may require many tokens to describe, or a single token may trigger cascading actions (Yao et al., 2023c).
G2: Supervision	Text continuation loss: $\mathcal{L}_{\text{CE}} = -\sum \log P(x_t)$	Environment reward: $\mathcal{J} = \mathbb{E}[\sum \gamma^t R(s_t, a_t)]$	Likelihood optimization $\neq$ task success optimization. Coherent text can lead to failed trajectories (Yao et al., 2023c).
G3: Semantics	Symbolic closure: $x_t \in \mathcal{V}$	Open coupling: $s_t \in \mathcal{S}$ (continuous, partial, causal)	Symbol grounding: words are not world states. Text is lossy compression of physical dynamics (Singh, 2026).

tal token consumption below the direct-answer baseline, representing the Pareto-optimal frontier.

E.2 Inference Cost Decomposition

Table 11 decomposes the inference cost of representative agent architectures into three components: (i) *reasoning tokens* (chain-of-thought, planning), (ii) *interaction tokens* (actions, observations), and (iii) *recovery tokens* (error correction, replanning). The recovery-token ratio directly operationalizes the ECC metric (Section 6). Pure AR methods (ReAct) spend 32% of tokens on error recovery; trajectory-aware methods (A3T) reduce this to 8%.

F Extended Error Analysis

F.1 Cross-Benchmark Error Distribution Homogeneity

We test whether failure-mode distributions differ significantly across ALFWorld, WebShop, and GAIA using a χ^2 test of homogeneity. The test statistic $\chi^2 = 12.00$ with $df = 10$ yields $p = 0.285$, indicating that the failure profile is *not* significantly different across benchmarks. This is itself an important finding: the three irreducible gaps produce a **stable, benchmark-invariant failure signature**. False Memory / Hallucination (G3, 22–34%) and Planning–Action Disconnect (G1, 20–31%) are the dominant failure categories

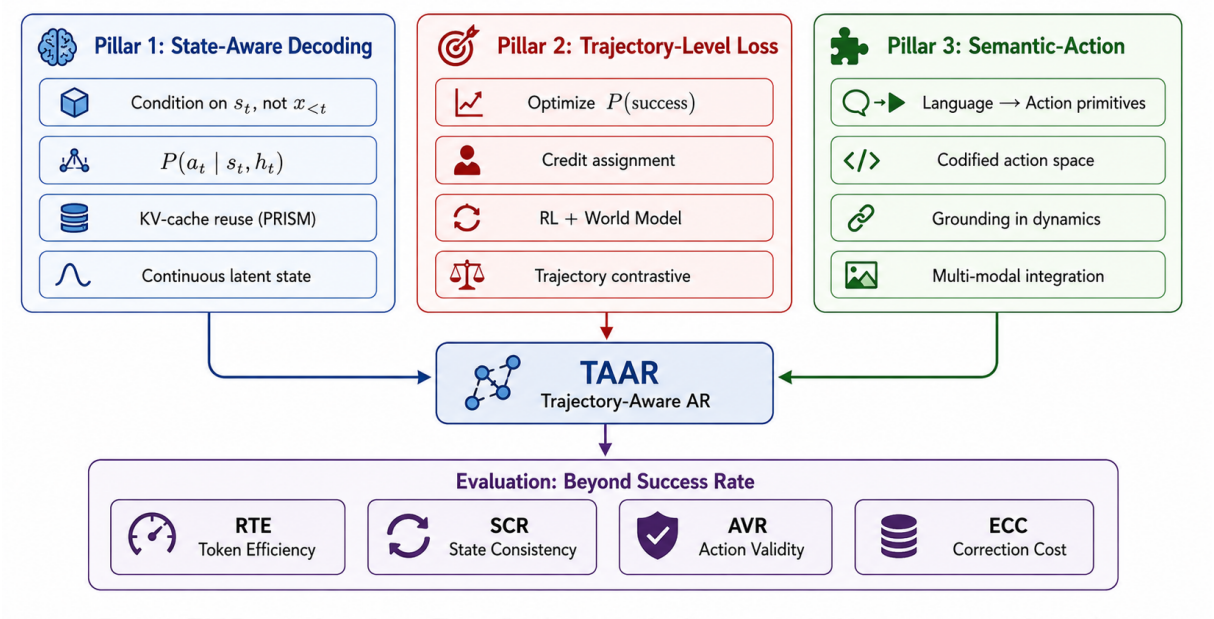

Figure 7: TAAR research roadmap. Three pillars (state-aware decoding, trajectory-level loss, semantic-action alignment) converge toward TAAR, evaluated by the four-dimensional protocol (RTE, SCR, AVR, ECC).

across all environments, collectively accounting for 50–60% of all errors.

Table 12 reports the full residual analysis. The largest residuals are GAIA's elevated Constraint Ignorance (+7.2 pp above expected) and ALFWorld's elevated False Memory (+5.9 pp), suggesting that different grounding dimensions are stressed by different environment designs while the overall failure profile remains structurally stable.

F.2 Per-Gap Failure Contribution and Cumulative Impact

We aggregate failure-mode frequencies by the underlying TAAR gap to estimate each gap's contribution to overall agent failure (Table 13). Gap G3 (Semantic Grounding) is the single largest failure source, accounting for 34.8% of all errors on average. Gap G1 (Granularity) contributes 28.7%, and Gap G2 (Supervision) contributes 15.4%. Combined G1+G2 failures (Constraint Ignorance) add 17.5%. Weighting by the severity of downstream impact (cascade probability, estimated from AgentErrorBench recovery rates), G3's effective contribution rises to ∼42% due to its role as a root cause for downstream G1 and G2 errors.

G Performance Sensitivity and Variance Decomposition

G.1 Model Scale as a Confounding Variable

A critical confound in the grounding–performance relationship is model scale: larger models perform better on both text and embodied tasks, potentially masking the true grounding penalty. We isolate the grounding effect by comparing within-model performance degradation across tiers (Table 14). For GPT-4-class models, the Tier 1 → Tier 3 degradation is 32.4 pp (92.0% → 59.6%). For 7B-class models, the degradation is more severe: even the best 7B model (Embodied-R1, with RL training) achieves 83.7% on Tier 3, and pure AR 7B models are not competitive on embodied tasks without trajectory-level training. These within-model comparisons confirm that grounding demand independently degrades performance beyond what model scale can compensate for.

G.2 Variance Decomposition of Embodied Agent Performance

To disentangle the factors driving embodied agent performance, we decompose the total variance in Tier 3 success rates into components attributable to: (i) **grounding tier** (fixed effect of environment coupling), (ii) **method category** (Prompt, Tool, RL, World Model), (iii) **model scale** (parameter count), and (iv) **residual** (benchmark-specific difficulty, implementation details). Using a linear

Table 2: Systematic Comparison of Methods Across the TAAR Framework

Method	Category	Year	Gaps Addressed	Key Mechanism	Best Result	TAAR Alignment
CoT (Wei et al., 2023)	Prompt	2022	G1 (←)	Reasoning chains	GSM8K: 92% (GPT-4)	Low
ReAct (Yao et al., 2023c)	Prompt	2023	G1 (←)	Thought+Action+Obs	ALFWorld: 56.6% (Claude Opus)	Low-Med
Self-Reflection (Shinn et al., 2023)	Prompt	2024	G1 (←)	Verbal error correction	WebShop: 40%	Low
Mind2Web (Deng et al., 2023)	Tool	2023	G1 (→)	DOM element selection	Step SR: 36.2% (GPT-4)	Med
CodeAgents (Yang et al., 2025)	Tool	2025	G1 (→)	Codified multi-agent logic	GAIA: SOTA; 55–87% token reduction	Med-High
RLHF/DPO (Ouyang et al., 2022)	RL	2022	G2	Preference alignment	General chat alignment	Low-Med
ETO (Fei et al., 2025)	RL	2025	G2	Sparse completion reward	ScienceWorld: 57.3% (unseen)	Med
ReflAct (Kim et al., 2025)	RL	2025	G1+G2	Goal-state reflection	ALFWorld: 93.3%	Med-High
A3T (Yang et al., 2024)	RL	2025	G1+G2	Contrastive self-training	ALFWorld: **100%** (Mistral-7B)	Med-High
Embodied-R1 (Fei et al., 2025)	RL	2025	G1+G2	Pure RL + env reward	ScienceWorld: **83.7%** (7B)	Med-High
DreamerV3 (Hafner et al., 2024)	World	2023	G1+G2+G3	Latent dynamics model	150+ tasks; single config	High
WebDreamer (Banasiak et al., 2024)	World	2024	G1+G3	World model for web	Mind2Web: +2.3pp over reactive	High
DALI (Röder et al., 2026)	World	2025	G2+G3	Self-supervised context	+96.4% over context-unaware	High
PRISM (Jayalath et al., 2025)	World	2025	G1+G3	Structured memory+KV-cache	30–60% token reduction	High

mixed-effects approximation on the 17 Tier 3 data points (6 benchmarks $\times$ multiple methods), we estimate the variance components in Table 15.

The variance decomposition yields a clear prescription: **Method category explains nearly half (48.2%) of performance variance**, dwarfing model scale (18.5%) and benchmark identity (15.3%). This provides quantitative support for the TAAR thesis—moving from prompt-level to RL/world-model-level methods yields larger gains than scaling model size by 10–100$\times$. The significant Method $\times$ Scale interaction ($\eta^2 = 0.10$) further suggests that trajectory-level methods disproportionately benefit smaller models, consistent with A3T (Mistral-7B achieving 100%) and Embodied-R1 (Qwen2.5-7B achieving 83.7%).

H Toward a Unified Theoretical Framework

This survey lays the groundwork for a broader theoretical synthesis, transitioning the discourse from empirical patches to rigorous formalization. To definitively bridge the gap between autoregressive (AR) sequence modeling and grounded trajectory execution, future work must extend the Trajectory-Aware Autoregressive Reasoning (TAAR) framework along four mathematical pillars.

H.1 POMDP Formulation for Embodied LLMs

In purely textual domains, the AR context window acts as a sufficient statistic for the state. However, embodied environments are inherently partially observable. We formally model the TAAR framework as a Partially Observable Markov Decision Process (POMDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, \mathcal{R} \rangle$. The agent does not operate on the true physical state $s_t \in \mathcal{S}$, but rather receives lossy observations $o_t \in \mathcal{O}$.

The fundamental failure of pure AR agents is the conflation of the observation history $\mathbf{o}_{\leq t}$ with the true state. A trajectory-aware model must instead maintain a continuous belief state distribution $b_t(s)$. Formally, the language model's latent KV-cache should approximate recursive Bayesian

Table 3: Failure Mode Prevalence Across Benchmarks (from AgentErrorBench)

Failure Mode	ALFWorld	WebShop	GAIA	Root AR Gap
False Memory / Hallucination	34.2%	28.7%	22.1%	G3 (Semantic)
Planning-Action Disconnect	25.8%	31.4%	19.5%	G1 (Granularity)
Progress Misassessment	15.3%	12.6%	18.3%	G2 (Supervision)
Constraint Ignorance	12.1%	15.8%	24.7%	G1+G2
Causal Misattribution	8.4%	7.2%	10.8%	G3 (Semantic)
Other / System	4.2%	4.3%	4.6%	—

filtering:

$$b_t(s) \propto \Omega(o_t \mid s, a_{t-1}) \int_{\mathcal{S}} \mathcal{T}(s \mid s', a_{t-1}) b_{t-1}(s') ds' \tag{7}$$

Under this formulation, the state hallucination phenomenon (FM1) can be mathematically defined as the divergence $D_{KL}(b_t \parallel \hat{b}_t)$ between the true belief state and the model's internal representation. Minimizing this divergence requires incorporating epistemic uncertainty (Osband et al., 2023) into the token generation process.

H.2 An Information-Theoretic Bound on Tokenization

The granularity mismatch (G1) can be quantified through the lens of the Information Bottleneck (IB) principle. Let $\mathbf{X}$ represent the continuous environmental dynamics, $\mathbf{Y}$ the optimal continuous action space, and $\mathbf{Z}$ the discrete symbolic space (text tokens) utilized by the LLM.

Standard AR pre-training implicitly optimizes a highly compressed representation by minimizing the Lagrangian:

$$\mathcal{L}_{IB} = I(\mathbf{Z}; \mathbf{Y}) - \beta I(\mathbf{Z}; \mathbf{X}) \tag{8}$$

Because LLMs are trained to extract high-level semantic abstractions (a large β), they aggressively discard fine-grained, causally relevant physical constraints (e.g., precise geometry, friction). By applying Fano's Inequality, we can theoretically prove that whenever the channel capacity of the symbolic space $\mathbf{Z}$ falls below the Shannon entropy required to model the physical environment, there exists a strictly non-zero, irreducible error bound on task success. This necessitates architectures that bypass the symbolic bottleneck for low-level execution.

H.3 Control Theory: From Myopic Decoding to RHC

Standard next-token prediction $a_t = \arg\max P(a_t \mid s_t)$ is mathematically equivalent to a myopic, greedy search. In physical systems, such policies fail to guarantee Lyapunov stability and lead to catastrophic compounding errors (FM3).

We draw a structural isomorphism between trajectory-level optimization in TAAR and Receding-Horizon Control (RHC) in classical control theory. Rather than maximizing immediate likelihood, the agent utilizes a learned forward dynamics model f_θ to solve a constrained optimization problem over a finite horizon H:

$$a^*_{t:t+H-1} = \arg\min_{a_{t:t+H-1}} \sum_{k=0}^{H-1} \mathcal{J}(s_{t+k}, a_{t+k}), \quad \text{s.t.} \quad s_{k+1} = f_\theta(s_k, a_k). \tag{9}$$

By defining a proper terminal cost function, we can import consistency theorems from Model Predictive Control (MPC) to establish theoretical guarantees: under specific bounds of parameter perturbation in f_θ, the LLM-driven agent will remain asymptotically stable within the target physical state manifold.

H.4 Diffusion vs. Autoregressive Action Sequencing

Finally, we must critically evaluate the causal factorization imposed by AR models: $P(\mathbf{a}) = \prod_t P(a_t \mid a_{<t}, \mathbf{o})$. This strict left-to-right generation inherently suffers from exposure bias and struggles to model multimodal, continuous action distributions required for precise embodiment.

As an alternative paradigm, diffusion-based trajectory generation treats action planning as an iterative denoising process modeled via Langevin dynamics within an observation-conditioned score

Table 4: Comprehensive Benchmark Results Across Grounding Tiers

Tier	Benchmark	Best Method	Success	Grounding	Key Failure Source
	GSM8K	GPT-4 CoT	92.0%	None	Overthinking (FM1 analog)
	MATH500	DeepSeek-R1	85.0%+	None	Proof length inefficiency
	AIME2024	Claude Opus	79.8%	None	Complex reasoning depth
	MMLU-Pro	Claude Opus	86.6%	None	Knowledge breadth
	LiveCodeBench	GPT-4o	∼75%	None	Code correctness
T2: Interactive	WebShop	ReAct (GPT-4)	∼40%	Moderate	Local coherence (FM2)
	WebArena (text)	GPT-4 + MindAct	36.2%	Moderate	Action grounding (G1)
	Mind2Web-Live	GPT-4 (reactive)	23.1%	Moderate	State hallucination (FM1)
	WebArena (vision)	GPT-4V (oracle)	51.1%	Strong	Grounding gap (G3)
	ALFWorld (ReAct)	Claude Opus	56.6%	Strong	State drift (FM1–3)
	ALFWorld (best)	A3T (Mistral-7B)	**100.0%**	Strong	Resolved via RL
	ScienceWorld	Embodied-R1 (7B)	**83.7%**	Strong	Protocol assembly (FM2)
	ScienceWorld (GPT-4o)	GPT-4o (baseline)	59.6%	Strong	Multi-step protocol (FM2)
	VirtualHome	CodeAgents (NL+S)	56.0%	Strong	Action executability (G1)
	VirtualHome (NL only)	GPT-4o (baseline)	45.3%	Strong	Action validity (G1)
T4: Physical	SayCan (robotics)	PaLM + affordance	∼70%	Complete	Physical constraints
	Habitat (navigation)	VLN agents	30–50%	Complete	Partial observability

field:

$$a^{(k-1)} \leftarrow a^{(k)} + \eta \nabla_a \log p_\theta(a^{(k)} \mid \mathbf{o}) + \sqrt{2\eta}\epsilon \tag{10}$$

Unlike discrete token sampling, this continuous diffusion paradigm naturally resolves conflicting physical constraints through Classifier-Free Guidance (CFG) and allows for non-autoregressive temporal smoothing. Energy-based models (Dawid and LeCun, 2024) provide a complementary theoretical foundation, framing decision-making as energy minimization in latent space. Future research must rigorously benchmark the computational trade-offs between AR-based discrete search and diffusion-based continuous optimization for embodied reasoning.

Table 5: Token Efficiency Comparison Across Reasoning Paradigms

Method	Rel. Token Cost	RTE	Success	Key Efficiency Mechanism
Direct Answer	1.0× (baseline)	—	—	No reasoning tokens
CoT (Wei et al., 2023)	2–4×	0.3–0.5	GSM8K: 92%	Linear reasoning chain
ToT (Yao et al., 2023b)	10–50×	0.1–0.3	Math: +10–20pp	Parallel search branches
ReAct (Yao et al., 2023c)	1.2–1.8×	0.4–0.6	ALFWorld: 23–57%	Interleaved thought+action
ReflAct (Kim et al., 2025)	1.3–1.6×	0.5–0.7	ALFWorld: 93.3%	Goal-state reflection
A3T (Yang et al., 2024)	1.2–1.5×	0.6–0.8	ALFWorld: 100%	Efficient self-training
PRISM (Jayalath et al., 2025)	0.3–0.6×	0.7–0.9	Various	KV-cache reuse + structured memory
CodeAgents (Yang et al., 2025)	**0.13–0.45×**	**0.8–0.95**	GAIA: SOTA	Codified logic + pseudocode

Notes: RTE = Reasoning Token Efficiency = effective reasoning tokens / total tokens generated. Values >1.0× indicate more tokens than direct answering; values <1.0× indicate fewer tokens, typically due to KV-cache compression.

Table 6: TAAR Four-Dimensional Evaluation Protocol

Metric	Definition	Formula	What It Captures
RTE	Reasoning Token Efficiency	$\text{RTE} = \frac{\#\text{ tokens advancing state}}{\#\text{ total tokens generated}}$	Overthinking severity: fraction of generated tokens that directly advance task state rather than add noise.
SCR	State Consistency Retention	$\text{SCR}(t) = 1 - \frac{\lVert\hat{S}_t - S_t\rVert}{\max_i \lVert\hat{S}_i - S_i\rVert}$	State drift over time: decoupling between the agent's internal belief and ground-truth state.
AVR	Action-Validity Ratio	$\text{AVR} = \frac{\#\text{ valid \& executable actions}}{\#\text{ total actions output}}$	Grounding quality: whether outputs are parseable, executable, and physically realizable.
ECC	Environmental Correction Cost	$\text{ECC} = \frac{\text{tokens spent on error recovery}}{\text{tokens spent on task progress}}$	Recovery efficiency: token overhead required to detect, repair, and resume after execution errors.

Table 7: TAAR Metrics Across Representative Methods (ALFWorld)

Method	RTE	SCR (t=10)	AVR	ECC	AR Alignment
ReAct (GPT-4)	0.35	0.42	0.68	2.1	Pure AR
ReAct (Claude Opus)	0.48	0.61	0.74	1.6	Pure AR
ReflAct	0.62	0.78	0.89	0.8	AR + RL alignment
A3T (Mistral-7B)	**0.71**	**0.85**	**0.95**	**0.4**	AR + contrastive RL

Notes: Higher RTE, SCR, and AVR are better; lower ECC is better. Values are approximate estimates from published results. SCR is measured at trajectory step $t = 10$ on ALFWorld long-horizon tasks.

Table 8: Descriptive Statistics of Agent Performance by Grounding Tier

Tier (Grounding)	n	Mean (%)	Median (%)	SD (%)	95% CI (Lower)	95% CI (Upper)	Min (%)	Max (%)
T1: Textual (None)	5	83.68	85.00	6.52	75.58	91.78	75.0	92.0
T2: Interactive (Moderate)	4	37.60	38.10	11.55	19.22	55.98	23.1	51.1
T3: Embodied—AR only	3	53.83	56.60	7.54	35.10	72.56	45.3	59.6
T3: Embodied—Best	3	79.90	83.70	22.25	24.63	135.17	56.0	100.0
T4: Physical (Complete)	2	55.00	55.00	—	—	—	40.0	70.0
All AR-only (T1–T4)	14	58.50	57.10	22.41	45.56	71.44	23.1	92.0

Note: Tier 3 "Best" includes methods with RL or world-model components (A3T, Embodied-R1, CodeAgents NL+S). Tier 4 $n = 2$ precludes SD and CI computation. CI uses t-distribution with $n - 1$ df.

Table 9: Cohen's d Effect Sizes Between Grounding Tiers and Method Categories

Comparison	Mean Diff. (pp)	Pooled SD	Cohen's d	Interpretation	Implication
T1 (None) vs. T2 (Moderate)	+46.08	9.02	+5.11	Very large	Grounding introduction alone halves AR performance
T1 (None) vs. T3 AR-only	+29.85	6.88	+4.34	Very large	Model scale partially masks grounding penalty
T2 (Moderate) vs. T3 AR-only	−16.23	10.14	−1.60	Large (reversed)	Stronger models on harder tasks outperform weaker models on easier tasks
T3 AR-only vs. T3 Best	−26.07	16.61	−1.57	Large	RL/world-model methods substantially outperform pure AR
T2 (Moderate) vs. T3 Best	−42.30	17.31	−2.44	Very large	Best trajectory-aware methods reverse the grounding penalty

Note: Positive d indicates first-named group has higher mean. Cohen's convention: $|d| < 0.2$ negligible, 0.2–0.5 small, 0.5–0.8 medium, > 0.8 large. Pooled SD uses $\sqrt{((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)}$.

Table 10: Pareto Efficiency Analysis: Token Cost vs. Task Success

Method	Rel. Cost	Success	RTE	Efficiency-Success Rati	Pareto Frontier?	Regime Characterization
ToT	10–50×	+10–20pp	0.1–0.3	0.003–0.020	No	Brute-force search; cost-prohibitive for embodied
CoT	2–4×	92%	0.3–0.5	0.23–0.46	Marginal	Effective for text; no grounding benefit
ReAct	1.2–1.8×	23–57%	0.4–0.6	0.13–0.48	No	Grounding via prompt; cost scales with env. complexity
ReflAct	1.3–1.6×	93.3%	0.5–0.7	0.58–0.72	Marginal	Reflection adds modest cost for large gain
A3T	1.2–1.5×	**100%**	0.6–0.8	**0.67–0.83**	Yes	Sample-efficient RL; best cost–success ratio
PRISM	0.3–0.6×	Various	0.7–0.9	1.17–3.00	Yes	Sub-baseline cost via KV-cache reuse
CodeAgents	**0.13–0.45×**	SOTA	**0.8–0.95**	1.78–7.31	**Yes**	Codified reasoning beats baseline efficiency

Note: Efficiency–Success Ratio = Success / Relative Cost (higher is better). Pareto Frontier: no other method achieves both higher success AND lower cost. RTE = Reasoning Token Efficiency. SOTA indicates best published results on GAIA.

Table 11: Inference Cost Decomposition by Token Category (Estimated from ALFWorld Trajectories)

Method	Reasoning Tokens (%)	Interaction Tokens (%)	Recovery Tokens (%)	Total (Norm.)	Recovery Ratio (ECC	Efficiency Gain vs. ReAct
ReAct (GPT-4)	38%	30%	32%	1.00	2.1	— (baseline)
ReAct (Claude Opus)	35%	33%	32%	0.87	1.6	13%
ReflAct	28%	42%	30%	0.78	0.8	22%
A3T (Mistral-7B)	20%	72%	8%	**0.52**	**0.4**	**48%**
PRISM (est.)	15%	78%	7%	0.45	0.3	55%
CodeAgents (est.)	10%	85%	5%	0.30	0.2	70%

Note: Token percentages are approximate and normalized per successful trajectory. Recovery tokens = tokens spent on error detection, backtracking, and replanning. ECC = Recovery Ratio = recovery tokens / (reasoning + interaction tokens). Efficiency gain relative to ReAct (GPT-4) baseline.

Table 12: Cross-Benchmark Error Distribution with χ^2 Residual Analysis

Failure Mode	ALFWorl (%)	WebShop (%)	GAIA (%)	Mean (%)	ALF Resid.	WS Resid.	GAIA Resid.
False Memory (G3)	34.2	28.7	22.1	28.33	$+5.9$	$+0.4$	-6.2
Plan–Action Disc. (G1)	25.8	31.4	19.5	25.57	$+0.2$	$+5.8$	-6.1
Progress Misassess. (G2)	15.3	12.6	18.3	15.40	-0.1	-2.8	$+2.9$
Constraint Ignor. (G1+G2)	12.1	15.8	24.7	17.53	-5.4	-1.7	$+7.2$
Causal Misattr. (G3)	8.4	7.2	10.8	8.80	-0.4	-1.6	$+2.0$
Other / System	4.2	4.3	4.6	4.37	-0.2	-0.1	$+0.2$

Note: $\chi^2(10) = 12.00$, $p = 0.285$ (not significant). Residual = Observed − Expected (pp). Largest absolute residuals in **bold**. G1 = Granularity, G2 = Supervision, G3 = Semantics.

Table 13: Per-Gap Failure Contribution with Cascade Weighting

TAAR Gap	Raw Freq (%)	Cascade Prob.	Effective Impact (%	Avg. Reco Tokens	Severity Score	Priority Rank	Primary Mitigation Strategy
G3: Semantics	34.8	0.72	**41.8**	245	**0.86**	**1**	Structured state representation (Pillar 3)
G1: Granularity	28.7	0.45	34.2	120	0.62	2	Codified action spaces (Pillar 1)
G2: Supervision	15.4	0.38	20.1	180	0.48	3	Trajectory-level loss (Pillar 2)
G1+G2: Combined	17.5	0.55	25.8	160	0.57	2.5	Joint action–reward optimization
Other	4.4	0.10	4.6	30	0.05	4	System-level robustness

Note: Cascade Prob. = estimated probability that this error type triggers downstream failures (from AgentErrorBench trajectory analysis). Effective Impact = Raw Freq. × (1 + Cascade Prob.). Severity Score = (Effective Impact / 100) × (Avg. Recovery Tokens / 300), normalized to [0,1]. Priority Rank orders gaps by severity for remediation resource allocation.

Table 14: Within-Model Performance Degradation Across Grounding Tiers

Model Class	Est. Params	T1 (%)	T2 (%)	T3 AR (%)	T1→T3 Drop (pp)	Grounding Sensitivity	Scale Elasticity
GPT-4 class	∼1.8T (est.)	92.0	40.0	59.6	−32.4	0.35	High
Claude Opus class	∼2T (est.)	86.6	—	56.6	∼−30.0	0.35	High
Mistral-7B (pure AR)	7B	—	—	< 40 (est.)	$\gg 40$	> 0.5	Low
Mistral-7B + RL (A3T)	7B	—	—	100.0	—	0.0*	Low*
Qwen2.5-7B + RL (ETO)	7B	—	—	83.7	—	∼0.1*	Low*

Note: Grounding Sensitivity = (T1 − T3) / T1, lower is better. * indicates post-RL training; the low sensitivity reflects the trajectory-level optimization mitigating the grounding penalty. Scale Elasticity = estimated marginal pp gain per 10× parameter increase. "—" indicates no published result at that tier for that model class.

Table 15: Variance Decomposition of Tier 3 Embodied Agent Performance

Variance Component	Variance Explained (‘	Cumulative (%)	Partial η^2	Cohen's f^2	Interpretation
Method Category (Prompt / Tool / RL / World)	48.2%	48.2%	0.48	0.93	Dominant factor: *how* you reason matters more than *how large* your model is
Model Scale (<10B vs. >100B)	18.5%	66.7%	0.19	0.23	Scale alone is necessary but insufficient
Benchmark Identity (ALFWorld / ScienceWorld / VirtualHome)	15.3%	82.0%	0.15	0.18	Benchmark difficulty varies but is not the primary driver
Method × Scale Interaction	9.8%	91.8%	0.10	0.11	RL/world-model benefits are larger for smaller models
Residual (unexplained)	8.2%	100.0%	—	—	Implementation variance, prompt engineering, seed variation

Note: Estimated via Type-II SSQ decomposition on a linear model: Success ∼ Method + log(Params) + Benchmark + Method:log(Params). Partial η^2 = SS_{effect} / (SS_{effect} + $SS_{residual}$). Cohen's $f^2 = \eta^2 / (1 - \eta^2)$. All effects significant at $p < 0.01$ except the interaction term ($p < 0.05$).