PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

Intent Hub: A Self-Healing Semantic Agent Routing System for Resolving Overlap in Agentic Systems

Chenrui Liang , **Peng Xu*** and **Xinyuan Liu**
Beijing University of Posts and Telecommunications
{liangchenrui, xupeng, xinyuanliu}@bupt.edu.cn

Abstract

Semantic overlap challenges accurate agent routing in large-scale agentic systems. We present Intent Hub, a self-healing semantic agent routing system that combines offline asynchronous HITL repair with online Dual Filtering. LLM-generated positive and adversarial negative utterances help construct explicit decision boundaries, enabling interpretable millisecond-level routing. Intent Hub further supports interactive semantic debugging, allowing developers to diagnose conflicts, repair rules, and observe online routing changes.

1 Introduction

The rise of the Agentic Enterprise marks a fundamental shift in how organizations adopt AI agents as integral components of end-to-end business workflows, rather than isolated tools. Reflecting this trend, industry reports indicate that enterprise-level agent deployments grew by 119% in the first half of 2025 alone [Salesforce Research, 2025], driven by the emergence of autonomous frameworks [Park *et al.*, 2023; Wang *et al.*, 2024]. However, this rapid expansion creates a discovery bottleneck, where it becomes increasingly difficult to manually identify the most appropriate agent that satisfies their specific business requirements amidst a growing sea of specialized tools [Qin *et al.*, 2024; Patil *et al.*, 2023]. This challenge necessitates the emergence of Agentic Semantic Routing Entities, as their routing strategies directly determine business success rates and response latency.

Early agent routing relied on rule-based patterns and decision trees [Wang *et al.*, 2023]. While deterministic and interpretable, these methods scale poorly and struggle with natural language variability. More recently, LLM-based approaches leverage prompt engineering and reasoning capabilities to enable flexible intent inference [Brown *et al.*, 2020; Yao *et al.*, 2023; Khattab *et al.*, 2024]. Despite their expressiveness, these methods incur high inference costs, lack transparent decision boundaries, and exhibit unstable latency under high concurrency [Arora *et al.*, 2024]. To balance efficiency and cost, vector-based routing methods match queries against agent descriptions in embedding spaces [Reimers and Gurevych, 2019]. Although efficient, they are inherently vulnerable to semantic overlap, where distinct intents occupy adjacent vector regions, leading to frequent misrouting and limited debuggability [Li *et al.*, 2025]. More critically, existing routing approaches largely lack systematic diagnosis and repair mechanisms. Resolving semantic conflicts often relies on ad hoc tuning or implicit updates, making large-scale routing entities increasingly fragile as agent inventories expand.

In this paper, we propose Intent Hub, a self-healing semantic agent routing system designed to explicitly diagnose, repair, and govern semantic overlap. Intent Hub employs a decoupled architecture: semantic conflicts are localized via visualization-based diagnostics, repaired through LLM-driven contrastive generation and human validation, and resolved by an efficient online filtering pipeline. Experimental results demonstrate that Intent Hub achieves higher routing accuracy than existing vector-based routers while maintaining millisecond-level latency, and approaches the performance of LLM-based routing at a fraction of the cost. Intent Hub is open-sourced on GitHub[1] with a demo video[2].

2 System Architecture

Intent Hub adopts a decoupled architecture in which a central vector database bridges offline configuration and online inference, ensuring that slower asynchronous optimization remains outside the online fast path, as illustrated in Figure 1.

2.1 Offline Plane: Indexing and Self-Healing

The offline plane focuses on controllable semantic boundaries and follows the Configuration-as-Truth principle, in which all routing rules are managed through a unified JSON configuration.

We define self-healing as a human-in-the-loop (HITL) closed-loop workflow that integrates visual semantic diagnosis, LLM-based contrastive generation, and human confirmation to incorporate validated updates.

When an agent is registered, its utterances are encoded into a high-dimensional vector space. Unlike conventional approaches that rely solely on positive intent descriptions, Intent Hub builds a dual index that explicitly captures both core intents (positive utterances) and boundary constraints (negative

*Corresponding author.

[1]https://github.com/Liangchenrui/intent-hub
[2]https://youtu.be/bWHMFci6Pkc?si=z7W_GVkbC3i0_Udp

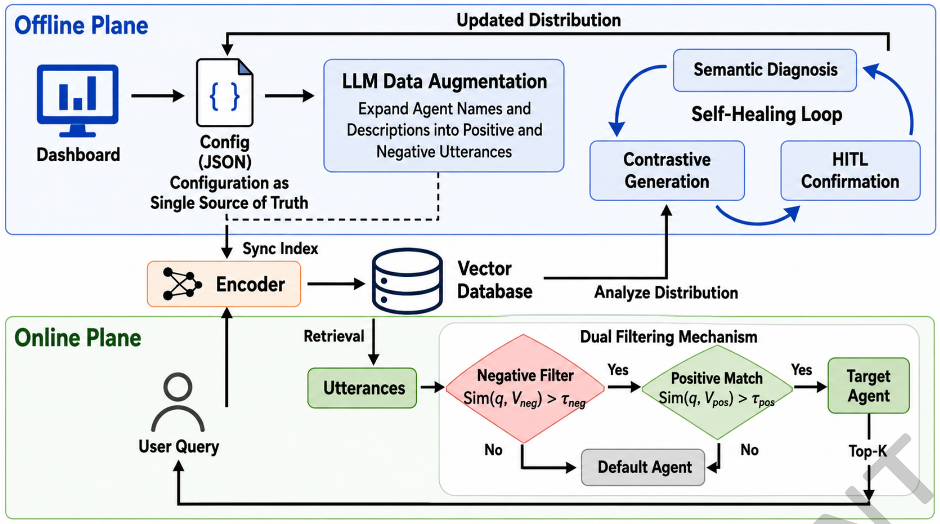


Figure 1: The architecture of Intent Hub. The offline plane asynchronously mitigates semantic overlap through HITL self-healing, while the online plane serves the fast routing path via Dual Filtering Mechanism. The two planes are decoupled and connected through a central vector database.

utterances) [Singhal *et al.*, 2023]. In the background, a monitoring process scans the semantic space for high-risk overlap regions. Once detected, an LLM-based optimizer generates candidate augmentative positive and adversarial negative utterances, which are reviewed through an interactive interface before being committed to the configuration.

2.2 Online Plane: Inference with Dual Filtering

During online inference, the vector database is treated as a read-only cache, and semantic overlap is resolved through Dual Filtering Mechanism. This latency-critical fast path does not call the LLM optimizer or wait for HITL confirmation.

Stage 1: Negative Filter. The query vector q is first matched against the negative index. If the similarity $\mathrm{Sim}(q, V_{\text{neg}}) > \tau_{\text{neg}}$, the request is immediately rejected, effectively filtering ambiguous or adversarial inputs.

Stage 2: Positive Match. Queries passing the negative filter are then matched against the positive utterances of candidate agents. Routing is triggered only if $\mathrm{Sim}(q, V_i) > \tau_i$, where τ_i is a per-agent confidence threshold. If no agent satisfies its respective threshold, the request falls back to a default agent. Intent Hub returns the top-K results that meet the criteria.

3 Key Features and Innovations

Intent Hub provides large-scale agentic systems with visualized, interactive, and repairable semantic routing. Unlike traditional vector-based “black-box” approaches, it emphasizes interpretable diagnostics and LLM-enhanced decision boundary optimization. Figure 2 illustrates the end-to-end workflow and semantic-space evolution.

3.1 Visual Semantic Diagnostics

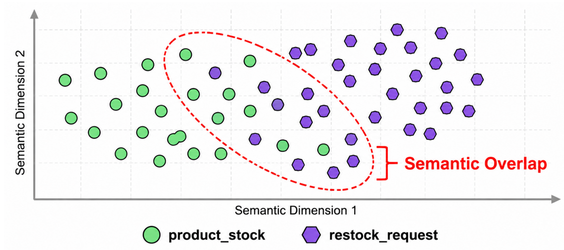
To address unpredictable semantic overlap introduced by new intents, Intent Hub exposes conflict evidence without automatically modifying routing rules. UMAP [McInnes *et al.*, 2020] projects high-dimensional utterance embeddings into a two-dimensional interactive view (Figure 2a), highlighting overlap between adjacent intents. Two metrics assess overlap risk: (1) *Region Overlap*, measuring centroid-level cosine similarity to identify intent-level structural overlap, and (2) *Instance Conflict*, locating specific utterance pairs with high neighborhood proximity. Together, these metrics shift routing maintenance from heuristic guesswork to data-driven decision-making.

3.2 LLM-Driven Self-Healing

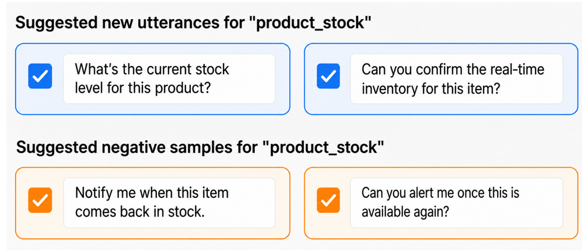
The core innovation is a HITL self-healing loop for semantic repair. Upon conflict detection, an LLM optimizer applies contrastive generation to propose repair candidates for developer confirmation (Figure 2b). In addition to generating augmentative positive utterances, it produces adversarial negative utterances–semantically similar boundary cases that must be explicitly rejected–to sharpen decision boundaries under semantic overlap [Mou *et al.*, 2022]. After human review, validated constraints are integrated into the routing configuration, and previously overlapping intents are separated by clearer decision margins (Figure 2c), substantially reducing false triggering during online routing.

3.3 The *Configuration-as-Truth* Paradigm

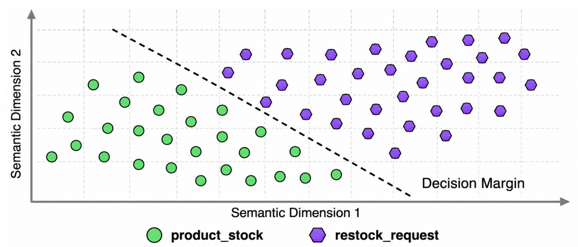
Intent Hub adopts a decoupled *Configuration-as-Truth* paradigm to ensure consistency. All routing logic, including utterances and thresholds, is defined in a static JSON file (Figure 2d) as the single source of truth. The vector database is treated as a derived state, enabling version control and rollback via standard Git workflows and ensuring consistency between development and production [Fowler, 2016].



(a) Initial Semantic Overlap Diagnosis



(b) LLM-Driven Contrastive Repair



(c) Reshaped Post-Repair Decision Margins

(d) Configuration as Single Truth

Figure 2: Visual semantic diagnostics and LLM-driven HITL self-healing workflow for semantic routing.

Method	B77 (Acc/Lat)	HWU64 (Acc/Lat)	FineEC (Acc/Lat)
LLM (3-shot)	86.02/0.61	**87.27**/0.63	92.50/0.64
LLM (0-shot)	59.74/0.62	79.07/0.68	91.00/0.63
Sem. Router	91.30/**0.04**	81.32/**0.04**	81.50/**0.01**
IH (Base)	92.82/0.05	83.92/0.05	91.00/0.04
IH (Repair)	**93.15**/0.10	85.78/0.10	**94.50**/0.09

Table 1: Accuracy (%) and latency (s). *Note:* LLM: DeepSeek-V3.2 [DeepSeek-AI *et al.*, 2025]. Sem. Router and Intent Hub use BAAI/bge-m3 [Chen *et al.*, 2024], Qdrant [Qdrant Solutions GmbH, 2020], Top-K=5. "IH (Base)" = positive index only; "IH (Repair)" = dual-stage filtering.

4 Evaluation

To validate the practicality of Intent Hub under high semantic overlap routing scenarios, we conduct a comparative evaluation against LLM-based routing methods and Semantic Router [Aurelio AI, 2023] under a unified experimental setup.

4.1 Datasets

We evaluate Intent Hub on three intent datasets with different semantic overlap characteristics:

- **Banking77** [Casanueva *et al.*, 2020], which contains 77 banking-domain intents with high semantic similarity;
- **HWU64** [Liu *et al.*, 2019], a 64-intent dataset covering multiple general-purpose domains;
- **FineEC-10**[3], a balanced high-overlap e-commerce dataset contributed by this work. It contains 500 JSONL utterances across 10 intents, with 300 training samples and 200 test samples, 50 samples per intent, and approximately 10% highly conflicting boundary samples designed to test fine-grained semantic discrimination.

4.2 Main Results and Analysis

The experimental results are summarized in Table 1. On the highly overlapping FineEC-10 dataset, Intent Hub with self-healing enabled achieves the highest accuracy, significantly outperforming the vector-based Semantic Router and matching or exceeding the performance of LLM-based methods. This demonstrates that offline generation of adversarial negative utterances, combined with online dual filtering, effectively mitigates routing errors caused by semantic overlap.

Although the self-healing variant introduces additional retrieval overhead due to dual filtering, its overall latency remains substantially lower than that of LLM-based routing. Compared to Semantic Router, the base variant exhibits a moderate latency increase, primarily due to finer-grained routing control and runtime configuration management, an overhead that is acceptable in exchange for improved robustness and maintainability.

5 Conclusion

We present Intent Hub, a self-healing semantic agent routing system that addresses semantic overlap and "black-box" opacity. By integrating visual diagnostics with HITL repair, Intent Hub achieves millisecond-level online routing with explicitly controllable decision boundaries. In the demonstration, users can diagnose conflicts, repair rules through the offline workflow, and observe online routing updates. Future work will study threshold calibration, repair-loop evaluation, stronger baselines, and component ablations.

[3]https://github.com/Liangchenrui/FineEC-10

Acknowledgments

This work was supported by National Key R&D Program of China (Grant No. 2024YFF0907400).

References

[Arora *et al.*, 2024] Gaurav Arora, Shreya Jain, and Srujana Merugu. Intent detection in the age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1559–1570, 2024.

[Aurelio AI, 2023] Aurelio AI. Semantic router. https://github.com/aurelio-labs/semantic-router, 2023. Open-source project.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Casanueva *et al.*, 2020] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online, 2020. Association for Computational Linguistics.

[Chen *et al.*, 2024] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[DeepSeek-AI *et al.*, 2025] DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, et al. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[Fowler, 2016] Martin Fowler. Infrastructure as code. https://martinfowler.com/bliki/InfrastructureAsCode.html, 2016. Accessed: 2025-02.

[Khattab *et al.*, 2024] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, et al. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*, 2024.

[Li *et al.*, 2025] Zhaoheng Li, Silu Huang, Wei Ding, Yongjoo Park, and Jianjun Chen. SIEVE: Effective filtered vector search with collection of indexes. *Proceedings of the VLDB Endowment*, 18(11):4723–4736, 2025.

[Liu *et al.*, 2019] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the 10th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2019.

[McInnes *et al.*, 2020] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.

[Mou *et al.*, 2022] Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for OOD intent discovery. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1517–1529, 2022.

[Park *et al.*, 2023] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[Patil *et al.*, 2023] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

[Qdrant Solutions GmbH, 2020] Qdrant Solutions GmbH. Qdrant vector database. https://qdrant.tech, 2020. Accessed: 2025-02.

[Qin *et al.*, 2024] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*, 2024.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

[Salesforce Research, 2025] Salesforce Research. Agentic enterprise index: H1 2025. https://www.salesforce.com/news/stories/agentic-enterprise-index-insights-h1-2025/, 2025. Industry Report.

[Singhal *et al.*, 2023] Bhavuk Singhal, Ashim Gupta, V P Shivasankaran, and Amrith Krishna. IntenDD: A unified contrastive learning approach for intent detection and discovery. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14204–14216, 2023.

[Wang *et al.*, 2023] Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems: Data, task and models, 2023.

[Wang *et al.*, 2024] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

[Yao *et al.*, 2023] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.