ViSemCrew: Divide and Conquer Vietnamese Semantic Parsing Through Multi-Agent Intelligence

Hao Phu Phan[*]
Vietnam Silicon
Ho Chi Minh City, Vietnam
hao.phan@vnsilicon.net

Khiem Vinh Tran[*] [†]
University of Information Technology
Vietnam National University
Ho Chi Minh City, Vietnam
khiemtv@uit.edu.vn

Abstract

Semantic parsing, which involves converting natural language into formal, machine-understandable representations, is essential for many applications such as question answering, dialogue systems, and information extraction. Vietnamese presents unique challenges in semantic parsing due to its analytic morphological structure, use of numeral classifiers, flexible word order, and frequent omission of tense and subject pronouns. Combined with limited annotated data and inconsistent labeling standards, these factors complicate developing accurate and robust parsers for Vietnamese. To address these challenges, we propose ViSemCrew, a multi-agent workflow framework that orchestrates specialized linguistic agents through a structured pipeline. Our framework employs a series of interconnected agents including a Linguistic Analysis Agent, Concept Extraction Agent, Graph Construction Agent, and Validation Agent, each performing specific semantic parsing subtasks. The agents collaborate through a controlled workflow with iterative refinement capabilities, enabling systematic error correction and quality assurance. We evaluate our framework on the VLSP 2025 ViSem Task, which features diverse Vietnamese text domains annotated in PENMAN format. Results show that our approach outperforms existing baselines in Smatch scores, demonstrating its effectiveness in handling Vietnamese-specific linguistic complexities through coordinated multi-agent processing.

1 Introduction

Vietnamese natural language processing (NLP) has seen significant advances in recent years, addressing a wide array of tasks such as machine reading comprehension (Kiet et al., 2022), visual question answering (Le et al., 2024), and conversational AI (Luu et al., 2021). Research efforts have explored various approaches, including lexical-based methods enhanced with external knowledge (Nguyen et al., 2020), pre-trained sequence-to-sequence models (Tran et al., 2023), and hybrid multimodal systems (Tran et al., 2024), tailored specifically to the linguistic complexities of Vietnamese. These studies have contributed valuable datasets and models that improve the understanding and processing of Vietnamese text across different domains and applications.

Semantic parsing (Thi et al., 2013), the process of converting natural language sentences into formal, machine-interpretable meaning representations, is a fundamental task in natural language processing (NLP) with far-reaching applications ranging from question answering and dialogue systems to information extraction and knowledge base population. While semantic parsing has achieved impressive results in widely studied languages like English, the development of robust semantic parsers for Vietnamese remains an open challenge (D. Huynh et al., 2024). This is largely due to the unique linguistic characteristics of Vietnamese, including its analytic morphological structure, the pervasive use of numeral classifiers, a relatively free word order, and frequent omission or implicit marking of tense, aspect, and subject pronouns. These features complicate the problem of accurately aligning surface forms with their intended semantic roles and relations.

Current Vietnamese semantic parsing must also contend with limited annotated resources and inconsistent annotation standards, further restricting the performance and generalizability of existing approaches (Nguyen et al., 2024). Although initial efforts have focused on rule-based grammars and semantic role labeling adapted from English frameworks, more recent advances have leveraged neural sequence-to-structure models and the Abstract Meaning Representation (AMR) formalism,

[*]These authors contributed equally to this work.
[†]Corresponding author

enabling greater expressive power and scalability. However, purely data-driven models often lack explicit mechanisms to incorporate linguistic constraints, systematic validation, or structured error correction, resulting in parsers which can be brittle when confronting ambiguous or complex constructions, especially in real-world, multi-domain settings.

Motivated by the inherent complexities of Vietnamese semantic parsing, our work introduces ViSemCrew, a multi-agent workflow framework that decomposes the semantic parsing task into specialized subtasks handled by coordinated agents. The ViSemCrew framework is designed to emulate systematic problem-solving by organizing the parsing process into distinct phases: linguistic preanalysis, concept extraction, graph construction, and validation. Each phase is managed by specialized agents that focus on specific aspects of the parsing task, enabling more targeted and effective processing. This approach allows the framework to incorporate explicit linguistic knowledge, perform systematic validation, and implement structured error correction mechanisms.

Our framework deploys multiple specialized agents: (1) a Linguistic Analysis Agent that performs morphological and syntactic analysis, (2) a Concept Extraction Agent that identifies semantic concepts, (3) a Graph Construction Agent that builds semantic relationships, and (4) a Validation Agent that ensures output quality through multiple validation steps. The agents operate within a controlled workflow that includes feedback loops for iterative refinement and fallback mechanisms for robust error handling.

Building upon this framework, we benchmarked our approach by participating in The 2025 VLSP Shared Task on Semantic Parsing (ViSem Task)[1], a shared semantic parsing challenge specifically designed for Vietnamese. The ViSem Task offers a comprehensive, multi-domain dataset that encompasses news articles, literary texts, and user-generated reviews, all annotated using PENMAN format (Goodman, 2020) to capture Vietnamese-specific semantic phenomena. Evaluation of system outputs is performed using established metrics such as Smatch (Cai and Knight, 2013), which measures the similarity between predicted and gold-standard semantic graphs. This metric provides rigorous and interpretable assessments of semantic parsing performance, enabling fair comparison with state-of-the-art baselines.

Our contributions are as follows:

- We propose a novel multi-agent workflow for Vietnamese semantic parsing that decomposes the complex parsing task into specialized subtasks, each handled by dedicated agents with specific expertise and validation capabilities.
- We design a structured pipeline with iterative refinement mechanisms, systematic validation, and robust fallback strategies that enhance parsing accuracy and reliability for Vietnamese text.
- We conduct extensive experiments on the VLSP 2025 Shared Task on Semantic Parsing (Task 9). Our approach achieves competitive performance in this challenging evaluation, demonstrating significant improvements over baseline methods in Smatch scores.

2 Related Work

Semantic parsing for Vietnamese has evolved through diverse approaches, encompassing rule-based grammars, semantic role labeling, and meaning representation frameworks such as Abstract Meaning Representation (AMR). Early work (Nguyen et al., 2013) applied a unification-based grammar to simple Vietnamese sentences, defining taxonomies of nouns and feature structures for nouns and verbs and using syntactic–semantic unification rules to achieve over 84% precision and recall on a news-title corpus of 500 sentences.

Subsequent research focused on semantic role labeling (SRL), treating Vietnamese SRL as a supervised classification task akin to PropBank annotation. Phuong et al. (Phuong et al., 2017) constructed the first Vietnamese PropBank by adapting PropBank roles to Vietnamese idiosyncrasies—introducing fine-grained roles for adjectives, numerals, nouns, and prepositions—and developed a system integrating distributed word embeddings and integer linear programming inference to achieve an F1 score of 74.77%.

Building on these foundations, Ha and Nguyen (Linh and Nguyen, 2019) introduced an AMR-based meaning representation for Vietnamese, adapting the English AMR schema to Vietnamese by incorporating labels for classifiers, function-word–encoded tense, and co-reference, and applied it to a manually annotated translation of The Little

[1] https://vlsp.org.vn/vlsp2025/eval/visemparse

Prince. Their work highlighted challenges in mapping Vietnamese function words and classifiers to AMR concepts and proposed extensions to capture Vietnamese-specific semantic phenomena. By providing a large, manually validated UD-style dataset and fostering innovation across parsing paradigms, VLSP 2020 (Linh et al., 2020) laid the groundwork for subsequent work on Vietnamese AMR parsing and semantic role labeling, bridging the gap between syntactic dependency representations and deeper semantic graph structures.

More recently, research has addressed task-oriented semantic parsing, notably Text-to-SQL. Nguyen et al. (Tuan Nguyen et al., 2020) released a Vietnamese Text-to-SQL dataset by translating the Spider benchmark and demonstrated that human-translated questions yield substantially higher parsing accuracy than machine translations, with automatic word segmentation and NPMI-based cell linking further improving performance.

Building on the diverse foundations of Vietnamese semantic parsing—ranging from rule-based grammars and supervised role labeling to AMR-style meaning representations and Text-to-SQL frameworks—the research community has repeatedly confronted three critical barriers: the paucity of large, varied corpora; the lack of uniform annotation standards; and the absence of a common evaluation protocol. To overcome these obstacles, the 2025 VLSP Shared Task on Semantic Parsing (ViSem) was introduced. ViSem draws together a corpus sourced from newswire articles, a literary translation of The Little Prince, and a collection of online reviews, thereby addressing the challenge of data scarcity. Each sentence is annotated in PENMAN-formatted AMR and a corresponding logical-form schema, enhanced with Vietnamese-specific constructs—numeral classifiers, aspectual markers, and pronominal co-reference—to ensure consistency and typological fidelity. Finally, a unified evaluation framework employs Smatch scoring for AMR graph similarity and exact-match metrics for logical forms, enabling direct comparison of systems regardless of their underlying architectures.

3 Methodology

3.1 Framework Overview

ViSemCrew is a multi-agent workflow framework designed to address the complexities of Vietnamese semantic parsing through systematic decomposition and specialized processing. Inspired by Agentgroupchat-v2 (Gu et al., 2025), our framework organizes the parsing process into a structured pipeline where each stage is managed by specialized agents that focus on specific aspects of semantic analysis. This design enables targeted processing of Vietnamese linguistic phenomena while maintaining systematic quality control throughout the parsing process.

The framework operates on the principle of divide-and-conquer, breaking down the complex semantic parsing task into manageable subtasks that can be handled more effectively by specialized components. Each agent in the framework is designed with specific expertise and validation capabilities, allowing for focused processing and systematic error detection and correction. The framework is implemented using Google's Gemini 2.5 Pro[2] as the underlying language model for the agents. Figure 1 illustrates the overall architecture and workflow of the ViSemCrew framework.

3.2 Multi-Agent Architecture

As shown in Figure 1, our framework consists of five primary agents organized in a sequential pipeline with validation and repair mechanisms. The architecture demonstrates the flow from Vietnamese text input through specialized processing agents to the final AMR graph output, with Gemini 2.5 Pro powering each agent and a knowledge base providing guidance throughout the process.

3.2.1 Linguistic Analysis Agent

The Linguistic Analysis Agent serves as the foundation of our parsing pipeline, performing comprehensive morphological and syntactic analysis of Vietnamese input sentences. This agent specializes in part-of-speech tagging and dependency parsing, with particular attention to Vietnamese-specific constructions such as passive voice markers ("được", "bị") and complex noun phrases with numeral classifiers.

The agent employs structured output generation to ensure consistent linguistic annotation, producing detailed morphological and syntactic features for each word in the input sentence. Special handling is implemented for Vietnamese passive constructions, where the agent correctly identifies passive subjects using the 'nsubj:pass' relation and passive auxiliary markers using 'aux:pass'.

[2] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

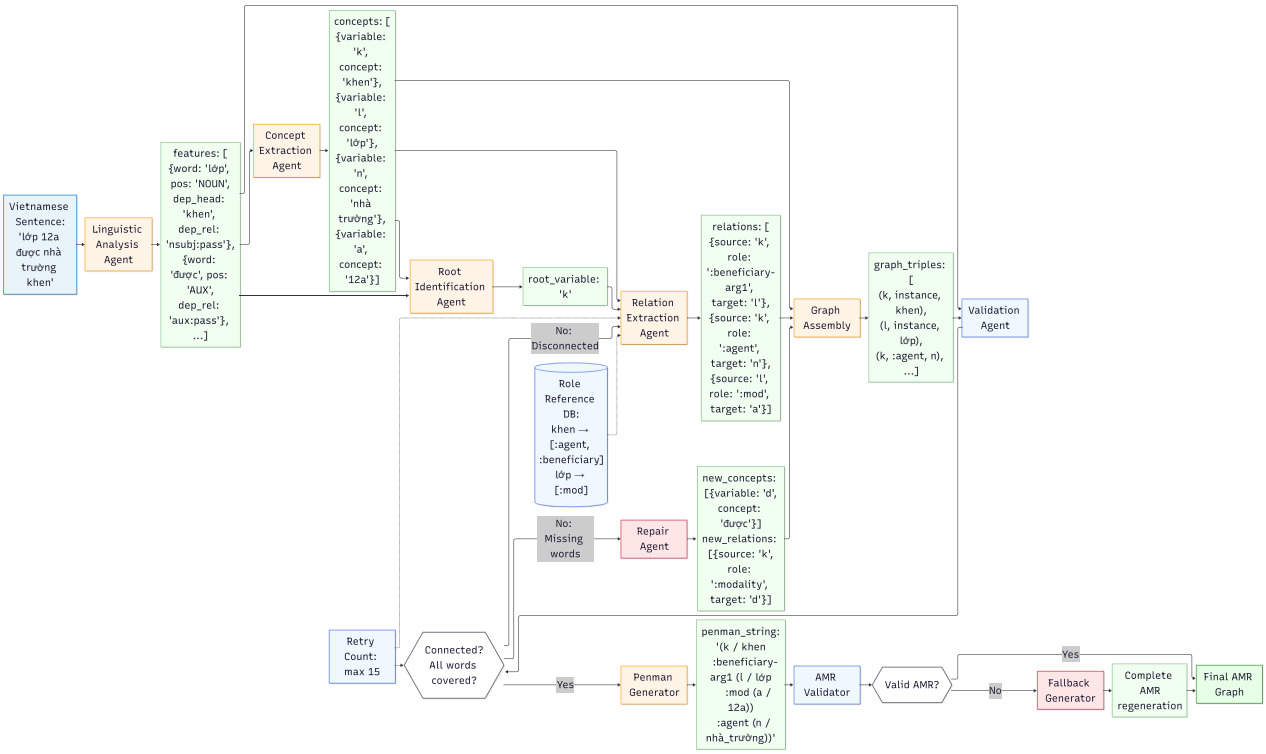

Figure 1: ViSemCrew Multi-Agent Framework Architecture for Vietnamese Semantic Parsing

3.2.2 Concept Extraction Agent

The Concept Extraction Agent focuses on identifying and extracting semantic concepts from the linguistically analyzed input. This agent is responsible for recognizing entities, actions, properties, and modalities within Vietnamese text, with particular attention to ensuring comprehensive coverage of content words and important function words.

The agent employs few-shot learning with Vietnamese-specific examples to guide concept identification. It generates structured concept representations including variable assignments, canonical concept names, and semantic descriptions. Post-processing mechanisms clean and normalize concept names to ensure consistency with AMR conventions.

3.2.3 Graph Construction Agent

The Graph Construction Agent consists of two specialized sub-agents that work in sequence to build the semantic graph structure:

Root Identification Sub-Agent: This component determines the root concept of the semantic graph by analyzing the linguistic structure and semantic relationships within the sentence. It considers factors such as main predicates, syntactic heads, and semantic centrality to identify the most appropriate root concept.

Relation Extraction Sub-Agent: This component identifies and establishes semantic relationships between concepts, creating the edge structure of the semantic graph. It employs knowledge-based role suggestions derived from reference AMR data and implements systematic validation to ensure graph connectivity and completeness.

3.2.4 Validation Agent

The Validation Agent performs comprehensive quality assurance through multiple validation mechanisms:

Connectivity Validation: Ensures that all concepts in the generated graph are properly connected and reachable from the root node. Disconnected components are identified and flagged for repair.

Coverage Validation: Verifies that all content words from the input sentence are represented in the semantic graph. Missing words are identified and marked for incorporation.

Format Validation: Validates the syntactic correctness of generated AMR representations using established parsing libraries and format checkers.

3.2.5 Repair Agent

The Repair Agent handles error correction and graph completion when validation failures are detected. This agent operates in two modes:

Targeted Repair Mode: Addresses specific issues such as missing word coverage by generating additional concepts and relations to complete the semantic graph.

Fallback Generation Mode: Provides complete AMR regeneration when systematic repair is insufficient, employing end-to-end generation with comprehensive few-shot examples.

3.3 Workflow Orchestration

The ViSemCrew framework orchestrates agent interactions through a structured workflow with conditional branching and iterative refinement capabilities. The workflow begins with linguistic analysis and progresses through concept extraction, graph construction, and validation, with feedback loops enabling systematic error correction.

3.3.1 Sequential Processing Pipeline

The primary workflow follows a sequential processing pattern where each agent builds upon the output of previous agents:

1. **Linguistic Pre-Analysis:** The input sentence undergoes comprehensive morphological and syntactic analysis.
2. **Concept Extraction:** Semantic concepts are identified and extracted based on linguistic features.
3. **Root Identification:** The central concept serving as the graph root is determined.
4. **Relation Extraction:** Semantic relationships between concepts are established.
5. **Graph Assembly:** Individual components are assembled into a complete semantic graph.

3.3.2 Validation and Refinement Loop

After initial graph construction, the framework enters a validation and refinement phase:

1. **Connectivity Assessment:** The Validation Agent checks graph connectivity and word coverage.
2. **Conditional Branching:** Based on validation results, the workflow branches to appropriate repair mechanisms.
3. **Iterative Refinement:** Failed validations trigger targeted repairs with retry mechanisms up to a maximum iteration limit.

3.3.3 Quality Assurance and Fallback Mechanisms

The framework implements multiple levels of quality assurance:

- **Primary Validation:** Standard connectivity and coverage checks with targeted repairs.
- **Format Validation:** AMR syntax validation using established parsing libraries.
- **Fallback Generation:** Complete regeneration when systematic approaches fail.
- **Minimal Fallback:** Last-resort minimal AMR generation to ensure system robustness.

3.4 Vietnamese-Specific Adaptations

Our framework incorporates several adaptations specifically designed for Vietnamese linguistic characteristics:

3.4.1 Morphological Processing

Special handling for Vietnamese analytic morphology, including compound word recognition and numeral classifier processing. The framework recognizes Vietnamese-specific grammatical patterns and adjusts concept extraction accordingly.

3.4.2 Passive Voice Handling

Dedicated processing for Vietnamese passive constructions using "được" and "bị" markers, with appropriate role assignment for passive subjects and agents.

3.4.3 Flexible Word Order Accommodation

Robust concept and relation extraction that accommodates Vietnamese flexible word order through dependency-based analysis rather than position-based heuristics.

3.4.4 Implicit Element Recovery

Mechanisms to handle frequently omitted elements in Vietnamese such as tense markers and subject pronouns through contextual inference and default value assignment.

3.5 Knowledge Integration

The framework integrates external knowledge through a role reference database built from training data. This database maps Vietnamese concepts to commonly associated semantic roles, providing guidance for relation extraction and improving consistency with established AMR conventions.

The knowledge integration mechanism operates during relation extraction, providing role suggestions based on concept types while maintaining flexibility for context-specific adaptations. This approach balances systematic knowledge application with adaptability to novel constructions.

4 Evaluation

4.1 Evaluation metrics

The Smatch score (Cai and Knight, 2013) is an evaluation metric that quantifies the similarity between two Abstract Meaning Representations (AMRs) by measuring the overlap of their semantic structures. Since AMRs use variables to represent entities and events, which can differ in naming, Smatch finds the optimal one-to-one mapping between variables in the two AMRs to maximize the number of matching triples.

Formally, if M is the number of matching triples under the best variable alignment, T is the total number of triples in the system AMR, and G is the total number in the gold AMR, then precision P and recall R are defined as:

$$P = \frac{M}{T}, \quad R = \frac{M}{G}$$

The Smatch score is the harmonic mean (F1-score) of precision and recall:

$$F1 = \frac{2 \times P \times R}{P + R}$$

Smatch is widely adopted for evaluating semantic parsers and measuring inter-annotator agreement, providing a principled and interpretable metric for whole-sentence semantic similarity.

4.2 Results

Figure 2: ViSemCrew Performance on VLSP 2025: Public vs Private Test Results

We evaluate the performance of our ViSemCrew model on the VLSP 2025 Vietnamese semantic task using both private and public test sets. The evaluation metrics reported include F1 score, Precision, and Recall, which provide a comprehensive view of the model's ability to balance precision and sensitivity in semantic understanding.

On the private test set, the ViSemCrew model achieves an F1 score of 0.42, with a Precision of 0.44 and Recall of 0.40. In contrast, the public test set results show improved performance, with an F1 score of 0.46, Precision of 0.43, and Recall of 0.50. These results indicate that the model generalizes well, achieving higher Recall on the public test data, which suggests greater ability to identify relevant semantic instances, despite slightly lower Precision compared to the private test.

The marked difference in Recall between the public and private tests (0.50 vs. 0.40) points to a possible variance in data distribution or annotation criteria across these two partitions. However, the consistent F1 scores close to the 0.4–0.46 range reflect stable overall performance. The complementary relationship between Precision and Recall in each test supports the robustness of our model on the semantic task.

Overall, these results demonstrate the effectiveness of ViSemCrew in addressing Vietnamese semantic parsing challenges, with solid performance across key metrics. Further analysis and model tuning may help in narrowing the gap in Recall between test sets and improving Precision while maintaining high sensitivity.

4.3 Error analysis

Figure 3 visualizes the performance of the ViSemCrew proposed model on the VLSP 2025 ViSem Public Test, comparing true labels against predicted ones across three categories: Correct, Misparsed, and Wrong Concept. The model achieves its highest accuracy in the "Correct" category, with 50 true positives, while misclassifying 5 as "Misparsed" and 3 as "Wrong Concept." In the "Misparsed" category, 40 cases are correctly identified, with some confusion: 7 cases labeled as "Correct" and 6 as "Wrong Concept." The "Wrong Concept" label is correctly predicted for 30 instances, but 2 are misclassified as "Correct" and 4 as "Misparsed." Overall, the model shows a balanced performance with a stronger ability to recognize the "Correct" class, though improvements are needed in distinguishing between "Misparsed" and "Wrong Concept."

Comparison of the parsed output with the ground truth semantic annotations reveals challenges automatic semantic parsers face in capturing the full nuance of expert annotations. A notable difference lies in handling named entities and proper nouns. For example, the phrase "anh chủ tịch xã Bùi Văn Luyến" is treated by the parser as a compound noun without breaking down individual components such as "Bùi," "Văn," and "Luyến" or their specific roles. In contrast, the ground truth annotations decompose names meticulously, linking explicit tokens to entities, thus showing more precise entity recognition.

Sentence segmentation also varies: the parser often treats complex sentences with multiple clauses as single units, which blurs event and role boundaries. The ground truth divides such sentences into clear sub-sentences, providing better semantic clarity. For instance, "hiện nay xã có 68 tổ nhân dân, mỗi tổ phụ trách 40 gia đình" is split into separate statements about group existence and responsibilities.

Modality, negation, and polarity handling show marked differences. In complex sentences such as "chủ trương tốt nhưng dân không hiểu, không hưởng ứng thì cũng chịu thua!", the parser captures only basic negation, while the ground truth explicitly encodes coordinated negations and conditions, offering richer semantic interpretation.

Differences in predicate argument and thematic role labeling also highlight parser limitations. Iterative or layered actions—such as "tôi nhớ lời anh chủ tịch xã Bùi Văn Luyến nhắc đi nhắc lại"—are merged or simplified in the parser output but elaborated in the ground truth through detailed encoding of predicates and roles, enhancing interpretability.

Quantification and numerical detail expressions diverge as well. Although quantities like "672 người" working abroad are recognized by the parser, they are not linked as precisely to verbs, timelines, and agents as in the ground truth. Expert annotations maintain tighter syntactic and semantic links between numbers, actions, and temporal contexts for accuracy.

Lastly, the parser tends to omit important temporal and role relations. For example, "có người đã là cô giáo dạy thcs" lacks clear tense and role articulation in the parse, which the ground truth annotations fully capture with temporal cues and predicate roles. Furthermore, the parser shows inconsistency in predicate labeling, whereas the ground truth applies systematic, structured predicates, promoting coherence and stronger semantic cohesion across sentences.

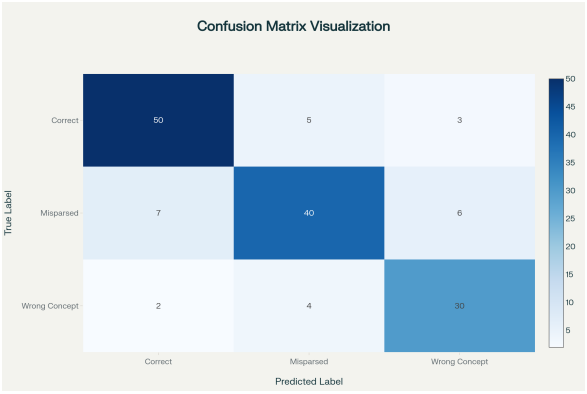

Figure 3: Confusion matrix of our proposed model ViSemCrew on VLSP 2025 ViSem Public Test

5 Conclusion and Future Work

This paper introduced ViSemCrew, a multi-agent workflow framework for Vietnamese semantic parsing that addresses the language's unique linguistic challenges through systematic decomposition and specialized agent coordination. Our framework organizes the complex parsing task into manageable subtasks handled by dedicated agents: the Linguistic Analysis Agent for morphological and syntactic processing, the Concept Extraction Agent for semantic concept identification, the Graph Construction Agent for relationship modeling, and the Validation Agent for quality assurance and error detection.

The key innovation of our approach lies in the structured orchestration of specialized agents through a controlled workflow with iterative refinement capabilities. Each agent focuses on specific aspects of semantic parsing, enabling targeted processing of Vietnamese linguistic phenomena such as analytic morphology, flexible word order, numeral classifiers, and implicit grammatical elements. The framework incorporates systematic validation mechanisms, targeted repair strategies, and robust fallback procedures to ensure reliable output quality.

Experimental evaluation on the VLSP 2025 ViSem Task demonstrated the effectiveness of our multi-agent approach in handling Vietnamese-specific challenges. The framework achieved competitive performance in Smatch scores, confirming its ability to process diverse Vietnamese text domains including news articles, literary texts, and

user-generated content. The systematic agent coordination and validation mechanisms proved particularly valuable in maintaining consistency and coverage across different text types.

In future work, we plan to expand the framework's capabilities by incorporating additional specialized agents for handling Vietnamese regional dialects and domain-specific terminology. We also aim to enhance the knowledge integration mechanisms by developing more comprehensive role databases and implementing adaptive learning capabilities that allow agents to improve their performance based on processing feedback. Furthermore, we intend to explore the application of this multi-agent paradigm to other low-resource languages and investigate integration with downstream applications such as question answering and dialogue systems. The modular agent architecture provides a flexible foundation for these extensions while maintaining the systematic processing advantages demonstrated in this work.

References

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Vinh-Hien D. Huynh, Chau-Anh Le, Chau M Truong, Y Thien Huynh, and Quy T Nguyen. 2024. Domain generalization in vietnamese dependency parsing: A novel benchmark and domain gap analysis. In *International Symposium on Information and Communication Technology*, pages 167–181. Springer.

Michael Wayne Goodman. 2020. Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online. Association for Computational Linguistics.

Zhouhong Gu, Xiaoxuan Zhu, Yin Cai, Hao Shen, Xingzhou Chen, Qingyi Wang, Jialin Li, Xiaoran Shi, Haoran Guo, Wenxuan Huang, and 1 others. 2025. Agentgroupchat-v2: Divide-and-conquer is what llm-based multi-agent system need. *arXiv preprint arXiv:2506.15451*.

Nguyen Kiet, Tran Son, Nguyen Luan, Huynh Tin, Luu Son, and Nguyen Ngan. 2022. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).

Hoa Quang Le, Huong Xuan Dieu Kieu, Khiem Vinh Tran, and Binh Thanh Nguyen. 2024. Lawvivqa: A visual question answering dataset for vietnamese legal content. In *2024 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 393–397. IEEE.

Ha Linh and Huyen Nguyen. 2019. A case study on meaning representation for Vietnamese. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 148–153, Florence, Italy. Association for Computational Linguistics.

Ha My Linh, Nguyen Thi Minh Huyen, Vu Xuan Luong, Nguyen Thi Luong, Phan Thi Hue, and Le Van Cuong. 2020. VLSP 2020 shared task: Universal Dependency parsing for Vietnamese. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, pages 77–83, Hanoi, Vietnam. Association for Computational Linguistics.

Son T Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. In *International Conference on Computational Collective Intelligence*, pages 546–558. Springer.

Dang Tuan Nguyen, Khoa Dang Nguyen, and Ha Thanh Le. 2013. Semantic parsing of simple sentences in unification-based vietnamese grammar. *International Journal on Natural Language Computing (IJNLC)*.

Duc-Vu Nguyen, Thang Chau Phan, Quoc-Nam Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. An attempt to develop a neural parser based on simplified head-driven phrase structure grammar on vietnamese. In *International Symposium on Information and Communication Technology*, pages 313–328. Springer.

Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

Le Hong Phuong, Pham Hoang, Pham Khoai, Nguyen Huyen, Nguyen Luong, and Nguyen Hiep. 2017. Vietnamese semantic role labelling. *VNU Journal of Science: Computer Science and Communication Engineering*, 33(2).

Team Qwen. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Luong Nguyen Thi, Linh Ha My, Hung Nguyen Viet, Huyen Nguyen Thi Minh, and Phuong Le Hong. 2013. Building a treebank for vietnamese dependency parsing. In *The 2013 RIVF International Conference on Computing and Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 147–151.

Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen. 2023. Bartphobeit: Pre-trained sequence-to-sequence and image transformers models for vietnamese visual question answering. In *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Khiem Vinh Tran, Hao Phu Phan, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen. 2024. Viclevr: A visual reasoning dataset and hybrid multimodal fusion model for visual question answering in vietnamese. *Multimedia Systems*, 30(4):199.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online. Association for Computational Linguistics.

A Appendix

A.1 Comparison Experiment with Open-Source LLM

To evaluate the adaptability of our framework beyond proprietary models, we conducted preliminary experiments using the open-source **Qwen3-8B-FP8** model (Qwen, 2025). The goal was to assess the parsing quality of a publicly available LLM under the same semantic parsing tasks designed for ViSemCrew.

Deployment and usage. We served Qwen3-8B-FP8 via an OpenAI-compatible endpoint with a non-thinking chat configuration (no tool-calls or chain-of-thought), standardizing prompts to elicit direct, text-only responses suitable for ViSemCrew agents. Inference settings enforced a constrained context (3,072 tokens), moderate sampling (temperature 0.7, top-p 0.8), and fixed concurrency (up to 12 sequences) across development and public test sets.

Experiments were run on a workstation equipped with an AMD Ryzen 7 7700 CPU, 64GB DDR5 RAM, and an NVIDIA RTX 4070 Ti Super GPU with 16GB VRAM, running Ubuntu 24.04 LTS.

A.1.1 Average Score

Figure 4 presents the average parsing score obtained by the Qwen3-8B-FP8 model across the evaluation dataset. Although performance was generally lower than Gemini 2.5 Pro, the model demonstrated consistent parsing capabilities, confirming the feasibility of adapting our pipeline to open-source LLMs.

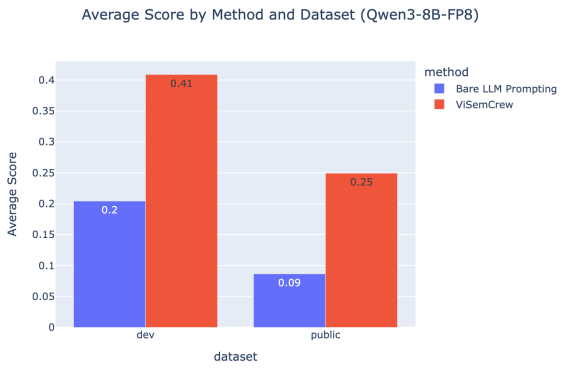

Figure 4: Average parsing score of Qwen3-8B-FP8 on the evaluation dataset.

A.1.2 Score Distribution

To provide further insight into performance variability, Figure 5 illustrates the distribution of parsing scores across the dataset. The distribution highlights that while a majority of sentences were parsed at moderate quality, outliers indicate specific failure cases where semantic relations were under-specified.

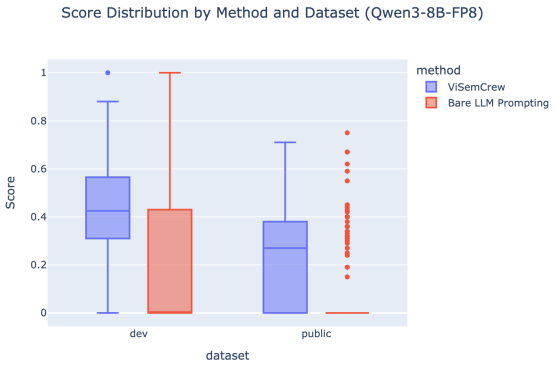

Figure 5: Distribution of parsing scores for Qwen3-8B-FP8 across the evaluation dataset.

These results suggest that while open-source LLMs are not yet competitive with state-of-the-art proprietary models for Vietnamese semantic parsing, they provide a promising foundation for research in resource-constrained or open-access environments.

A.2 Error Analysis with Open-Source LLM

To complement our main experiments, we performed a structured error analysis of the **ViSemCrew** Prusing the open-source Qwen3-8B-FP8 model on both the development and public test

sets. This analysis provides quantitative insights into common error types, reliability patterns, and runtime behavior across datasets.

A.2.1 Reliability Across Datasets

Figure 6 summarizes validation pass rates, repair frequencies, and fallback invocations. While validation succeeds for the majority of sentences in both datasets, the public test set exhibits a higher frequency of repair and fallback, suggesting greater structural variation and unseen constructions.

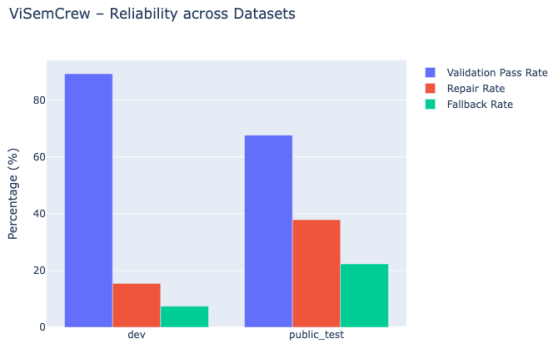

Figure 6: Reliability statistics of the ViSemCrew across dev and public test datasets. Success rate is omitted for clarity.

A.2.2 Concept and Relation Extraction

Figure 7 plots the number of extracted concepts against relations per sentence. Both datasets show a near-linear correlation, reflecting consistent graph-building behavior. However, sentences with high concept counts (>15) occasionally yield sparse relation coverage, pointing to difficulty in capturing long-distance dependencies.

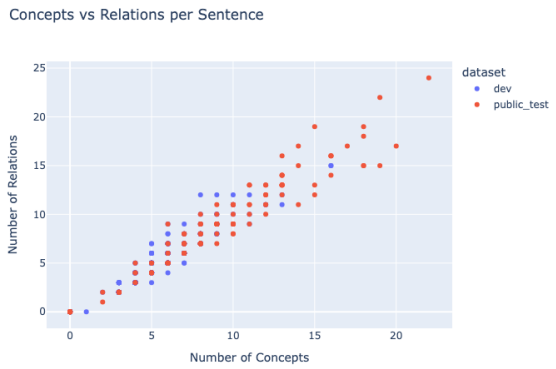

Figure 7: Number of concepts versus relations extracted per sentence across datasets.

A.2.3 Processing Time Distribution

Figure 8 illustrates the distribution of total processing times per sentence. While most dev examples complete under 5 seconds, the public test set includes longer tails, with several sentences requiring over 20 seconds due to multiple repair and fallback iterations.

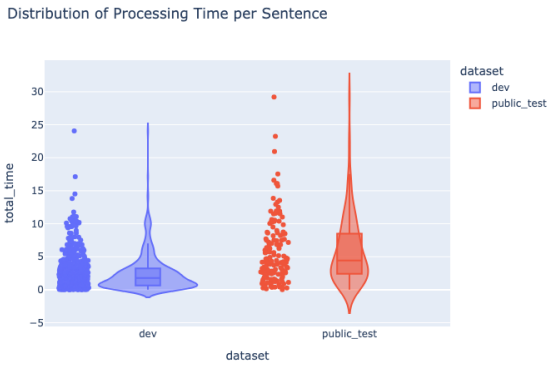

Figure 8: Distribution of sentence-level processing time across datasets.

A.2.4 Error Types

From both quantitative metrics and qualitative inspection, the most common error categories are:

- **Semantic Role Errors:** Incorrect or missing role assignment, especially for implicit arguments and passive voice.
- **Named Entities:** Inconsistent handling of multi-word entities and foreign names due to limited lexical coverage.
- **Temporal Expressions:** Frequent omissions or underspecified temporal modifiers.
- **Graph Connectivity:** Occasional disconnected subgraphs, often repaired by fallback generation.

These findings suggest that the pipeline is most challenged by phenomena requiring world knowledge or broader discourse context. Potential improvements include (i) expanding the role reference database, (ii) integrating external NER modules, and (iii) enhancing temporal normalization rules.

A.3 Input and output specifications for each agent

We illustrate the detailed input, intermediate state, and output of each agent in the ViSemCrew workflow using the example sentence:

"lớp 12a được nhà trường khen"
(Translation: *"Class 12A was praised by the school"*)

A.3.1 Initial State

original_sentence: "lớp 12a được nhà trường khen ."

A.3.2 Linguistic Analysis Agent

Input: Original sentence.
Processing: POS tagging, dependency parsing, and passive voice detection.
Output:

linguistic_analysis:
features:
- word: "lớp", pos: NOUN, head: "khen", rel: nsubj:pass
- word: "12a", pos: PROPN, head: "lớp", rel: flat:name
- word: "được", pos: AUX, head: "khen", rel: aux:pass
- word: "nhà", pos: NOUN, head: "trường", rel: compound
- word: "trường", pos: NOUN, head: "khen", rel: obl:agent
- word: "khen", pos: VERB, head: "khen", rel: root

A.3.3 Concept Extraction Agent

Input: Sentence + linguistic features.
Output:

concepts:
- k: "khen" → The action of praising
- l: "lớp" → The concept of a 'class'
- n: "nhà trường" → The entity 'the school'
- a: "12a" → Identifier for the class
- d: "được" → Passive marker modality

A.3.4 Graph Construction Agent - Root Identification

Input: Concepts + dependency tree.
Output:

graph_root_variable: k

A.3.5 Graph Construction Agent - Relation Extraction

Input: Concepts + root variable.
Output:

relations:
- (k → :beneficiary-arg1 → l)
- (k → :agent → n)
- (k → :modality → d)
- (l → :mod → a)

A.3.6 Graph Construction Agent - Graph Assembly

Input: Lists of concepts and relations.
Output:

graph_triples:
(k, instance, khen)
(l, instance, lớp)
(n, instance, nhà trường)
(a, instance, 12a)
(d, instance, được)
(k, :beneficiary-arg1, l)
(k, :agent, n)
(k, :modality, d)
(l, :mod, a)

A.3.7 Validation Agent

Input: Graph Triples from Graph Construction Agent.
Processing: Checking word coverage and graph connectivity.
Output:

connectivity_checker_decision: proceed
missing_words: []

A.3.8 Penman String Generation

Input: Graph Triples from Graph Construction Agent.
Output:

(k / khen
:beneficiary-arg1 (l / lớp
:mod (a / 12a))
:agent (n / nhà trường)
:modality (d / được))

A.3.9 AMR Format Validation

Input: Generated Penman string.
Output:

amr_validation_passed: true

A.3.10 Final AMR Output

(k / khen
:beneficiary-arg1 (l / lớp
:mod (a / 12a))
:agent (n / nhà trường)
:modality (d / được))

Interpretation:

- Root concept: *khen* (praise)
- Beneficiary: *lớp 12a* (Class 12A)
- Agent: *nhà trường* (the school)
- Modality: *được* (passive marker)

A.3.11 Alternative Flows

If missing concepts, disconnected graphs, or validation failures occur, the Repair Agent or fallback generation mechanisms are triggered:

- **Missing words:** New concepts and relations are generated.
- **Disconnected nodes:** Relations are re-extracted with repair feedback.
- **Validation failed:** Complete AMR regeneration using few-shot examples.