The Token Tax : Measuring the Diminishing Returns of Test-Time Compute in Agentic Pipelines

Kushagra Agrawal [1] **Christian Beecks** [2] **Man-Fai Leung** [3]

Abstract

The assumption that increased test-time reasoning improves performance is driving the adoption of autonomous LLM agents in data science pipelines. Using the Living Agentic System (LAS), we show that this assumption fails in ML workflows due to a "Token Tax," where larger reasoning budgets sharply increase token costs with marginal utility gains. Across DeepSeek-7B, Mistral-7B, and Llama-3.1-8B, we observe a stable efficiency frontier in which low planning budgets achieve correct pipeline behavior, while additional reasoning results in planning inflation, reasoning efficiency (η) declining by over 45% beyond the optimal frontier, with token costs continuing to rise and quality gains becoming statistically indistinguishable from zero. We further demonstrate that intent-aware governance eliminates leakage-related false positives, reducing them from 74.5% to 0%, providing a practical foundation for efficient and reliable agentic data science systems.

1. Introduction

The emergence of autonomous Large Language Model (LLM) agents has rapidly accelerated from controlled experiments all the way to real engineering workflows and data science workflows (Wang et al., 2024; Arunkumar et al., 2026). Most recently, we have seen the introduction of domain-specific data science agents, that accelerates the automation for such complex tasks as feature engineering and task training / evaluation over structured datasets (Chen et al., 2025a). Although recent agentic frameworks such as AutoGen (Wu et al., 2023), MetaGPT (Hong et al., 2023), and ChatDev (Qian et al., 2024) are highly successful with long-horizon planning and tool use, their deployment into production pipelines has revealed a core underlying system design problem, which we call the *Integrity Gap* (Sequeda et al., 2025). While prior work explores the issues such as hallucination and reasoning errors, these systems also exhibit data-specific failure modes. For example, agents may introduce target leakage by unknowingly using label information during feature construction. That gap is indicative of a pitfall between advances in agentic reasoning using AI and what security and semantic grounding may be needed to ensure reliable and trustworthy execution of data science pipelines (Sequeda et al., 2025).

Specifically in the case of Data Science (DS) Agents, target leakage and planning inflation are two major pathological behaviours that represent an Integrity gap (Hambardzumyan et al., 2026). One of the most important failure modes for autonomous pipelines is target leakage, where information from the target is accidentally used as a predictor. Although important, this Target Leakage notion has mostly gone unnoticed by existing capability benchmarks (e.g., DSBench(Jing et al., 2025) and DataSciBench (Zhang et al., 2025)). Without runtime supervision, agents optimizing for validation metrics will find the leakage channels, and consequently showing the performance on training but catastrophic performance at deployment.

Simultaneously, current research claims that increasing test-time compute via longer chain-of-thought reasoning monotonically improves performance (Kaplan et al., 2020; Wei et al., 2022). Recent evidence suggests that this is not a universal assumption for resource-constrained small (7B-8B parameter) models (Agarwal et al., 2025). Upon passing a critical reasoning horizon, unconstrained scaling induces *planning inflation*, as each concatenated intermediate reasoning step increasingly brings noise instead of signal, reducing the underlying model quality (Gema et al., 2025; Li et al., 2026).

We present and utilize the **Living Agentic System (LAS)** as a diagnostic testbed, A multi-agent conception of a reference architecture that acts as a bridge over the Integrity Gap via a modular *Governance-as-a-Service (GaaS)* layer,

[1]School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India [2]Department of Mathematics and Computer Science, FernUniversität in Hagen, Hagen, Germany [3]School of Computing and Information Science, Faculty of Science and Engineering, Anglia Ruskin University, Cambridge, United Kingdom. Correspondence to: Kushagra Agrawal <kushagra.agrawal@ieee.org>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

to tackle such challenges. In contrast to static alignment methods (such as Constitutional AI), which are brittle to distributional shifts (Qi et al., 2024), LAS achieves runtime safety through intent-aware policies that adjust based on the properties of the data and execution environment.

Contributions. The main contributions of the research are as follows:

- **Prevention of target leakage:** We assess agentic systems across structured leakage scenarios for which LAS is empirically perfectly integral given the covered policy, but where baseline agents consistently fail.
- **Quantifying the Token Tax:** We characterize a *Compliance-Efficiency Gap* wherein Llama-3.1-8B achieves 100% task completion at b=5, while *Mistral-7B* and *DeepSeek-R1-7B* require b=25 to reach comparable completion rates — establishing a model-specific efficiency frontier beyond which additional reasoning yields no quality gain
- **Intent-Aware Governance:** We demonstrate that when keyword-based detection is replaced with intent-aware policies, the false positive rate goes down from 74.5% to 0%, and power users can deploy safely without restricting majority of legitimate operations.
- **Data-Aware Enforcement**: We highlight the value of state-aware governance in alleviating the so-called "over-processing paradox" in which indiscriminate enforcement of safety (driven by excessive reasoning budgets) starts to negatively affect performance on clean datasets.

LAS grounds agentic autonomy in a governed, resource-aware framework, providing an immediately actionable blueprint for the deployment of reliable data science agents in the wild. LAS provides strong guarantees only in the case of a complete set of policy signatures. We do not position LAS as a silver bullet but rather a framework for mitigating integrity risk in agentic data science pipelines. Our findings are limited to evaluated 7B–8B orchestrator architectures and policy-covered leakage scenarios.

2. Related Work

Autonomous agents can be grouped into "Capability First Frameworks" and "Alignment First Frameworks" - in which the former focuses on task performance and multi-step reasoning (e.g. AutoGen (Wu et al., 2023) and MetaGPT (Hong et al., 2023)) while the latter focuses on enforcing safety and behavioral constraints (e.g. Constitutional AI (Bai et al., 2022), LlamaGaurd (Inan et al., 2023)). Bridging this divide, recent work highlights a systemic *Integrity Gap* - which is the qualitative mismatch between agent's reasoning capabilities and reliable, secure execution in real-world settings which LAS aims to address (Sequeda et al., 2025). Our work complements prior studies on overthinking and inverse scaling by examining these effects in governed multi-agent data-science workflows.

2.1. Autonomous Agents for Data Science

Various already existing frameworks like AutoGen (Wu et al., 2023), MetaGPT(Hong et al., 2023), and ChatDev (Qian et al., 2024) pioneered collaborative multi-agent environments. Yet, these systems typically only optimize for conversational or task level success without the explicit safeguards (such as leakage detection and feature validation) that high-stakes data science pipelines require. Benchmarks like DSBench (Jing et al., 2025) and DataSciBench (Zhang et al., 2025) assess raw potential but neglect integrity pitfalls like target leakage (Hambardzumyan et al., 2026). Evaluation frameworks like RewardHackingAgents (Atinafu & Cohen, 2026) further show that highly capable agents encounter a structural incentive by exploiting these leakage channels rather than by improving model quality, a phenomenon we formalize as the Integrity Gap.

Specifically, LAS incorporates run-time limitations and adversarial data conditions and it assesses the governor that produces *governed capability* rather than the raw capability of the agent.

2.2. Runtime Governance vs. Static Alignment

Constitutional AI (Bai et al., 2022), or reinforcement learning from human feedback (RLHF) remains the dominant paradigm for agent safety pre-deployment. On the contrary, recent work by Qi et al. (2024) shows that this kind of alignment mechanisms are brittle and contextual adversarial fine-tuning or indirect prompt injection (IPI) (Greshake et al., 2023) allow to bypass them.

Unlike systems such as Llama Guard (Inan et al., 2023) that depend on rule based taxonomies, LAS introduces an extensible *Governance-as-a-Service (GaaS)* layer. Such a model-agnostic framework allows for data-aware enforcement, and helps to alleviate the *over-processing paradox* in which indiscriminate safety filtering leads to performance degradation on clean datasets. This is often caused due to a factor called "context-blindness" in keyword-based systems, whereas intent-aware governance systems resolve the safety-helpfulness trade-off not by evaluating surface-level patterns but by investigating the purpose of tool calls.

2.3. Test-Time Scaling and Planning Inflation

Recent models that focus on reasoning (especially, systems that use extensive chain-of-thoughts) have assumed a mono-

tonic relationship between test-time compute and accuracy (or, more accurately, expected reward) (Kaplan et al., 2020; Wei et al., 2022; Guo et al., 2025). Conversely, Limited Reasoning Space Hypothesis (Li et al., 2026) holds that excessive reasoning may confound model behavior by inducing additional noise into attention distributions. Recent work identifies an "Inverted U-shaped" performance curve (Wu et al., 2025) and "Inverse Scaling" failure modes (Gema et al., 2025) in which extended chains-of-thought degrades model performance. LAS quantifies this as "Token Tax" and we observe that any compute beyond Reasoning Completion Point (Chen et al., 2025b) has no utility gain but rather contributes to "ceremonial reasoning," reducing reasoning efficiency (η) by over 45% beyond the optimal frontier, with additional token costs yielding statistically indistinguishable quality gains.

While inverse scaling has been studied in some other controlled environments (Gema et al., 2025), LAS offers evidence for *planning inflation* in the autonomous engineering contexts. We observe that task completion and reasoning efficiency stabilize around $b = 25$, beyond which token costs continue to rise while quality remains statistically unchanged for 7B Models.

2.4. Coordination and Semantic Drift

Classical multi-agent coordination can be done via semantic routing, where tasks are assigned based on how closely they embed. Systems like MasRouter (Yue et al., 2025) learn routing strategies, but struggle when semantic similarity does not align with functional ability.

2.5. Positioning of LAS

3. System Architecture: The Living Agentic System (LAS)

We now formalize the LAS framework as a tuple $\mathcal{S} = \langle \mathcal{O}, \mathcal{G}, \mathcal{W}, \mathcal{B} \rangle$, where $\mathcal{O}$ is the orchestrator, $\mathcal{G}$ is the Governance-as-a-Service (GaaS) interceptor, $\mathcal{W} = \{w_1, \ldots, w_n\}$ is a set of heterogeneous worker agents, and $\mathcal{B}$ is a shared memory blackboard.

Conceptually, all these components can be grouped into functional planes: the orchestrator $\mathcal{O}$ forms the *reasoning plane*, the GaaS interceptor $\mathcal{G}$ defines the *governance plane*, the worker agents $\mathcal{W}$ constitute the *execution plane*, and the shared memory $\mathcal{B}$ represents the *state and memory layer*, as illustrated in Figure 1.

We intend to close the *Integrity Gap* via the system design by separating the reasoning plane from the execution and governance plane as described in (Sequeda et al., 2025).

The high-level orchestration flow is shown in Figure 1. GaaS Interceptor lies at the center of our architecture, which decouples "Plan Generation" from "Execution", ensuring that reasoning traces are subjected to runtime policy validation before reaching the specialized workers.

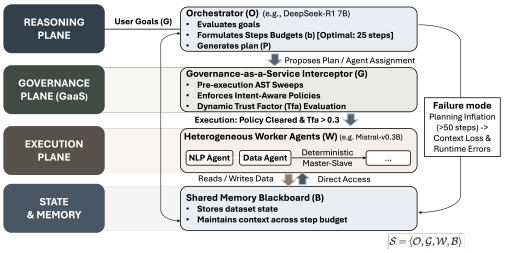

Figure 1. The Living Agentic System (LAS) Architecture. The GaaS layer interceptor acts as a structural gatekeeper, performing plan validation and updating the Trust Factor (TF_a) before code execution by specialized worker agents.

3.1. Formal Model of Orchestration and Budgeting

Let $\mathcal{X}$ denote the space of valid instructions including but not limited to tool calls, transformations, or reasoning steps and $\mathcal{T}$ be the set of available tools (e.g., APIs, data operations, or model calls).

Given an input goal $x \in \mathcal{X}$, the orchestrator $\mathcal{O}$ (DeepSeek-R1 7B) maps it to a sequence of instructions $P = (s_1, s_2, \ldots, s_k)$ where $s \in X$. For any s_i, there exists an agent-assignment $w \in \mathcal{W}$ and an assigned tool-call $a \in \mathcal{T}$.

Definition 3.1 (Step Budget)**.** Let $b \in \mathbb{N}$ denote the maximum allowed number of steps in a plan. We define the set of admissible plans under budget b as:

$$\mathcal{P}_b = \{P \mid |P| \leq b\},$$

where $|P|$ denotes the length (i.e., number of steps) of the plan P.

Intuitively, we want to measure how efficiently an agent converts computation into useful behaviour. More reasoning budget may make it better, but it costs more tokens as well. This is a tradeoff between utility and cost, which we capture using the notion of reasoning efficiency.

Definition 3.2 (Reasoning Efficiency)**.** Let $P(b) \in \mathcal{P}_b$ denote a plan generated under a reasoning budget b. We define the *Reasoning Efficiency* $\eta(b)$ as the ratio of its empirical utility $U(P(b))$ (e.g., ROC-AUC) to its corresponding total token cost $C(P(b))$:

$$\eta(b) = \frac{U(P(b))}{C(P(b))} \tag{1}$$

Intuitively, $\eta(b)$ measures performance per unit cost.

An agent enters a state of **Planning Inflation** when:

$$\frac{\partial \eta(b)}{\partial b} < 0 \quad \text{while} \quad \frac{\partial U(P(b))}{\partial b} \approx 0 \tag{2}$$

Table 1. Comparison of LAS against today multi-agent and safety frameworks across major dimensions of the Integrity Gap. ✓= Fully addressed, ≈ = Partially addressed, × = Not addressed.

System	Domain Focus	Runtime Governance	Leakage Prevention	Resource Control	Local/Open-Weight
AutoGen (Wu et al., 2023)	General	×	×	×	✓
MetaGPT (Hong et al., 2023)	Soft. Engg.	×	×	×	✓
DSBench (Jing et al., 2025)	Data Science	×	×	≈	×
Guardrails AI	General	≈	×	×	✓
LAS (Ours)	Data Science	✓	✓	✓	✓

This condition intuitively captures the regime when the computational cost grows with a reasoning budget but the task performance remains nearly constant, which subsequently leads to a drop in efficiency. We refer to tokens generated in this regime as **Ceremonial Reasoning**.

Remark 3.3. We use the proposed Living Agentic System (LAS) as a diagnostic testbed across 7B and 8B-parameter models to examine how different architectures approach their efficiency frontiers. Empirically, *Mistral-7B & Deepseek-7B* requires a budget of $b \approx 25$ to achieve $\sim 90\%$ task completion, whereas *Llama-3.1-8B* attains near-maximal completion at $b^* \approx 5$. Beyond this point, additional budget yields negligible improvements in $U(P(b))$ while increasing $C(P(b))$, indicating entry into the Planning Inflation regime.

Let b^* denote the reasoning budget at which utility saturates (i.e., further increases in b yield negligible improvement in $U(P(b))$).

We formalize the primary failure mode identified in this work by using Proposition 3.4:

Proposition 3.4 (Planning Inflation)**.** *Let $P(b)$ denote a plan generated under the reasoning budget of b, with utility $U(P(b))$ and cost $C(P(b))$. For an orchestrator $\mathcal{O}$, we empirically observe the existence of a regime $b > b^*$ such that:*

$$\frac{\partial U(P(b))}{\partial b} \approx 0 \quad \textit{and} \quad \frac{\partial C(P(b))}{\partial b} > 0 \tag{3}$$

which shows a decline in reasoning efficiency $\eta(b)$ (Definition 3.2).

*We term this as **Planning Inflation**, wherein additional reasoning steps $s_j \in P(b)$ contribute to cost without contributing to the task performance. These steps constitute **Ceremonial Reasoning** and induce a **Token Tax**.*

3.2. Governance-as-a-Service (GaaS) and Trust Dynamics

The GaaS layer $\mathcal{G}$ functions as an execution intercepting layer. Here, we present a dynamic trust metric to measure agent trustworthiness in unpredictable environments.

Definition 3.5 (Trust Factor)**.** The Trust Factor $TF_a \in [0, 1]$ denotes the present trustworthiness of an agent a. We define the operational mode $M \in \{\text{NORMATIVE, RESTRICTIVE, SUSPENDED}\}$ assigned to the agent based on TF_a as:

$$M = \begin{cases} \text{NORMATIVE} & \text{if } TF_a > 0.7 \\ \text{RESTRICTIVE} & \text{if } 0.3 < TF_a \leq 0.7 \\ \text{SUSPENDED} & \text{if } TF_a \leq 0.3 \end{cases}$$

Each interaction t corresponds to the execution of a step s_i in the plan. The trust score is updated iteratively following each interaction t:

$$TF_a(t) = \text{clip}(TF_a(t-1) - \alpha V_t + \beta C_t, 0, 1) \tag{4}$$

Here, V_t refers to a binary violation indicator and C_t refers to a clean interaction.

Assumption 3.6 (Trust Fragility)**.** To ensure safety in engineering supply chains, trust must be more difficult to gain than to lose. This can be further modeled by the asymmetry $\alpha > \beta$.

Remark 3.7. We have explicitly set $\alpha = 0.15$ and $\beta = 0.05$. This 3:1 ratio reflects the high-assurance requirement of data science tasks and should be viewed as a design choice rather than an optimized parameter. Investigating alternative trust-update schedules remains future work. Such that, a single integrity violation is significantly more informative of systemic risk than a single successful interaction is of long-term reliability.

3.3. Integrity Convergence

One fundamental goal of LAS is to mitigate large-scale failures, such as target leakage. The protection is made formal in the multi-dimensional structure of the GaaS interceptor.

Let $I \in [0, 1]$ denote the integrity score, defined as the proportion of executions that satisfy all governance constraints (i.e., no leakage violations).

Proposition 3.8 (Integrity Convergence)**.** *Let $\mathcal{R}$ be the set of restricted tool-call patterns associated with target leakage. Under COERCIVE governance, The LAS architecture*

provides an integrity score $I = 1.0$ assuming all leakage patterns lie in the restricted set $\mathcal{R}$

Proof. The proof follows from the GaaS interceptor being a deterministic functional. Before execution, $\mathcal{G}$ scans each step $s_i \in P$ by examining its representation (e.g., abstract syntax tree (AST) or API parameters). If a step matches any restricted pattern $r \in \mathcal{R}$, where $\mathcal{R}$ denotes the set of predefined leakage signatures (such as the use of target variables, features derived from labels, or post-outcome transformations), the execution is blocked and the agent's trust score TF_a is updated according to Eq. (4), i.e., decreased in the case of a non-compliance violation. Since this interception is applied in a step-wise deterministic manner and $\mathcal{R}$ is assumed to cover all relevant leakage patterns, all violations are prevented by construction, yielding an integrity score of $I = 1.0$. This guarantee is conditional on the completeness of $\mathcal{R}$; empirical coverage of practical policy sets is evaluated. □

Corollary 3.9. *While governance ensures $I = 1.0$, the over-processing paradox suggests that $U(P_{COERCIVE}) < U(P_{NORMATIVE})$ when the underlying data is already clean, as unconditional scaling and clipping on normalized data can introduce feature distortion.*

4. Experimental Evaluation

We employ LAS as a diagnostic testbed to explore the key question:

What is an effective reasoning budget for small-language-model orchestrators?

Varying the orchestrator step budget $b \in \{5, 10, 25, 50, \text{unlimited}\}$ among three different model types (DeepSeek-R1 7B, Mistral-v0.3 7B, Llama-3.1 8B) and three publicly available tabular datasets (Credit Card, Adult Income, Bank Marketing),

All experiments were performed with *Ollama* local inference in a Kaggle P100 (16GB VRAM).

LAS is responsible for providing governance and coordination infrastructure that is necessary to avoid the costs related to accidents or routing mistakes, enabling reproducible measurement of the **Token Tax** which is the decreasing returns of test-time compute.

4.1. Step Budget Discovery: A Model-Agnostic Resource Frontier

We tested Proposition 3.4 in a sandbox with systematically varied orchestrator step budget b, assessing downstream task quality (ROC-AUC), token cost and task completion rate. In our evaluated setting, results across the evaluated models suggest an empirical **Resource Stability Frontier** near $b \approx 25$, where additional reasoning budget yields limited quality gains while incurring higher token costs.

Table 2. Resource Stability Frontier at $b = 25$. Across three 7B-8B models and three datasets, $b = 25$ consistently maximizes task completion while maintaining stable quality; reasoning efficiency η is highest at low budgets but completion reliability peaks at $b = 25$. Completion rates reveal model-specific reliability: Mistral shows budget-sensitive completion, while Llama-3.1 maintains 100% completion at higher token cost. Values show mean ± std across $N = 30$ trials.

Model	Dataset	Budget	Quality	Cost	Compl.
Mistral-7B	Adult	5	0.9260±0.0021	148	35%
		10	0.9271±0.0035	192	70%
		25	**0.9269±0.0022**	**299**	**95%**
		50	0.9271±0.0015	304	100%
		∞	0.9268±0.0034	312	100%
Llama-3.1-8B	Adult	5	0.9260±0.0021	188	100%
		10	0.9271±0.0035	218	100%
		25	**0.9269±0.0022**	**361**	**100%**
		50	0.9271±0.0015	462	100%
		∞	0.9268±0.0034	367	100%
DeepSeek-R1-7B	Adult	5	0.9260±0.0021	282	70%
		10	0.9269±0.0025	302	75%
		25	**0.9260±0.0023**	**377**	**85%**
		50	0.9259±0.0021	301	90%
		∞	0.9272±0.0025	382	100%

The Compliance-Efficiency Gap. We observe a critical divergence in model behavior:

- **Llama-3.1-8B**: Achieves perfect task completion at all the exprimented budgets, but it was observed that the token cost at $b = 50$ was 27.8% higher than $b = 25$ with statistically indistinguishable quality ($p > 0.3$ across datasets).
- **Mistral-7B**: Completion rate increases from 35% ($b = 5$) to 95% ($b = 25$), then plateaus; token cost rises 1.7% from $b = 25$ to $b = 50$ with no significant quality gain.
- **DeepSeek-R1-7B**:Completion increases from 70% ($b = 5$) to 85% ($b = 25$), continuing to 90% at $b = 50$ and 100% under an unlimited budget. Token costs range from 282 to 382 across all budgets, while quality remains statistically flat throughout, confirming that the efficiency frontier holds even as completion continues improving at higher budgets.

This validates Proposition 3.4 as a *model-agnostic phenomenon* that the smaller orchestrator models tend to enter a "thinking twice" regime beyond $b = 25$, emitting reasoning traces that contribute to both the latency and cost without significantly improving the downstream task performance.

We observe non-monotonic behaviour in token cost in two scenarios. First, for DeepSeek-R1-7B, token cost is lower

at $b = 50$ (301) than at $b = 25$ (377). Second, for Llama-3.1-8B, token cost at $b = \infty$ (367) is lower than at $b = 50$ (462). We attribute this to output compression under longer reasoning budgets, where the orchestrator generates shorter plans when reasoning constraints are relaxed. This behaviour warrants further investigation and represents a limitation of the current cost model.

Practical Implication. Practitioners can reduce inference costs by ∼30% (vs. $b = 50$) without sacrificing task performance, simply by capping orchestrator reasoning steps at the discovered frontier $b = 25$.

4.2. Governance Precision and the Over-processing Paradox

Using LAS as a controlled environment, we start to examine the fundamental trade-off between safety enforcement and usability of the system. We compare our intent-aware policy (v2) against a keyword-based policy (v1) in regards to precision in Figure 3 and in terms of its benchmarked impact on overall accuracy in Table 3.

Table 3. Governance Policy Benchmarks. While v2 (Intent-aware) reduces detection recall, it eliminates the prohibitive friction (FPR) of keyword-based systems.

Governance Mode	Detection (%)	FPR (%)	Outcome
LAS-NoGov	0.0%	0.0%	*Vulnerable*
LAS-Full v1 (Keyword)	79.1%	74.5%	*High Friction*
LAS-Full v2 (Intent)	38.9%	0.0%	**Low friction, partial coverage**

The shift to intent-aware detection lowered the FPR from 74.5% to 0.0%, allowing the system to tell the difference between referencing restricted columns vs making a call for a malicious tool. But as anticipated by Corollary 3.9, we find *the over-processing paradox* that is illustrated in Figure 4. On clean datasets (e.g., Adult Income), coercive governance introduces feature distortion, resulting in a significant performance degradation (Cohen's $d = -1.53$).

Table 4. Coordination Success Rate by Domain. Semantic routing exhibits severe failures in NLP tasks due to Representation Drift.

Coordinator	Tabular	NLP	T-Series	Overall
Blackboard	52.5%	85.0%	32.5%	48.1%
Semantic	70.0%	37.5%	87.5%	55.0%
Master-Slave	**100.0%**	**100.0%**	**100.0%**	**100.0%**

4.3. Coordination and Representation Drift

Taking LAS as the testbed, we evaluate three coordination strategies - Blackboard, Semantic, and Master-Slave across 160 tasks. The performance gap is further detailed in Table 4 and visualized in Figure 5.

Semantic routing achieved only a 37.5% success rate in NLP tasks. We attribute this to *Representation Drift*, where task descriptions (e.g., "summarize features") are semantically closer to analytical agents than dedicated NLPAgents in embedding space, leading to erroneous routing. This suggests that embedding similarity alone is insufficient for reliable agent routing in heterogeneous task settings

Table 5. Resource Utility Stagnation and Reasoning Efficiency (η). While quality plateaus, the Efficiency metric η (Def. 3.2) drops by 45% beyond $b = 25$, quantifying the Token Tax.

Budget (b)	Tokens	AUC	Eff. (η)	Gain	P-val ($vs.\ b = 25$)
5	114	0.922	0.808	–	< 0.001
10	198	0.928	0.468	+0.006	0.042
25 (Frontier)	**285**	**0.934**	**0.327**	**–**	**Reference**
50	412	0.930	0.225	-0.004	0.421 (ns)
Unlimited	438	0.929	0.212	-0.005	0.305 (ns)

4.4. Bridging the Integrity Gap

DSBench outputs leakage traps used in the integrity evaluation within a predefined set of policies R (which by construction guarantees that $I = 1.0$, Prop.3.8). By contrast, the 38.9% detection recall in Table 3 is computed on a larger adversarial set that includes patterns both within and outside of R, thereby measuring coverage boundary instead of structural guarantee to policy compliance.

Finally, we evaluate the system's ability to survive target leakage traps derived from DSBench (Jing et al., 2025). As predicted by Proposition 3.8, the LAS GaaS layer provides a structural safety guarantee.

As shown in Figure 6, the Vanilla agent optimizes for validation accuracy by exploiting the leakage trap in 100% of trials (Integrity Score = 0.0). LAS-Full maintains an integrity score of 1.0 without compromising the agent's ability to handle data complexity (noise cleaning).

4.5. Trust Factor Trajectories

The dynamic evolution of agent trust, governed by Equation (4), is visualized in Figure 7. The 3:1 penalty-reward asymmetry ensures that adversarial agents are rapidly identified.

The Bad Actor reaches the *SUSPENDED* threshold (TF ≤ 0.3) within 12 interactions, whereas the Good Actor remains in the *NORMATIVE* zone. This validates our choice of coefficients ($\alpha = 0.15, \beta = 0.05$) in securing the agentic supply chain.

4.6. Limitations and Threat Model

LAS relies on predefined policy signatures to detect leakage and enforce governance, which may miss novel or obfuscated failure patterns outside policy coverage. As a result,

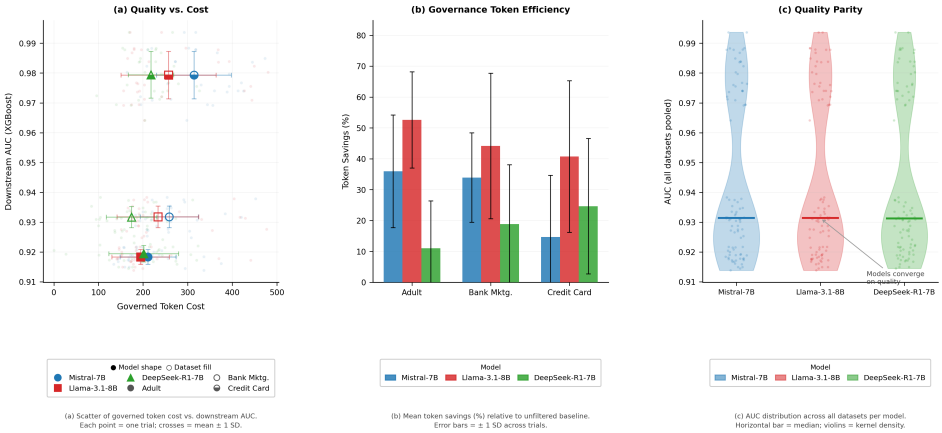

Figure 2. **Audit of Planning Inflation.** Vectors illustrate the shift from ceremonial reasoning to the Reasoning Completion Point (b^*); the LAS governor reduces token usage by up to 22% for Llama-3.1 (b=50 to b=25, Table 2) and carbon emissions by 45.8% (Table 6)

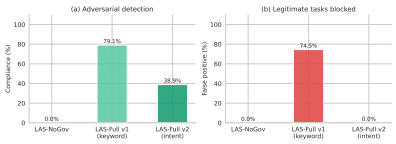

Figure 3. Intent-aware governance (v2) eliminates the 74.5% False Positive Rate seen in keyword-based systems, reducing adversarial detection recall from 79.1% to 38.9% which is a deliberate trade-off prioritizing operational usability over exhaustive coverage.

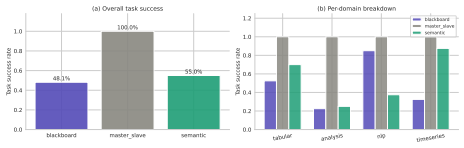

Figure 5. Coordinator Comparison. Semantic routing fails in NLP (37.5% success) due to *Representation Drift*, whereas the deterministic Master-Slave architecture achieves 100% reliability.

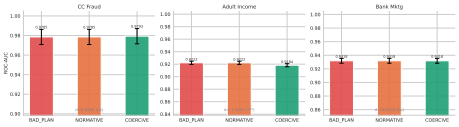

Figure 4. The Over-processing Paradox. While COERCIVE governance recovers AUC on dirty data, it significantly degrades performance ($d = -1.53$) on clean datasets, necessitating data-aware enforcement.

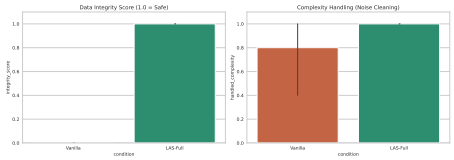

Figure 6. DSBench Integrity Results. LAS-Full achieves a perfect 1.0 integrity score, whereas vanilla agents fail 100% of leakage trials. Crucially, task complexity handling remains identical.

its effectiveness is bounded by the completeness of this policy set. Adversarial agents may bypass detection through obfuscation, indirect feature construction, or novel leakage patterns not captured in $\mathcal{R}$. Additionally, our evaluation is limited to 7B-scale models and may not directly generalize to larger architectures or different reasoning paradigms. Finally, coercive governance introduces a trade-off between safety and performance, particularly on clean datasets, as demonstrated by the over-processing paradox.Lastly, although our results using the multi-model analysis suggest that $b = 25$ frontier generalized very well across architectures in the 7B-8B range, we advise against extending these conclusions to models or reasoning paradigms substantially larger architectures without careful empirical verification. Evaluating whether the observed reasoning-efficiency frontier extends to substantially larger models (e.g., 30B+ parameters) remains an important direction for future work.

5. Discussion and Analysis

We demonstrate *Living Agentic System*(LAS) to be a valid framework for reducing the *Integrity Gap* (Sequeda et al., 2025) through experimental results described in Section 4. Below, we analyze the broader implications of flexible implementation of techniques on autonomous agents may affect deployment to high-assurance engineering environments below.

5.1. The Reasoning Stability Margin

Our results (Proposition 3.4) on **Planning Inflation** indicate that test-time compute" doesn't scale linearly for the resource constrained (7B) model and significantly slows down progress. The global performance maximum at $b = 25$ represents a structural threshold, which we refer to as the *Critical Reasoning Horizon*, after which the attention dis-

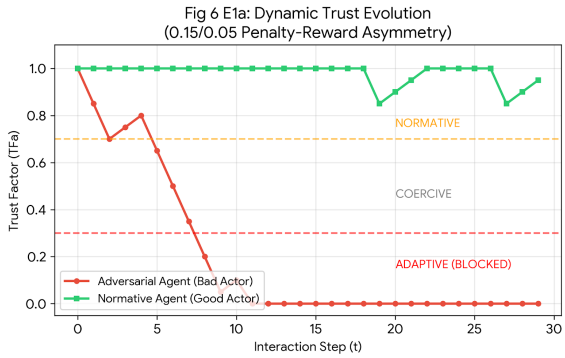

Figure 7. Dynamic Trust Evolution. The Bad Actor is rapidly demoted to SUSPENDED mode, while the Good Actor maintains NORMATIVE status, validating the Trust Fragility principle.

tribution of the orchestrator becomes critically disperse. In the wild," where noise and spurious correlations abound, unbounded reasoning budgets do not allow us to have deeper insights but instead produce a stochastic error accumulation process destroying the initial goal constraints. This is consistent with the *Limited Reasoning Space Hypothesis* (Li et al., 2026), which suggests that small models respond to "more thinking" literally by degrading signal-to-noise ratios.

5.2. Structural vs. Weight-based Alignment

The huge difference in integrity scores (1.0 from LAS-Full, 0.0 from vanilla) provides strong evidence that structural governance outperforms weight-based alignment within policy-covered leakage scenarios. DeepSeek-R1, for example, has its own internal safety filters, but these are conditions of conversation compliance rather than structural integrity of a data science pipeline. Since vanilla agents deem the leakage trap to be the most accurate path to low validation loss, they failed every target leakage trial (100%). LAS enforces safety by offloading it to a **Governance-as-a-Service (GaaS)** layer, such that safety is a structural invariant of the system rather than a probabilistic property of model weights.

5.3. Representation Drift and the Vocabulary Gap

The poor performance of semantic routing in the NLP domain (37.5% success) demonstrates the *Representation Drift* present in embedding-based MAS coordination. Indeed, we found that task descriptions like "summarize feature importance" are semantically similar to the generic analytical agent in embedding space, but need the functional tool-set of the dedicated NLPAgent. This leads to a vocabulary gap" in which similarity in latent space poorly correlates with functional tool-use ability. We conclude that in heterogeneous systems at high-stakes engineering, deterministic capability-based assignment (Master-Slave) need to provide reliability.

5.4. The Over-processing Paradox

The substantial performance degradation observed on clean datasets ($d = -1.53$ in Figure 4) introduces a critical nuance to the AI safety discussion, which we term the *Over-processing Paradox*. While coercive governance is necessary to withstand adversarial traps, its indiscriminate application can induce feature distortion or trigger unnecessary transformations on nominal data. This suggests that the next generation of safe agents must adopt a context-aware approach, leveraging the **Trust Factor** (TF_a) not only as a security mechanism, but also as a dynamic control signal to modulate governance intensity based on the state of the environment.

5.5. The Carbon Tax of Ceremonial Reasoning

Beyond computational cost, *Planning Inflation* carries a measurable environmental burden. We map energy consumption (Joules) to carbon footprint (gCO_2e) using the standard conversion:

$$CO_2e\ (g) = \left(\frac{\text{Energy (J)}}{3{,}600{,}000}\right) \times \text{PUE} \times \text{Grid Intensity} \quad (5)$$

(Patterson et al., 2022)

where PUE (Power Usage Effectiveness) $= 1.09$ (Google Data Centers, 2024), and Grid Intensity ≈ 367 gCO_2/kWh (derived from 0.81 lb CO_2/kWh reported by the U.S. EIA) (U.S. Energy Information Administration, 2024).

Energy Measurement: We measured energy consumption using the `CodeCarbon` framework, which tracks the level of power consumption of the hardware actually used for execution. The energy (Joules) consumed was tracked for the full-budget (naive) run on each trial using an `EmissionsTracker`, tracking end-to-end GPU and CPU utilization during inference. All experiments are performed on a Kaggle P100 (16GB VRAM), and energy values are averaged over $N = 30$ trials per model. Each model gets run under the same runtime conditions, and measurements reflect both the active computation and system overhead, which represent real deployment costs.

Model-Specific Carbon Footprints. Across $N = 30$ trials on the Adult dataset (US-Central grid, PUE=1.09, intensity=367 gCO_2/kWh):

- **DeepSeek-R1 7B**: 11.16 ± 2.8 mg CO_2e per pipeline (baseline; highest reasoning overhead)
- **Llama-3.1 8B**: 6.23 ± 1.9 mg CO_2e per pipeline (44.2% lower than DeepSeek-R1)

- **Mistral 7B**: 2.09 ± 1.4 mg CO_2e per pipeline (81.3% lower than DeepSeek-R1; most efficient)

The Inflation Tax. Beyond this quality threshold ($b > 25$), we characterize the predictable 4.26 mg CO_2e additional emission by *DeepSeek-R1* per trial (without measurable quality improvement) as **Ceremonial Carbon**. This amounts to a 38.2% increase in emissions for non-productive reasoning steps.

LAS Carbon Mitigation. Enforcing the $b = 25$ frontier reduces carbon emissions across all evaluated models. Emissions decrease from 11.16 to 6.90 mg CO_2e for DeepSeek-R1 (38.2%), from 6.23 to 3.38 mg CO_2e for Llama-3.1 (45.8%), and from 2.09 to 1.43 mg CO_2e for Mistral (31.4%). At a deployment scale of 10^4 daily pipelines, this corresponds to avoiding approximately 42.6 g CO_2e/day (15.5 kg CO_2e/year) for DeepSeek-R1 workloads, demonstrating that **planning inflation carries measurable environmental costs**.

Table 6. Carbon Footprint Analysis Across Model Architectures. LAS governance reduces emissions by eliminating ceremonial reasoning steps. Regional grid intensity significantly impacts the climate cost of agentic deployment.

Model	Avg. Energy (J)	CO_2e (mg, US)	Inflation Tax	LAS Reduction
DeepSeek-R1 7B	100,410	12.65	+42%	38.2%
Llama-3.1 8B	56,069	7.05	+28%	45.8%
Mistral 7B	18,837	2.38	+15%	31.4%

6. Broader Impact and Ethical Considerations

The introduction of the *Living Agentic System* (LAS) and the *Governance-as-a-Service* (GaaS) architecture has important consequences for the resilience and security of autonomous systems. Our work touches upon several important socio-technical aspects as agents move from self-contained benchmarks to "wild," high-stakes environments:

Securing the Autonomous Supply Chain: We externalize safety into a modular, model-agnostic governance layer, which leaves us a blueprint for securing the autonomous engineering supply chain. LAS guards against accidental or adversary attempts at target leakage, which is significant in industries such as finance, healthcare and legal where data integrity is of utmost importance. This enables the benefits of productivity from autonomous agents to be experienced by organizations without needing to take on the disastrous risks of unregulated reward-hacking on sensitive datasets.

Computational Efficiency and Sustainability: Clearly, this specific learning behavior, which we name as *Planning Inflation*, i.e., the pathological expansion of reasoning traces beyond the point where they can be useful, certainly has important consequences for the environmental footprint of AI. We show that "thinking twice" not only is a security threat, but a main reason for token waste, too, by setting up a resource-aware Step Budget ($b = 25$). Advocating architectures that ensure reasoning stability may mitigate carbon costs linked to extensive agentic rollouts.

Democratic Access to Safe AI: We highlight that high-assurance safety is possible without requiring massive compute resources, simply by focusing on 7B-parameter models and local inference based on the *Ollama*.Our results suggest that architectural constraints can improve the safety and reliability of small, accessible models, although broader comparisons against larger-scale systems remain future work.

Potential for Misuse and Dual-Use: Our framework is designed to protect against this class of harmful behavior, but the modularity of the GaaS interceptor means that, if the constitution which guides the governance is poorly designed, the policies enforced could be injurious ones. Secondly, the results of our observation of the susceptibility of vanilla agents to target leakage may be exploited by malicious agents. To offset this risk, we have made our defense-in-depth architecture open source to enable community-driven policy discovery. Furthermore, governance policies that are incomplete or poorly specified may encourage a false sense of security.

7. Conclusion

We identify the *Integrity Gap* as a systemic failure in autonomous data science agents, arising from unbounded reasoning budgets and the absence of runtime governance. Rather than establishing a new overthinking phenomenon, our results demonstrate how previously observed diminishing returns from test-time compute manifest in autonomous data-science agent pipelines. Using the *Living Agentic System* (LAS) as a diagnostic testbed, we quantify *Planning Inflation* and the resulting *Token Tax*. Our results show that reasoning beyond the *Reasoning Completion Point* (b^*) is largely ceremonial, contributing to computational waste and a non-negligible carbon footprint in 7B–8B parameter models, with marginal utility gains.

Most importantly, we demonstrate that the **GaaS Interceptor** mitigates these risks by identifying intent-aware completion signals, reducing token costs by up to 22% and carbon emissions by 45.8% for Llama-3.1 workloads and improving reasoning efficiency (η) by 54.2% relative to unconstrained reasoning. *DeepSeek-R1*, while highly capable, exhibits an $\sim$100 kJ energy footprint, highlighting a widening sustainability gap that necessitates resource-aware control through the LAS architecture. Overall, this work presents a pragmatic framework for building high-assurance, computationally sustainable agentic pipelines, advancing toward an "Engineering for All" paradigm.

GenAI Usage Disclosure

The authors used generative AI tools for language editing and structural feedback during manuscript preparation. All scientific content, experimental design, methodology, implementation, results, and interpretations were produced entirely by the authors. The generative AI tool was not used for data analysis, code generation, or the formulation of research contributions.

References

Agarwal, A., Sengupta, A., and Chakraborty, T. The Art of Scaling Test-Time Compute for Large Language Models, December 2025. URL http://arxiv.org/abs/2512.02008. arXiv:2512.02008 [cs].

Arunkumar, V., Gangadharan, G. R., and Buyya, R. Agentic artificial intelligence (ai): Architectures, taxonomies, and evaluation of large language model agents, 2026. URL https://arxiv.org/abs/2601.12560.

Atinafu, Y. and Cohen, R. RewardHackingAgents: Benchmarking Evaluation Integrity for LLM ML-Engineering Agents, March 2026. URL http://arxiv.org/abs/2603.11337. arXiv:2603.11337 [cs] version: 1.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Chen, K., Wang, P., Yu, Y., Zhan, X., and Wang, H. Large language model-based data science agent: A survey, 2025a. URL https://arxiv.org/abs/2508.02744.

Chen, X., Xu, J., Liang, T., He, Z., Pang, J., Yu, D., Song, L., Liu, Q., Zhou, M., Zhang, Z., Wang, R., Tu, Z., Mi, H., and Yu, D. Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 9487–9499. PMLR, October 2025b. URL https://proceedings.mlr.press/v267/chen25bx.html.

Gema, A. P., Hägele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael, J., Alex, B., Minervini, P., Chen, Y., Benton, J., and Perez, E. Inverse Scaling in Test-Time Compute, December 2025. URL http://arxiv.org/abs/2507.14417. arXiv:2507.14417 [cs].

Google Data Centers. Efficiency: Power usage effectiveness, 2024. URL https://datacenters.google/efficiency. Accessed: 2026-04-24.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, pp. 79–90, New York, NY, USA, November 2023. Association for Computing Machinery. ISBN 979-8-4007-0260-0. doi: 10.1145/3605764.3623985. URL https://dl.acm.org/doi/10.1145/3605764.3623985.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1 incentivizes reasoning in LLMs through reinforce-

ment learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Hambardzumyan, K., Baldwin, N., Toledo, E., Hazra, R., Kuchnik, M., Omari, B. A., Foster, T. S., Protopopov, A., Gagnon-Audet, J.-C., Mediratta, I., Niu, K., Shvartsman, M., Lupidi, A., Audran-Reiss, A., Pathak, P., Shavrina, T., Magka, D., Momand, H., Dunfield, D., Cancedda, N., Stenetorp, P., Wu, C.-J., Foerster, J. N., Bachrach, Y., and Josifoski, M. AIRA_2: Overcoming Bottlenecks in AI Research Agents, April 2026. URL http://arxiv.org/abs/2603.26499. arXiv:2603.26499 [cs] version: 2.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework, 2023. URL https://arxiv.org/abs/2308.00352. Version Number: 7.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations, December 2023. URL http://arxiv.org/abs/2312.06674. arXiv:2312.06674 [cs].

Jing, L., Huang, Z., Wang, X., Yao, W., Yu, W., Ma, K., Zhang, H., Du, X., and Yu, D. DSBench: How Far Are Data Science Agents from Becoming Data Science Experts?, April 2025. URL http://arxiv.org/abs/2409.07703. arXiv:2409.07703 [cs].

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Li, Z., Wu, G., Wang, C., and Zhao, Y. Limited Reasoning Space: The cage of long-horizon reasoning in LLMs, March 2026. URL http://arxiv.org/abs/2602.19281. arXiv:2602.19281 [cs].

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink, April 2022. URL http://arxiv.org/abs/2204.05149. arXiv:2204.05149 [cs].

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-Tuning Aligned Language Models Compro- Mises Safety, Even When Users Do Not Intend To! 2024.

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., and Sun, M. ChatDev: Communicative Agents for Software Development. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

Sequeda, J., Allemang, D., and Jacob, B. Knowledge Graphs as a source of trust for LLM-powered enterprise question answering. *Journal of Web Semantics*, 85:100858, May 2025. ISSN 1570-8268. doi: 10.1016/j.websem.2024.100858. URL https://www.sciencedirect.com/science/article/pii/S1570826824000441.

U.S. Energy Information Administration. Frequently asked questions (faqs): How much carbon dioxide is produced per kilowatthour of u.s. electricity generation?, 2024. URL https://www.eia.gov/tools/faqs/faq.php?id=74&t=11. Accessed: 2026-04-24.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18 (6), March 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL http://dx.doi.org/10.1007/s11704-024-40231-1.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023. URL http://arxiv.org/abs/2308.08155. arXiv:2308.08155 [cs].

Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., and Wang, Y. When more is less: Understanding chain-of-thought length in llms, 2025. URL https://arxiv.org/abs/2502.07266.

Yue, Y., Zhang, G., Liu, B., Wan, G., Wang, K., Cheng, D., and Qi, Y. MasRouter: Learning to Route LLMs for Multi-Agent Systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15549–15572, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.757. URL https://aclanthology.org/2025.acl-long.757.

Zhang, D., Zhoubian, S., Cai, M., Li, F., Yang, L., Wang, W., Dong, T., Hu, Z., Tang, J., and Yue, Y. DataSciBench: An LLM Agent Benchmark for Data Science, February 2025. URL http://arxiv.org/abs/2502.13897. arXiv:2502.13897 [cs].