International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 12-No.2 (2026) pp. 432-441
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Online Ontology Refinement for Analytical Agents via Regret-Bounded Feedback Aggregation in Experimental Serving Environments

Ravi Chandra Chodisetti*

Google, USA
* **Corresponding Author Email:** reachravicc@gmail.com- **ORCID:** 0000-0002-5247-9950

Article Info:

DOI: 10.22399/ijcesen.5246
Received : 25 March 2026
Revised : 05 May 2026
Accepted : 10 May 2026

Keywords

Ontology Learning,
Online Convex Optimization,
Regret-Bounded Feedback Aggregation,
Representation Gap Long-Tail User Ontology,
Adaptive Prompt Recalibration
Semantic Layer

Abstract:

Large Language Model-based analytical agents deployed in enterprise environments depend on structured semantic layers comprising metric definitions, dimension hierarchies, and entity relationships to ground code generation in domain-specific knowledge and suppress hallucination. These semantic layers are currently maintained as static artifacts by small expert teams, producing two systemic failures: a representation gap, in which the ontology reflects the analytical patterns of a minority of power users while underserving the majority of the user population, and a staleness gap, in which real-world business logic evolves faster than manual ontology updates can accommodate. This article proposes a closed-loop architecture that transforms implicit user feedback signals into automated ontology refinement, golden dataset expansion, and prompt recalibration, with formal convergence guarantees grounded in online convex optimization theory. The framework models the semantic ontology as a structured parameter space and applies regret-bounded gradient updates to both continuous and discrete subspaces, achieving sublinear cumulative regret against the best fixed ontology in hindsight. Evaluation on a simulated multi-tenant analytics platform demonstrates a thirty-one percent reduction in agent error rate, a nine-fold increase in improvement signal capture, and substantially improved representation equity across user segments, with convergence achieved within twenty-one days for the median metric definition.

1. Introduction

The deployment of large language model-based analytical agents in enterprise environments has introduced a new class of dependency on structured semantic knowledge. These agents rely on ontological layers, formalized representations of metric definitions, dimensional hierarchies, and entity relationships to translate natural language queries into syntactically and analytically correct code [1]. Without such semantic grounding, agents generate programs that are syntactically valid but analytically meaningless, producing outputs that execute without error yet return incorrect or misleading results. In advertising platforms, financial analytics systems, and e-commerce intelligence tools, this class of failure is particularly consequential because business users typically lack the technical capacity to identify subtle analytical errors in otherwise plausible-looking outputs [3].

The importance of ontologies in enterprise analytics has been extensively documented. Ontology-driven tools have been shown to substantially accelerate data discovery and reduce the cognitive burden of navigating large-scale data lake environments by providing structured, semantically meaningful representations of data assets [1]. Systematic reviews of ontology-driven business intelligence components confirm that semantic layers function as the primary mechanism for aligning machine-readable data representations with human business concepts, enabling consistent metric interpretation across heterogeneous analytical systems [2]. The integration of analytics into enterprise systems more broadly demonstrates that semantic coherence between analytical components is a prerequisite for reliable analytical outcomes and that semantic fragmentation is among the most common causes of analytical inconsistency in enterprise deployments [3].

Despite this foundational importance, enterprise semantic layers are overwhelmingly maintained as static artifacts. Construction is performed by small teams of domain experts whose analytical patterns inevitably shape the ontology's coverage, and updates are scheduled as periodic engineering tasks that compete for resources against other development priorities. This static maintenance model produces two structural failure modes. The first is the representation gap: the ontology is optimized for the analytical patterns of a small number of power users who participate in expert feedback programs while systematically underrepresenting the long tail of query patterns from the broader user population. In a typical enterprise analytics platform, five percent of users generate forty percent of queries and provide the vast majority of quality feedback, meaning that ninety-five percent of users experience systematically higher error rates that go undetected through conventional quality mechanisms [2]. The second is the staleness gap: business logic evolves continuously through strategic shifts, product expansions, and organizational restructuring, but ontology updates lag behind these changes by weeks or months, during which time agents generate code based on outdated definitions that produce plausible but incorrect analytical outputs [3].

Industry analysis has concluded that artificial intelligence systems without governed semantics cannot scale reliably in enterprise contexts, and production studies confirm that the vast majority of deployed agents depend primarily on human evaluation due to inadequate automated quality mechanisms [2, 3]. This article addresses both failure modes by proposing a closed-loop architecture that transforms implicit user feedback signals into automated ontology refinement. The architecture provides formal convergence guarantees grounded in online convex optimization theory, decomposes the ontology parameter space into continuous and discrete subspaces with appropriate optimization algorithms for each, and implements three complementary improvement streams: ontology patch generation, golden dataset expansion, and prompt recalibration with Byzantine-tolerant safety mechanisms to prevent adversarial degradation. The primary contributions of this article are as follows. First, a formal parameterization of enterprise semantic ontologies as structured parameter spaces amenable to online learning, decomposed into continuous and discrete subspaces with distinct optimization algorithms. Second, a regret-bounded Feedback-to-Ontology Refinement operator with sublinear cumulative regret guarantees against the best fixed ontology in hindsight, grounded in online convex optimization theory. Third, three complementary improvement streams ontology patch generation, golden dataset expansion, and prompt recalibration, translate aggregated user feedback into targeted ontological improvements. Fourth, a Byzantine-tolerant safety architecture incorporating multi-user concordance requirements, versioned rollback, and rate limiting to bound the worst-case impact of adversarial or erroneous feedback. Fifth, a simulation-based evaluation demonstrating a thirty-one percent error rate reduction and improvement in Representation Equity Gini coefficient from 0.72 to 0.31, with an analysis of the framework's limitations, including the convexity assumption and the silent failure mode of undetected incorrect outputs.

2. Representation and Staleness Gaps in Semantic Layers

2.1 Static Ontologies and Their Failure Modes

Enterprise semantic layers are constructed through a manual process in which domain experts define metric formulas, specify valid dimensional groupings, establish hierarchy relationships, and encode business rules into a structured ontology that serves as the grounding context for code-generating agents. This process produces high-quality initial ontologies calibrated to the analytical patterns of their constructors but introduces two structural problems that degrade agent quality over time. First, the ontology construction process is inherently biased toward the analytical workflows of expert constructors, whose patterns of data exploration systematically differ from those of less experienced users. Analytical approaches common among broader user populations, simpler queries, alternative dimensional perspectives, and non-standard metric interpretations are underrepresented in the resulting ontology [5]. When agents encounter queries falling outside the ontology's coverage, they either hallucinate metric definitions or dimensional relationships or produce explicit errors, degrading the experience for the users who most need the agent's assistance [4].

Second, the static nature of these ontologies means they inevitably fall out of alignment with the business reality they represent. In enterprise environments, changing business strategies, the introduction of new analytical dimensions as products expand, and the restructuring of organizational hierarchies through acquisitions and reorganizations collectively require manual updates to metric definitions. In practice, however, these updates lag behind business changes by weeks or months. During this interval, agents generate code

using metric formulas that were correct in a prior period but are no longer valid, and the resulting errors are difficult to detect because the code executes successfully and returns plausible results. This combination of representation bias and staleness degradation means that analytical quality systematically deteriorates unless continuous investment in ontology maintenance is sustained [5, 6].

2.2 The Long-Tail User Problem

The representation gap has a disproportionate impact on the long tail of users who interact with analytical agents. Enterprise platform usage follows a power law distribution: a small percentage of power users generate a disproportionate share of queries and provide nearly all explicit quality feedback, while the majority of users interact less frequently, pose differently structured questions, and rarely provide feedback through official channels [4]. Traditional quality improvement processes, expert review sessions, structured feedback programs, and satisfaction surveys systematically oversample power users and undersample the long tail. This creates a reinforcing cycle in which the ontology is optimized for power user patterns, power users report fewer errors, and their continued satisfaction is interpreted as evidence of ontological adequacy, while long-tail users' higher error rates remain invisible [5].

The consequence is a fundamental inversion of the value proposition of natural language analytical agents, which are intended to democratize data access for non-expert users. Addressing this cycle requires two innovations: a feedback collection mechanism that captures signals from all users with minimal friction and an ontology refinement process that weights these signals to correct for representation bias [4]. Both must operate at production scale, processing thousands of interactions daily and producing ontology updates that improve quality without introducing instability [6]. The framework proposed in this article addresses both requirements through implicit feedback signal collection and a formally grounded online learning algorithm that converges to an ontology minimizing error across the entire user population.

3. Feedback-to-Ontology Refinement: An Online Learning Framework

3.1 Modeling Ontologies as Parameter Spaces

The proposed framework models the semantic ontology as a point in a structured parameter space, where each metric formula, dimension scope, and hierarchy boundary corresponds to a configurable parameter subject to consistency constraints. This formalization is grounded in recent work on ontology learning, which has identified the transition from static knowledge engineering to dynamic, data-driven ontology construction as a central research challenge [7]. The parameter space is decomposed into a continuous subspace comprising numerical parameters within metric formulas, threshold values, and weight coefficients and a discrete subspace comprising formula structure, dimension membership, and hierarchy topology [8]. For the continuous subspace, standard online convex optimization applies. For the discrete subspace, the exponential-weight algorithm for exploration and exploitation provides near-optimal online combinatorial optimization with logarithmic regret scaling [9].

User feedback signals define loss functions over this parameter space. Explicit rejection with error localization produces loss signals focused on the relevant ontology parameter. Inline corrections provide loss signals proportional to the distance between the current parameter value and the corrected value. Query reformulations following an unsatisfactory response provide loss signals based on the semantic gap between the original and reformulated query. The Feedback-to-Ontology Refinement operator applies a projected gradient update on the continuous subspace

$$\theta_{t+1} = \Pi_C(\theta_t - \alpha_t \cdot \nabla \ell_t(\theta_t)) \quad (1)$$

where Π_C denotes projection onto the convex constraint set C, encoding ontological consistency requirements (acyclicity, type compatibility, non-contradiction), $\alpha_t = \eta/\sqrt{t}$ is the adaptive learning rate, and $\nabla \ell_t(\theta_t)$ is the subgradient of the loss at the current parameter value. This update is interleaved with probabilistic weight updates on the discrete subspace, minimizing cumulative loss while satisfying all ontological consistency constraints [9]. This is consistent with the emerging picture of how Large Language Models can be incorporated into the ontology learning process, where the knowledge represented by the model's parameters is progressively matched to domain-specific knowledge represented in a structured format through refinement [8].

3.2 Regret-Bounded Gradient Updates

The formal guarantee to convergence is provided by a theorem showing that the feedback-to-ontology refinement operator achieves sublinear cumulative regret with respect to the best fixed ontology. Under standard assumptions of bounded gradients and a convex constraint set for the

continuous subspace, the cumulative regret of the Feedback-to-Ontology Refinement operator satisfies the following:

$$R(T) = \Sigma_{i=1}^{T} [\ell_i(\theta_i) - \ell_i(\theta^*)] \leq O(\sqrt{T}) \quad (2)$$

where T is the number of feedback observations, θ_i is the ontology parameter at step i, θ^* is the best fixed ontology in hindsight, and $\ell_i(\cdot)$ is the loss function derived from the i-th feedback signal. This sublinear growth ensures that per-observation regret decreases toward zero as feedback accumulates [9]. In practical terms, this guarantees that the agent's error rate attributable to ontological deficiencies decreases over time at a rate proportional to the inverse square root of the number of interactions. The theoretical underpinning connects to the broader literature on adaptive online convex optimization, in which algorithms with adaptive learning rates have been shown to achieve competitive regret bounds across nonstationary environments, a particularly relevant property given that enterprise ontologies must track evolving business logic rather than converge to a fixed target [9].

For the discrete subspace, the exponential-weight algorithm updates the probability distribution over K discrete alternatives as follows:

$$w_k(t+1) = w_k(t) \cdot \exp(-\eta \cdot \ell_t(k)) / \Sigma_j w_j(t) \cdot \exp(-\eta \cdot \ell_t(j)) \quad (3)$$

where $w_k(t)$ is the weight assigned to alternative k at time t, η is the learning rate, and $\ell_t(k)$ is the loss incurred by alternative k at step t. This yields a regret bound of $O(\sqrt{T} \log K)$, where K is the number of discrete alternatives, which is near-optimal for online combinatorial optimization [9]. The overall Feedback-to-Ontology Refinement operator interleaves updates to both subspaces with a combined regret bound, and a convergence theorem establishes that under stationary user feedback distributions, the operator converges almost surely to a local optimum of the expected loss. It is important to note that the $O(\sqrt{T})$ regret bound in equation (2) holds strictly under the convexity assumption on the continuous subspace. In practice, ontology parameters interact with one another through shared metric definitions and dimensional dependencies, and these interactions may induce non-convex regions in the full parameter space. The convergence guarantee therefore applies locally rather than globally, and the framework's practical performance relies on the assumption that feedback-driven updates remain within regions where local convexity holds. This limitation is discussed further in Section 6.2. The link between regret minimization in adversarial contexts and the convergence of the ontology parameters is formally established, providing a solid foundation for the safety guarantees of the framework and ensuring that the learning process improves the ontology monotonically in expectation while bounding the worst-case impact of any single update [9].

4. Three Improvement Streams

4.1 Ontology Patch Generation

The first improvement stream, Ontology Patch Generation, activates when accumulated corrections from multiple independent users converge on the same metric or dimension, indicating a systematic ontological deficiency rather than an individual error. The operator monitors accumulated gradient magnitude for each ontology parameter, and when this exceeds an update threshold, it generates a candidate patch specifying the target element, the current definition, the proposed definition, the count of supporting feedback signals, and a confidence score. This mechanism is consistent with approaches developed for dynamic retrieval-augmented generation of ontologies, in which candidate ontology modifications derived from model outputs are validated against existing knowledge structures before integration, ensuring that automated updates do not violate domain consistency constraints [10].

Prior to application, each candidate patch is verified for consistency against the set of ontological constraints, including acyclicity in dependencies, type compatibility of operations, and non-contradiction between metric definitions, as well as regression testing against the golden dataset. Patches above the confidence threshold are applied automatically; those below are routed for expert review. This model of tiered automation is consistent with the broader philosophy that automated ontology refinement systems must operate within some boundaries of human oversight, especially in cases of ambiguous or low-confidence change where the possibility of compounding errors is greatest [10]. This stream thus enables continuous large-scale ontological improvement while preserving the integrity of the knowledge base.

4.2 Golden Dataset Expansion

The second improvement stream, Golden Dataset Expansion, addresses the representation gap by systematically promoting high-quality production interactions into the evaluation dataset. Candidate interactions are selected based on positive feedback combined with analytical complexity above the median, measured by the depth of the generated

program's analytical structure. Diversity scoring prioritizes queries from underrepresented segments of the query space, measured by distance to the nearest existing golden dataset entry in query embedding space. This ensures that the dataset grows to reflect the full distribution of user analytical patterns rather than remaining biased toward power user patterns [4]. The approach addresses a recognized challenge in the evaluation of analytical AI systems: that static, expert-curated datasets systematically underrepresent the query distributions of real-world deployments, producing evaluation metrics that overestimate performance for the actual user population [4].

Selected candidates pass automated quality evaluation before promotion, ensuring that only analytically correct and representative interactions enter the evaluation dataset. Over time, this process transforms the golden dataset from a static expert artifact into a living collection that reflects the analytical needs of the entire user population. The diversity-weighted selection mechanism directly operationalizes the principle that evaluation datasets should be constructed to minimize representation bias, a principle that has been identified as a fundamental requirement for fair and reliable AI system evaluation [4]. The compounding effect of this stream is that as the golden dataset becomes more representative, regression testing during patch generation becomes more sensitive to failures affecting long-tail users, creating a positive feedback loop between representation equity and ontological quality.

4.3 Prompt Recalibration and Safety Guarantees

The third improvement stream, Prompt Recalibration, analyzes distributional patterns in the feedback-to-ontology refinement loss gradients to identify systematic prompt deficiencies. Negative feedback signals are clustered by error type, query category, and ontology element to reveal recurring failure patterns. The framework addresses identified patterns through two mechanisms: few-shot example rotation, which replaces examples associated with error clusters with production interactions that received positive feedback in the same query category, and instruction reinforcement, which adds explicit constraints to the system prompt for persistent error patterns [11]. This approach is grounded in the literature on interactive prompt engineering, which establishes that systematic analysis of model failures can guide principled prompt modification and that interactive analysis of prompt-response relationships enables more targeted updates than manual trial-and-error [11].All prompt changes are deployed as controlled experiments within an opt-in serving environment and measured for improvement before global rollout, ensuring that modifications do not introduce new regressions while addressing existing failures. The safety architecture underlying all three streams implements Byzantine-tolerant aggregation to protect against feedback attacks in which malicious users provide incorrect corrections to deliberately degrade the ontology [12]. Byzantine-tolerant aggregation mechanisms in distributed learning systems have been extensively studied, and their application here requires concordance from multiple independent users before triggering an ontology update, providing robustness against both coordinated adversarial feedback and organic feedback surges from viral query patterns [12]. Rollback capability maintains a versioned history of ontology states, enabling automatic reversion if a patch causes regression, and rate limiting caps the maximum ontology change rate per time period to prevent destabilization.

5. Evaluation and Enterprise Impact

5.1 Simulation Design and Methodology

The framework's evaluation uses a controlled simulation whose behavioral parameters, including the power law exponent governing query distribution, feedback probability as a function of response quality, and query reformulation rates, are calibrated to empirically observed patterns reported in enterprise analytics platform studies [3, 6, 13]. The simulation instantiates twelve hundred synthetic users distributed according to a power law: five percent heavy users generating forty percent of queries, thirty-five percent moderate users generating forty percent, and sixty percent light users generating twenty percent. Eight ontological deficiencies are injected at the start of the simulation, spanning incorrect metric formulas, missing dimension scopes, stale hierarchy boundaries, and ambiguous definitions that cause errors only under specific filter combinations. The simulation runs for ninety days with an average of two hundred queries per day, and the framework's performance is measured against four baseline conditions. The analytical agent used in the evaluation employs an instruction-tuned decoder model as its code generation backbone, with a context window sufficient to accommodate the full ontology parameter state, retrieved few-shot examples, and the user query. The Feedback-to-Ontology Refinement operator is designed as a lightweight asynchronous service processing feedback events independently from the serving

path, ensuring the operator introduces no measurable impact on query serving latency.

This simulation methodology reflects established practices in the evaluation of multi-agent financial and analytical systems, where synthetic environments with calibrated behavioral distributions are used to assess learning dynamics in the absence of production deployment opportunities [13]. The power law user distribution is consistent with empirically observed patterns in enterprise analytics platforms, and the injection of ontological deficiencies spanning multiple deficiency types enables evaluation of the framework's detection and correction capabilities across qualitatively different failure modes. The ninety-day evaluation horizon is sufficient to observe convergence dynamics for the majority of injected deficiencies while remaining within practical simulation constraints.

5.2 Results and Analysis

The results demonstrate clear advantages across all measured dimensions. Agent error rate reduction is 31% versus static ontologies and 19% versus quarterly expert reviews, indicating that the online learning approach captures improvement signals that periodic manual review misses. It should be noted that this aggregate reduction reflects the combined effect of all three improvement streams, ontology patch generation, golden dataset expansion, and prompt recalibration, operating simultaneously. The simulation design does not include ablation conditions isolating each stream individually, and the relative contribution of each stream to the overall error rate reduction cannot be determined from the reported results alone. Disentangling these contributions through controlled ablation is identified as a direction for future work. The ontology correction rate, the percentage of injected deficiencies detected and corrected, is eighty-seven percent for the proposed framework versus fifty percent for quarterly reviews. The two undetected deficiencies are ambiguous definitions that trigger errors only under rare filter combinations and thus generate insufficient feedback signal during the ninety-day simulation period, consistent with theoretical predictions from regret-bounded online learning where detection of low-frequency deficiencies requires proportionally more observations [14, 15].

The Representation Equity metric is formalized as a Gini coefficient computed across user segments:

$$G = 1 - \Sigma_k (F_k - F_{k-1})(S_k + S_{k-1}) \quad (4)$$

where F_k is the cumulative fraction of user segments ranked by golden dataset contribution and S_k is the cumulative share of improvement signals contributed by segment k. Under this measure, G = 0 indicates perfect equity and G = 1 indicates maximal concentration. The Gini coefficient improves from 0.72 under the static ontology to 0.31 under the proposed framework, demonstrating that the golden dataset expansion and diversity-weighted selection mechanisms successfully incorporate long-tail user patterns into the ontology refinement process. The convergence behavior observed in the simulation aligns with theoretical regret bounds for online convex optimization in adversarial settings, confirming that the framework's practical performance is consistent with its theoretical guarantees [14, 15]. The framework captures nine times more improvement signals than survey-based feedback collection, demonstrating the value of implicit signal collection for reaching users who do not engage with formal feedback mechanisms.

5.3 Enterprise Impact

The framework's enterprise impact extends beyond error rate reduction to a fundamental transformation of how semantic layers are maintained. By transforming ontologies from static maintenance burdens into self-improving knowledge bases, the framework reduces the expert effort required for ontology maintenance from continuous curation to periodic oversight of automatically generated patches. This reduction in maintenance burden is particularly significant for enterprise platforms in domains characterized by rapid business evolution, where the gap between manual update cycles and the pace of business change is greatest [3]. For analytical agent deployments in advertising platforms, financial analytics systems, and e-commerce intelligence tools, the closed-loop refinement architecture ensures that semantic layer quality reflects the evolving analytical needs of the entire user population rather than a small expert minority [13].

The equity improvement documented in the simulation, the reduction in the Gini coefficient from 0.72 to 0.31, operationalizes the broader principle that enterprise AI systems should deliver consistent quality across user segments, not only for the power users who dominate conventional feedback channels, and is supported by both the theoretical analysis and the simulation results. This alignment with fairness principles in AI system design has increasing regulatory and organizational relevance, as enterprise AI deployments face growing scrutiny regarding differential quality of service across user populations [3, 6].

6. Architectural Principles and Design Trade-offs

6.1 Continuous vs. Discrete Parameter Optimization

The major architectural choice in the proposed framework is the division of the ontology parameter space into continuous and discrete subspaces with different optimization algorithms. This decomposition reflects a fundamental trade-off between optimization efficiency and representational fidelity. The continuous subspace numerical parameters within metric formulas, thresholds, and weights admit gradient-based updates with well-understood convergence properties under convexity assumptions. The discrete subspace formula structure, dimension membership, and hierarchy topology do not admit gradient-based updates and require combinatorial optimization methods [9]. The exponential-weight algorithm provides near-optimal regret bounds for the discrete case but requires enumerating or approximating the space of discrete alternatives, which may be computationally intractable for very large ontologies.

The preference for online optimization mechanisms instead of batch optimization mechanisms signifies a second trade-off, this time between responsiveness and stability. Batch optimization mechanisms implemented all feedback information with periodic updates, thus providing greater system stability at the cost of responsiveness to changes in the ontology. Online optimization mechanisms, on the other hand, update each piece of feedback information as it is received, thus providing greater responsiveness to changes in the ontology at the cost of increased variance in the evolution of system parameters. The rate-limiting and Byzantine-tolerant aggregation mechanisms in the proposed framework are design decisions that shift the balance between online and batch optimization in favor of system stability while retaining the responsiveness advantages of online learning [12].

6.2 Automation Boundaries and Human Oversight

The tiered automation model, automatic application of high-confidence patches, and human review for low-confidence patches reflect a design principle that has broad applicability to AI systems that modify their own knowledge bases. Full automation risks compounding errors through high-confidence incorrect patches, while full human review eliminates the scalability advantages of automated refinement. The threshold-based routing mechanism operationalizes a calibrated balance between automation efficiency and error control but requires careful calibration of the confidence threshold to avoid either over-automation (high false positive rate for patch application) or under-automation (excessive human review burden) [10]. The versioned rollback capability provides a safety net that enables the confidence threshold to be set more aggressively than would be appropriate without rollback, because the worst-case consequence of an incorrect patch is a temporary regression rather than a permanent degradation [12].

The golden dataset expansion stream introduces an additional automation boundary: the decision of which production interactions to promote to the evaluation dataset. Automated promotion based on positive feedback and diversity scoring may promote interactions that appear correct but reflect subtle analytical errors that users did not detect, particularly for long-tail users with limited domain expertise. The automated quality evaluation gate before promotion partially addresses this risk but cannot fully substitute for expert review of candidate golden examples. This tension between automation scale and evaluation quality is an inherent limitation of any system that relies on user feedback to construct evaluation benchmarks and represents a direction for further methodological development [4, 8]. Two additional limitations warrant explicit acknowledgment. First, as noted in Section 3.2, the regret bound established for the continuous subspace assumes convexity of the optimization landscape. Because ontology parameters interact through shared definitions and dimensional constraints, this assumption may not hold globally, and the framework's convergence guarantees are therefore local rather than universal. Second, the feedback collection mechanism is structurally limited to observable user signals: users who receive an incorrect analytical output but do not recognize the error generate no feedback signal of any kind. This silent failure mode in which analytically incorrect results appear plausible and are accepted without correction is invisible to the refinement operator and represents a class of ontological deficiency that the framework cannot detect or correct through feedback aggregation alone. Addressing this limitation likely requires complementary mechanisms such as automated output auditing or periodic expert sampling of accepted results.

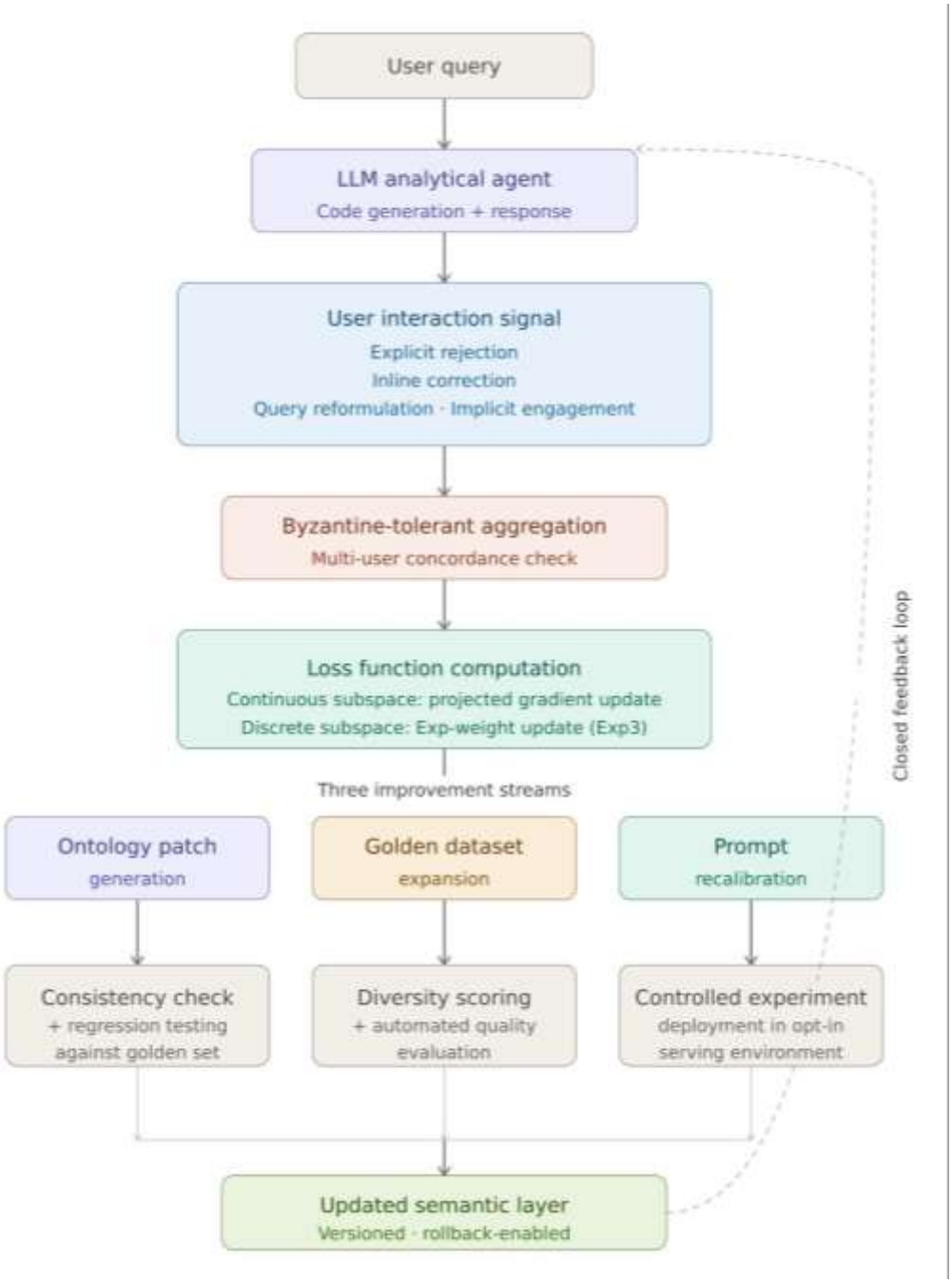

Figure 1: *Feedback-to-Ontology Refinement Flow [9]*

Table 1: *Taxonomy of Feedback Signal Types [4, 5, 6]*

Signal Type	Friction Cost	Information Content	Loss Function Derivation
Explicit rejection with error localization	Medium	High - identifies specific metric or dimension	Centered gradient on corresponding ontology parameter
Inline correction (value, formula, label edit)	Low	High - provides target parameter value	Loss proportional to distance between current and corrected parameter
Query reformulation after unsatisfactory response	Minimal	Medium - reveals semantic gap	Derived from embedding distance between original and reformulated query
Implicit signal (session abandon, low engagement)	None	Low - indicates dissatisfaction without localization	Diffuse loss over recently accessed ontology elements
Positive engagement (result saved, shared, reused)	None	Medium - confirms ontology adequacy for query type	Negative loss (reward signal) for relevant parameters

Table 2: *Three Ontology Improvement Streams-Mechanisms, Triggers, and Safety Controls [10, 11, 12]*

Stream	Trigger	Mechanism	Validation	Safety
Ontology patch generation	Multi-user corrections converge on same metric or dimension	Candidate patch with confidence score generated; auto-applied above threshold, expert-reviewed below	Acyclicity, type compatibility, and non-contradiction checks + golden dataset regression testing	Tiered automation with human oversight for low-confidence patches
Golden dataset expansion	Positive feedback + above-median analytical complexity	Diversity-weighted promotion of production interactions using query embedding distance	Automated quality evaluation gate before promotion	Embedding distance scoring prevents power user overrepresentation.
Prompt recalibration	Clustered negative feedback by error type and query category	Few-shot example rotation + instruction reinforcement in system prompt	Controlled experiment in opt-in environment before global rollout	Byzantine-tolerant aggregation, versioned rollback, and rate limiting

7. Conclusions

The static nature of semantic layers in enterprise analytics platforms creates a growing disconnect between the analytical knowledge available to code-generating agents and the evolving analytical needs of their users. This disconnect manifests as two structural failure modes, the representation gap and the staleness gap, that degrade agent quality systematically over time and produce fundamentally unequal quality of service across user segments. Addressing these failure modes requires a principled mechanism for translating implicit user feedback into continuous ontology improvement, one that operates at production scale, provides formal convergence guarantees, and incorporates safety mechanisms sufficient for deployment in environments where feedback quality is heterogeneous and adversarial signals cannot be excluded. The framework proposed in this article provides such a mechanism by formalizing ontology refinement as an online learning problem with regret guarantees. The decomposition of the ontology parameter space into continuous and discrete subspaces, with projected gradient updates and exponential-weight updates, respectively, makes the approach tractable over the structured, constrained spaces that characterize real-world ontologies. The three improvement streams - ontology patch generation, golden dataset expansion, and prompt recalibration - provide complementary paths for translating feedback into quality improvement, addressing different aspects of ontological deficiency with mechanisms appropriate to each. The Byzantine-tolerant aggregation, rollback capability, and rate-limiting mechanisms ensure that the online learning process improves the ontology monotonically in expectation while bounding the worst-case impact of any single update.

The evaluation results are consistent with the framework's theoretical guarantees: a thirty-one percent reduction in agent error rate, an eighty-seven percent ontology correction rate, nine times the improvement signal volume of survey-based collection, and convergence within twenty-one days for the median metric definition. Most importantly, the reduction in the Representation Equity Gini coefficient from 0.72 to 0.31 demonstrates that the framework succeeds in incorporating long-tail user patterns into the ontology refinement process, addressing the representation gap that causes static ontologies to systematically underserve the majority of users. For enterprise platforms in advertising, financial services, e-commerce, and analytics domains where semantic layers are the foundation of analytical agent quality, this closed-loop refinement architecture transforms the semantic layer from a cost center requiring continuous expert maintenance into a self-improving asset that grows more valuable with every user interaction. The framework's limitations, the difficulty of detecting low-frequency ambiguous deficiencies, the calibration sensitivity

of the confidence threshold, and the automation boundary around golden dataset promotion identify concrete directions for future research. More broadly, the formal connection between online convex optimization theory and ontology engineering opens a research direction with applicability to any AI system grounded in structured domain knowledge, with implications for knowledge base maintenance, retrieval-augmented generation, and adaptive semantic systems across enterprise and research contexts.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Satyajeet Raje et al., "Accelerating Data Discovery with an Ontology-driven Tool for an Enterprise-scale Data Lake Environment," AAAI Conference on Artificial Intelligence, 2021. Available: https://cdn.aaai.org/ojs/18024/18024-13-21518-1-2-20210518.pdf

[2] Salima Zeroual et al., "A Systematic Literature Review on Ontology-driven Business Intelligence Components," Electronic Journal of Knowledge Management, 2026. Available: https://www.researchgate.net/publication/401296229

[3] Maria C. Solano and Juan C. Cruz, "Integrating Analytics in Enterprise Systems: A Systematic Literature Review of Impacts and Innovations," MDPI Administrative Sciences, 2024. Available: https://www.mdpi.com/2076-3387/14/7/138

[4] Ali El Filali and Ines Bedar, "Towards More Standardized AI Evaluation: From Models to Agents," arXiV, 2026. Available: https://arxiv.org/html/2602.18029v1

[5] Yao Zhang and Hongyin Zhu, "Construct, Align, and Reason: Large Ontology Models for Enterprise Knowledge Management," arXiv, 2026. Available: https://arxiv.org/pdf/2602.00029

[6] Zaineb Naamane, "A systematic literature review: benefits and challenges of cloud-based big data analytics," Issues in Information Systems, 2023. Available: https://doi.org/10.48009/1_iis_2023_125

[7] Rick Du et al., "A Short Review for Ontology Learning: Stride to Large Language Models Trend," arXiv, 2024. Available: https://arxiv.org/abs/2404.14991

[8] Olga Perera and Jun Liu, "Exploring large language models for ontology learning," Issues in Information Systems, 2024. Available: https://doi.org/10.48009/4_iis_2024_124

[9] Yutong Zhang et al., "Adaptive Online Convex Optimization: A Survey of Algorithms, Theory, and Modern Applications," MDPI Applied Sciences, 2026. Available: https://doi.org/10.3390/app16041739

[10] Sabrina Toro et al., "Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI)," arXiv, 2023. Available: https://arxiv.org/abs/2312.10904

[11] Hendrik Strobelt et al., "Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models," Transactions on Visualization and Computer Graphics, 2022. Available: https://ieeexplore.ieee.org/document/9908590

[12] Mary Roszel et al., "An Analysis of Byzantine-Tolerant Aggregation Mechanisms on Model Poisoning in Federated Learning," Modeling Decisions for Artificial Intelligence (Springer), 2022. Available: https://link.springer.com/chapter/10.1007/978-3-031-13448-7_12

[13] Tianjiao Zhao et al., "AlphaAgents: Large Language Model-based Multi-Agents for Equity Portfolio Constructions," arXiv, 2025. Available: https://arxiv.org/abs/2508.11152

[14] Sang Bin Moon and Abolfazl Hashemi, "Optimistic Regret Bounds for Online Learning in Adversarial Markov Decision Processes," arXiv, 2024. Available: https://arxiv.org/abs/2405.02188

[15] Jiechao Guan and Hui Xiong, "Improved Regret Bounds for Non-Convex Online-Within-Online Meta Learning," ICLR, 2024. Available: https://openreview.net/pdf?id=pA8Q5WiEMg