SYNAPSE: Synthesizing Narratives from Agentic Path Spatial Exploration

Prateek Gupta[1], Daniel Antunes Pedrozo[1], Aleksandr Koviazin[1], Jorge Augusto Meira[1], and Antonio Ken Iannillo[1]

SnT, University of Luxembourg, 29 Av. John F. Kennedy, 1855 Luxembourg City, Luxembourg {prateek.gupta,daniel.antunes,aleksandr.koviazin,jorge.meira, antonioken.iannillo}@uni.lu

Abstract. We introduce Synapse, a modular, privacy-preserving framework for generating multilingual summaries of logistics routes, a task distinct from traditional routing optimization or forecasting. Our approach transforms static route definitions by synthesizing them with a variety of signals, including pre-computed failure predictions, real-time traffic, and weather, through a network of specialized AI agents, orchestrated using the FastMCP (Multi-Agent Communication Protocol) framework. To ensure GDPR compliance, we apply random perturbation to GPS coordinates before querying external APIs, a method that obfuscates precise locations while preserving local context. The system simulates the route's temporal progression via iterative calls to a traffic API, enabling time-aware queries for other agents gathering weather and urban context data. A generative agent fuses these multimodal signals into a core narrative. Parallel translation agents then render that narrative in multiple languages. This work presents a scalable pipeline for interpretable logistics intelligence, introducing an agentic architecture and a human-validated, multilingual evaluation benchmark.

Keywords: Multimodal Machine Learning · Privacy-Preserving Logistics · Route Summarization · Agent-Based Systems · Large Language Models (LLMs)

1 Introduction

The planning and execution of logistics routes in heterogeneous regions, such as Europe, involves navigating two persistent challenges: the synthesis of high-volume, multimodal data streams and the need for clear communication across multilingual teams. While substantial research has advanced route optimization planning, a complementary challenge is the automated generation of concise, multilingual narratives for real-time operational awareness. Such summaries can help human decision-makers adapt to dynamic conditions, a task that remains difficult to scale. The complexity of this task is enhanced by two primary obstacles. First, route data contains sensitive geolocations protected by frameworks like GDPR. This privacy constraint renders monolithic approaches impractical for real-world deployment, as they would require transmitting all raw, sensitive data to a single, often third-party, large language model. Second, the architectural challenge of reasoning across diverse data domains (e.g., service failure predictions, traffic, weather, workload) remains significant. A traditional approach would require a human planner or a monolithic application to manually query multiple REST APIs, each with its own unique payload structure, authentication, and data format. Our agentic framework is designed to abstract this complexity. It allows users to make requests in plain English, delegating the intricate task of data retrieval and synthesis to a network of specialized agents. This not only makes the system more scalable and easier to maintain but also democratizes access to complex logistics intelligence by removing the need for specialized API knowledge. In this work, we explore this problem space by introducing Synapse, a modular agentic framework built on the FastMCP library designed to navigate these constraints. To address privacy constraints, the framework applies random perturbation to GPS coordinates before they are sent to external services. The system's architecture uses specialized agents to reason over distinct data modalities; for instance, a traffic agent projects a route's temporal progression to enable time-aware queries for weather and urban context. These signals are then synthesized into a base narrative by a generative agent and localized by parallel translation agents. The contributions of this work are therefore twofold: (1) Synapse, a novel and extensible agentic architecture for multilingual narrative generation, and (2) a human-validated multilingual evaluation benchmark, which includes the challenging low-resource Luxembourgish language. We release both our framework and benchmark publicly to provide a foundation for future research and development in this new domain.[1]

1.1 Related Work

Logistics operations present numerous challenges stemming from the complexity and volume of multimodal data streams. Planners and drivers must continuously interpret inputs such as traffic conditions, weather forecasts, and delivery constraints to make time-sensitive decisions. However, raw streams of GPS coordinates and sensor signals often lack the semantic clarity required for intuitive, proactive risk assessment. This disconnect between low-level data and high-level situational awareness has positioned the automated summarization of geospatial data as a key research direction. Early work in this area explored the segmentation of GPS trajectories into semantically meaningful units, followed by the generation of textual summaries to highlight salient events along a route [15]. This concept was extended to summarizing network-level activity by identifying representative paths, such as the K-Main Routes approach used to cluster crime reports or traffic flows [13]. Subsequent efforts aimed to improve the human-centric utility of these summaries through personalization and multimodality; frameworks like PerNav and PaRE generated more intuitive directions from his-

[1] The evaluation benchmark and script are available at: https://github.com/Synapse-INTSYS2025/agentic-eval-benchmark

torical data [12],[11], while systems such as TREADS and Scenemash demonstrated the value of enriching summaries with real-time social media and image data [5],[1],[8]. While these foundational systems demonstrated the value of multimodal data integration, they predate the advanced cross-domain reasoning and synthesis capabilities of modern Large Language Models (LLMs), which have enabled a new paradigm for solving complex, multi-step problems through agentic workflows [4]. Advanced reasoning techniques like Chain-of-Thought (CoT) and its variants allows LLMs to function as orchestrators that can decompose problems and interact with external tools [3], [19], [17]. This has spurred the development of multi-agent systems built upon standardized communication frameworks like the Model Context Protocol (MCP), which provides a unified interface for model-tool interaction [7], [9], [16]. Applying these advanced systems to multilingual contexts, however, introduces significant challenges, particularly for low-resource languages, as comprehensive surveys of Neural Machine Translation (NMT) establish that model performance remains suboptimal for language pairs with scarce parallel data [14]. This limitation extends directly to generative tasks like summarization [10], with the difficulty being particularly acute for languages such as Luxembourgish, where even foundational NLP tasks require specialized models and human-in-the-loop validation to achieve acceptable performance [6]. Furthermore, the linguistic inequality in training data can degrade the effectiveness of safety mechanisms in LLMs, making robust evaluation in low-resource settings a critical area of study [18], [2].

2 System Architecture and Model

2.1 Overview

We propose Synapse, a modular agentic framework designed for the automated generation of multilingual, human-centered summaries for logistics operations. The framework transforms static route definitions and dynamic, real-time signals into natural language narratives through a network of specialized AI agents. Our architecture explicitly models the distinct domains of logistics data, with individual agents handling real-time traffic, weather, workload assessment, and urban context. A core component of our approach is a privacy-preserving mechanism that applies random perturbation to all GPS coordinates before they are sent to external services, ensuring GDPR compliance. Furthermore, this modular design offers a significant scalability advantage: users can interact with the entire system through a single, natural language endpoint, completely obviating the need to understand or construct the complex, individual REST API payloads required by the downstream specialist agents. While the underlying MCP framework supports dynamic tool discovery, our implementation uses a static registration model. This is a deliberate design choice to enhance security and privacy; by manually registering a limited, necessary set of tools for each agent, we prevent the potential for the underlying LLM to discover and misuse other available tools in unintended ways.

2.2 Input Representations

The system ingests an unstructured natural language query, such as the example shown in Listing 1.1, via a JSON API endpoint. A Natural Language Understanding (NLU) module, powered by an LLM, parses this query into a structured `tool_call`. This decouples the ambiguity of natural language from the deterministic execution required by the downstream agentic workflow.

```
curl -X POST http://localhost:8088/natural_query \
     -H "Content-Type: application/json" \
     -d '{"query": "Briefing for Route-001 tomorrow at 2 PM."}'
```

Listing 1.1: Example user query via `curl`.

2.3 Workflow

Our system's architecture is a multi-agent framework designed to process a natural language query and generate a multilingual briefing, as illustrated in Figure 1. The system comprises a user-facing API Gateway, a central `Pipeline Agent` orchestrator, and a suite of specialist agents for data retrieval and synthesis.

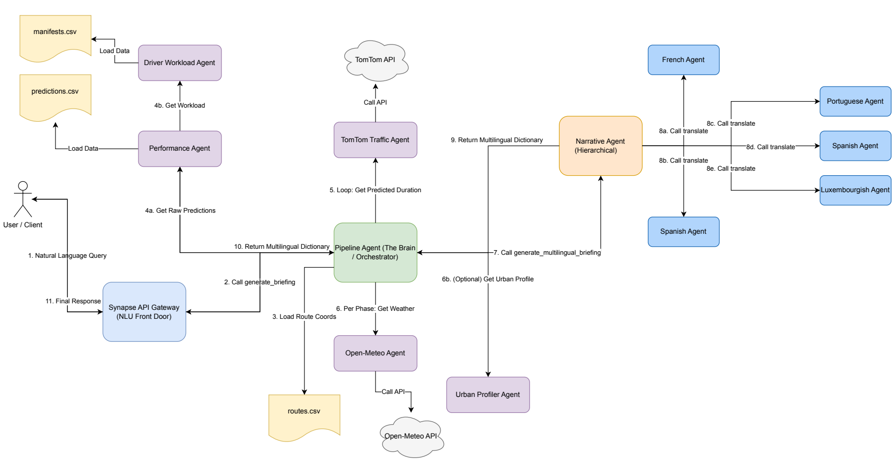

Fig. 1: Architecture and execution flow of the Synapse multi-agent system.

1. **Query Ingestion and Orchestration.** The workflow is initiated when a user's natural language request is ingested by the `Synapse API Gateway`, which performs initial NLU parsing. The gateway then invokes the central `Pipeline Agent`, a FastMCP service orchestrator, to begin the main workflow.
2. **Baseline Data and Prediction Loading.** The orchestrator begins by gathering the foundational data for the route. It calls the internal `Performance Agent`, which provides a set of pre-computed failure predictions for each stop

based on a model using historical data. Concurrently, it loads the route's static geospatial coordinates and manifest details. This initial step provides the system with a complete baseline of the planned route and its historically-informed risks before querying for live, dynamic signals.

3. **Dynamic Agent Interrogation.** The orchestrator enriches the static data by querying its specialist agents. It first projects the route's temporal progression by sequentially calling the `TomTom Traffic Agent` for leg-by-leg travel times. This simulation assigns each stop to an operational phase (e.g., Morning, Afternoon), enabling phase-specific queries. A geospatial centroid is calculated for the stops in each phase to serve as a representative point for a targeted `Open-Meteo Agent` weather forecast, assuming minimal intra-phase weather variation. Concurrently, the `Driver Workload Agent` provides manifest details. This culminates in a multi-factor risk assessment: if a phase combines adverse weather with a challenging load (heavy/oversized items), the `Urban Profiler Agent` is invoked to provide proactive parking guidance, mitigating potential delays from increased manual handling in areas with limited parking.

4. **Hierarchical Synthesis and Multilingual Generation.** With the complete data context assembled, the orchestrator invokes the `Narrative Agent`. This agent exemplifies a key architectural pattern in our framework, acting as both a service provider and a service consumer. First, in its server role, it receives the data and synthesizes all information into a coherent base summary in English. Then, acting as a client, it broadcasts this English summary concurrently via FastMCP to five single-language translation agents (e.g., `French Agent`, `Spanish Agent`) for localization. This hierarchical delegation allows the primary orchestrator to remain simple while enabling specialized sub-tasks like translation to be managed by the appropriate agent.

5. **Response Aggregation.** Finally, the `Pipeline Agent` collects the outputs from all translation agents into a single multilingual dictionary, which is returned to the user through the API Gateway.

The final aggregated response is a JSON object containing the synthesized narrative in English and its translations. An example of the `Luxembourgish` output, which is generated by the `Narrative Agent` to match the golden standard format, is shown in Listing 1.2.

```
D'Route-003 huet 10 Arrêten mat 1 virausgesoten Ausfall. D'
    Haapterausfuerderunge sinn Appartement-Arrêten: 2, Schwéier
    Artikelen: 1. --- Phasendetailer: - Moies: D'Wieder wäert kloer sinn
     mat Temperaturen ëm 27.9°C a mëttelméissegem Wand (39-43 km/h) an
    den éischten dräi Stonnen. 1 Ausfäll virausgesot op 7 Arrêten an dë
    ser Phas.
```

Listing 1.2: Example of the final Luxembourgish narrative output

3 Experimental Setup

Experiments were conducted on a server with an NVIDIA RTX A6000 GPU (48 GB VRAM) running Ubuntu 24.04.2 LTS. The entire multi-agent system was containerized using Docker to ensure reproducibility. All generative agents were powered by local language models served via Ollama v0.9.6, and agent communication was managed by the FastMCP framework. To ensure temporal consistency for the evaluation, all queries to time-sensitive external APIs (e.g., traffic, weather) were anchored to a static date of **May 15, 2025**, although the framework is designed for live data streams.

3.1 Hyperparameters and Prompts

All generative tasks were performed using large language models served via the Ollama v0.9.6 framework. We evaluated two primary models: `Mistral-7B Instruct` and `Llama 3.2`. To ensure a reproducible baseline, we used the official default inference parameters for both models. For all experimental runs, the `seed` was not fixed to a static value.

System Prompts. Each generative agent was assigned a distinct system prompt tailored to its specialized function. The key prompts are summarized below:

- **Narrative Agent:** Instructed to act as an expert "Senior Logistics Editor," this agent's sole task is to synthesize the incoming JSON data into a single, factually perfect string. The prompt enforces a strict output template ('Overall, [route name] has...') and provides explicit step-by-step instructions for calculating totals and formatting the phase details. This ensures the output matches the golden standard format used for evaluation.

- **Translation Agents:** Given a simple, direct instruction to "Translate the following English logistics briefing into professional, formal [Language]," followed by the English text. The prompt explicitly commands the agent to respond ONLY with the translated text to ensure a clean output for the final response aggregation.

- **Weather Agent:** Prompted to act as a specialized meteorologist, this agent is provided with a suite of tools for accessing different types of weather data (e.g., 'get forecast 3 hours', 'get daily summary', 'get forecast 6 hours'). Its primary task is to perform LLM-driven tool selection based on the specific temporal and geospatial context provided by the orchestrator.

- **API Gateway:** This agent is prompted to act as an "API routing bot" with the sole function of parsing an unstructured user query. The prompt provides a clear example of both the input query and the required structured JSON output ('"tool": "generate briefing", "parameters": ...'), forcing the LLM to act as a reliable Natural Language Understanding (NLU) component.

3.2 Baselines

We evaluate the performance of our Synapse framework by comparing two distinct large language models serving as the core generative engines for all agents:

- **Mistral-7B:** A widely-used, high-performance 7-billion parameter model, selected as our primary baseline due to its strong performance and the clear visibility of its default inference parameters, which facilitates a reproducible experimental setup. We use the `mistral:7b-instruct-q4_K_M` variant.
- **Llama 3.2:** A recent, state-of-the-art model used for comparison against the established baseline. We use the `llama3.2:8b-instruct-q4_K_M` variant.

A third potential baseline, a monolithic approach using a single powerful external LLM (e.g., GPT-4), was considered but explicitly excluded. Such an approach is fundamentally incompatible with the GDPR constraints central to our problem formulation, as transmitting precise geolocations to a third-party service would constitute a privacy violation. Our agentic architecture is specifically designed to mitigate this risk, and our evaluation therefore focuses on comparing different LLMs within our privacy-preserving framework.

4 Evaluation

To assess the performance of our framework, we conducted a quantitative analysis of the multilingual translation quality produced by our agentic pipeline.

4.1 Evaluation Setup

Dataset. Our evaluation benchmark consists of 100 semi-synthetic, 10-stop logistics routes. To ensure linguistic and geographic diversity, stop locations were sampled from regions across Luxembourg, France, and Belgium, corresponding to the target languages of the system. For each route, a canonical English summary was first manually authored by the researchers. This text then served as the source for creating and validating with the help of native speakers, the golden reference translations for our five target languages: French, Spanish, Italian, Portuguese, and Luxembourgish.

Metrics. To provide a comprehensive assessment of translation quality, we employed three standard automated metrics, each capturing a different aspect of performance.

COMET. We use the `Unbabel/wmt23-cometkiwi-da-xl` model, a neural regression metric that estimates translation quality based on semantic similarity to the source and reference texts. It is trained on direct human judgments, and higher scores indicate better performance. As a neural model, it leverages contextual embeddings rather than relying on explicit word overlap.

METEOR. This metric computes a score based on the harmonic mean of unigram precision (P) and recall (R), with a penalty for fragmentation. The score is calculated as:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \quad \text{where} \quad F_{\text{mean}} = \frac{10PR}{9P + R}$$

The penalty term increases with the number of contiguous chunks of matched words, rewarding longer, more coherent matches.

TER (Translation Edit Rate). TER is an error metric that measures the minimum number of edits required to transform a system hypothesis into the reference translation. It is defined as:

$$\text{TER} = \frac{\text{Number of edits}}{\text{Number of words in reference}}$$

Edits include insertions, deletions, substitutions, and shifts. Lower scores indicate better translation quality.

5 Results and Discussion

The performance of the Llama 3.2 and Mistral-7B models within the SYNAPSE framework is presented in Table 1. A horizontal analysis of these results comparing a model's performance across tasks reveals a critical finding: a significant and consistent degradation in quality occurs when the synthesized English narrative is passed to downstream translation agents. This suggests a systemic challenge in maintaining quality across a multi-agent generative pipeline. With the superior

Table 1: Comparative performance of Llama 3.2 (L3.2) and Mistral-7B (M7B). COMET scores are reported as mean (std. dev.). Best results are in bold.

	COMET ↑		**METEOR ↑**		**TER ↓**	
Language	**L3.2**	**M7B**	**L3.2**	**M7B**	**L3.2**	**M7B**
English (en)	**0.34** (0.03)	0.25 (0.04)	**0.70**	0.39	**43.30**	64.51
Spanish (es)	**0.46** (0.08)	0.38 (0.06)	**0.48**	0.30	**65.43**	75.68
French (fr)	**0.40** (0.07)	0.30 (0.05)	**0.48**	0.30	**68.56**	74.75
Italian (it)	**0.41** (0.08)	0.31 (0.06)	**0.42**	0.24	**78.68**	82.90
Portuguese (pt)	**0.43** (0.09)	0.37 (0.07)	**0.49**	0.30	**65.45**	76.09
Luxembourgish (lb)	0.22 (0.06)	**0.26** (0.07)	**0.20**	0.08	**93.37**	97.80

Llama 3.2 model, the METEOR score drops from a high of 0.70 on the English synthesis task to an average of 0.47 across the four high-resource languages. This degradation is even more pronounced in the Translation Edit Rate (TER), which increases from 43.30 for the English output to as high as 78.68 for the Italian translation. For a logistics planner who must coordinate multilingual teams, such a high error rate transforms the system from a tool for clarity into a source of

operational confusion, risking a loss of user trust and eventual abandonment of the system.

This finding validates the necessity of a modular agentic architecture. The data clearly shows that a single, general-purpose LLM is insufficient for reliably executing all sub-tasks in this specialized pipeline. The poor performance of the translation components highlights the need for a system where underperforming agents can be individually fine-tuned or replaced.

6 Conclusion and Future Work

We introduced SYNAPSE, a modular, privacy-preserving agentic framework for generating multilingual logistics summaries. Our system demonstrates a scalable approach to synthesizing real-time multimodal data, including traffic, weather, and pre-computed failure predictions, into a coherent, human-centered narrative while ensuring GDPR compliance through coordinate perturbation. The quality of the final output, however, remains dependent on the capabilities of the underlying LLMs, and the microservices architecture presents a trade-off between modularity and latency.

Future work will focus on two primary directions. First, we plan to fine-tune the generative agents on domain-specific logistics texts to improve narrative quality and mitigate factual inconsistencies. Second, we will investigate agent optimization techniques and the integration of additional data streams, such as vehicle telematics, to enhance the system's real-time contextual awareness and overall robustness.

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments

We wish to express our sincere gratitude to our colleagues at the SnT, University of Luxembourg. We are particularly indebted to our fellow researchers for their meticulous work in constructing our evaluation benchmark. Their efforts included curating the foundational logistics route data and, crucially, providing native-speaker validation for each of the target language translations. This research was greatly strengthened by their collaborative support and linguistic expertise.

References

1. van den Berg, J., Rudinac, S., Worring, M.: Scenemash: Multimodal Route Summarization for City Exploration. In: Ferro, N., Crestani, F., Moens, M.F., Mothe,

J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. pp. 833–836. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1$_7$5

2. Buscemi, A., Lothritz, C., Morales, S., Gomez-Vazquez, M., Clarisó, R., Cabot, J., Castignani, G.: Mind the Language Gap: Automated and Augmented Evaluation of Bias in LLMs for High- and Low-Resource Languages (Apr 2025). https://doi.org/10.48550/arXiv.2504.18560, http://arxiv.org/abs/2504.18560, arXiv:2504.18560 [cs]
3. Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., Che, W.: Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models (Apr 2025). https://doi.org/10.48550/arXiv.2503.09567, http://arxiv.org/abs/2503.09567, arXiv:2503.09567 [cs]
4. Chen, Q., Qin, L., Wang, J., Zhou, J., Che, W.: Unlocking the Capabilities of Thought: A Reasoning Boundary Framework to Quantify and Optimize Chain-of-Thought. Advances in Neural Information Processing Systems **37**, 54872–54904 (Dec 2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/62ab1c2cb4b03e717005479efb211841-Abstract-Conference.html
5. Fu, K., Lu, Y.C., Lu, C.T.: TREADS: a safe route recommender using social media mining and text summarization. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 557–560. SIGSPATIAL '14, Association for Computing Machinery, New York, NY, USA (Nov 2014). https://doi.org/10.1145/2666310.2666368, https://doi.org/10.1145/2666310.2666368
6. Hosseini-Kivanani, N., Kühn, J., Schommer, C.: Mapping Sentiments: A Journey into Low-Resource Luxembourgish Analysis. In: Sousa-Silva, R., Lopes Cardoso, H., Koponen, M., Pareja Lora, A., Seresi, M. (eds.) Proceedings of the First LUHME Workshop. pp. 20–27. CLUP, Centro de Linguística da Universidade do Porto FLUP - Faculdade de Letras da Universidade do Porto, Santiago de Compostela, Spain (Oct 2024), https://aclanthology.org/2024.luhme-1.3/
7. Hou, X., Zhao, Y., Wang, S., Wang, H.: Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions (Apr 2025). https://doi.org/10.48550/arXiv.2503.23278, http://arxiv.org/abs/2503.23278, arXiv:2503.23278 [cs]
8. Kanaya, Y., Kawanaka, S., Suwa, H., Arakawa, Y., Yasumoto, K.: Automatic Route Video Summarization based on Image Analysis for Intuitive Touristic Experience. Sensors and Materials **32**(2), 599 (Feb 2020). https://doi.org/10.18494/SAM.2020.2616, http://myukk.org/SM2017/article.php?ss=2616
9. Karimova, S., Dadashova, U.: The model context protocol: a standardization analysis for application integration. Journal of Computer Science and Digital Technologies **1**(1), 50–59 (Jun 2025), http://journals.unec.edu.az/index.php/jcsdt/article/view/27
10. Li, J., Chen, J., Chen, H., Zhao, D., Yan, R.: Multilingual Generation in Abstractive Summarization: A Comparative Study. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evalu-

ation (LREC-COLING 2024). pp. 11827–11837. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.1032/

11. Li, Y., Su, H., Demiryurek, U., Zheng, B., He, T., Shahabi, C.: PaRE: A System for Personalized Route Guidance. In: Proceedings of the 26th International Conference on World Wide Web. pp. 637–646. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (Apr 2017). https://doi.org/10.1145/3038912.3052717, https://doi.org/10.1145/3038912.3052717
12. Li, Y., Su, H., Demiryurek, U., Zheng, B., Zeng, K., Shahabi, C.: PerNav: A Route Summarization Framework for Personalized Navigation. In: Proceedings of the 2016 International Conference on Management of Data. pp. 2125–2128. SIGMOD '16, Association for Computing Machinery, New York, NY, USA (Jun 2016). https://doi.org/10.1145/2882903.2899384, https://doi.org/10.1145/2882903.2899384
13. Oliver, D., Bannur, A., Kang, J.M., Shekhar, S., Bousselaire, R.: A K-Main Routes Approach to Spatial Network Activity Summarization: A Summary of Results. In: 2010 IEEE International Conference on Data Mining Workshops. pp. 265–272 (Dec 2010). https://doi.org/10.1109/ICDMW.2010.156, https://ieeexplore.ieee.org/abstract/document/5693309, iSSN: 2375-9259
14. Ranathunga, S., Lee, E.S.A., Prifti Skenduli, M., Shekhar, R., Alam, M., Kaur, R.: Neural Machine Translation for Low-resource Languages: A Survey. ACM Comput. Surv. **55**(11), 229:1–229:37 (Feb 2023). https://doi.org/10.1145/3567592, https://doi.org/10.1145/3567592
15. Su, H., Zheng, K., Zeng, K., Huang, J., Sadiq, S., Yuan, N.J., Zhou, X.: Making sense of trajectory data: A partition-and-summarization approach. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 963–974 (Apr 2015). https://doi.org/10.1109/ICDE.2015.7113348, https://ieeexplore.ieee.org/abstract/document/7113348, iSSN: 2375-026X
16. Szeider, S.: MCP-Solver: Integrating Language Models with Constraint Programming Systems (Apr 2025). https://doi.org/10.48550/arXiv.2501.00539, http://arxiv.org/abs/2501.00539, arXiv:2501.00539 [cs]
17. Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., Fei, H.: Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey (Mar 2025). https://doi.org/10.48550/arXiv.2503.12605, http://arxiv.org/abs/2503.12605, arXiv:2503.12605 [cs]
18. Yong, Z.X., Menghini, C., Bach, S.H.: Low-Resource Languages Jailbreak GPT-4 (Jan 2024). https://doi.org/10.48550/arXiv.2310.02446, http://arxiv.org/abs/2310.02446, arXiv:2310.02446 [cs]
19. Zhang, X., Du, C., Pang, T., Liu, Q., Gao, W., Lin, M.: Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs. Advances in Neural Information Processing Systems **37**, 333–356 (Dec 2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/00d80722b756de0166523a87805dd00f-Abstract-Conference.html