Revisiting LLMs in Planning from Literature Review: a Semi-Automated Analysis Approach and Evolving Categories Representing Shifting Perspectives

Vishal Pallagani[1], **Nitin Gupta**[1], **Bharath Muppasani**[1], **Biplav Srivastava**[1]

[1] University of South Carolina

Abstract

Tracking the rapidly evolving literature at the intersection of large language models (LLMs) and planning has become increasingly complex due to significant growth in research output and shifting thematic focuses. Building on the survey by Pallagani et al. (2024), which organized 126 papers collected till November 2023 into eight categories, we present a platform that automates the extraction, categorization, and trend analysis of new papers. Our analysis reports on category drift, identifying evolving perspectives on the use of LLMs for planning. Our analysis reveals a decline in the percentage of papers for six categories, an increase in two, and the emergence of two new categories. Specifically, we contribute by (1) developing an automated system for categorizing new papers into existing or emergent categories, (2) reporting on category shifts with the addition of 47 new papers till September 2024, and (3) introducing a platform for continuous extraction, categorization, and trend tracking in LLM and planning research. This platform also features a leaderboard to encourage innovations in automated paper categorization.

Introduction

The field of LLMs in planning is experiencing rapid growth, evolving well beyond its initial foundations. In recent years, research at the intersection of LLMs and planning has expanded both in volume and complexity, prompting a need to reconsider existing frameworks for organizing and understanding this literature. Early surveys, such as the one by Pallagani et al. (2024), categorized over a hundred studies across eight established categories, laying a foundational taxonomy that enabled researchers to contextualize and build upon prior work. However, with the pace of advancements and the emergence of diverse approaches, new categories are emerging that reflect shifting perspectives on the role of LLMs in planning. These emerging categories, which extend beyond the eight defined in the previous survey, underscore the limitations of traditional, static taxonomies in capturing the nuances of this rapidly evolving domain. This progression highlights the need for a more flexible, responsive structure capable of accommodating new developments as they arise.

In this position paper, we address these evolving dynamics by contributing: (a) **automation in the categorization process**, introducing an automated tool that leverages natural language processing techniques to streamline the extraction and categorization of relevant papers, providing a scalable solution for dynamically updating taxonomies in this fast-evolving domain; (b) an **updated categorization of literature in LLM and planning**, re-evaluating the existing taxonomy to accommodate the growing breadth of research and identifying emerging themes and trends in the deployment of LLMs for planning tasks; and (c) **a testbed to support future research**, including baseline tools, submission guidelines, and visualization mechanisms. This framework allows other researchers to benchmark and enhance automated categorization methods using the provided baseline data and leaderboard.

Our work underscores the importance of a responsive, adaptive approach to literature organization, particularly in the context of LLMs and planning, where research progress is both rapid and multifaceted. By documenting the shifts within existing categories and introducing automation, we aim to set a new standard for tracking academic trends in a way that is both structured and capable of evolving alongside the field.

Automation of Category Identification

Let $C = \{c_0, c_1, c_2, \ldots, c_n\}$ be a set of category labels, where each label $c_i \in C$ represents a distinct research area or topic, such as *Plan Generation* or *Language Translation*. A **taxonomy** T is defined as a subset of these labels that represents a comprehensive classification structure for a particular field. This taxonomy might be derived by analyzing a set of documents $D = \{p_1, p_2, \ldots, p_k\}$, such as research papers, relevant to a research domain. For example, a taxonomy could be {Brain-inspired Planning, Heuristic Optimization, Interactive Planning, Language Translation, Model Construction, Multi-agent Planning, Plan Generation, Tool Integration}, as reported in Pallagani et al. (2024), where D consists of 126 documents.

In this automated classification system, each document in a new set of papers $D' = \{p'_1, p'_2, \ldots, p'_m\}$ is categorized based on relevance to an existing taxonomy T or an *uncategorized* option if no suitable match exists. To categorize each new paper $p' \in D'$, we define a **goodness metric**

$g(p', c_i)$, which evaluates the relevance of p' to each category $c_i \in C \cup \{c_0\}$, where c_0 represents the *uncategorized* category. The **single label** categorization problem assigns each paper p' to the category c_i^* with the highest relevance score, provided it exceeds a threshold τ:

$$c_i^* = \underset{c_i \in C \cup \{c_0\}}{\arg\max}\ g(p', c_i)$$

If $g(p', c_i) < \tau$ for all $c_i \in C$, the paper p' is assigned to c_0, as it does not sufficiently match any predefined category.

The **multi-label** categorization problem assigns each document to one or more categories that reflect the reality that research papers often cover multiple, overlapping topics. In this setup, each paper $p' \in D'$ is assigned to a subset of categories $C' \subseteq C$ where the goodness metric $g(p', c_i)$ exceeds the threshold τ:

$$C' = \{c_i \in C \mid g(p', c_i) \geq \tau\}$$

If C' is empty, the paper is assigned to the uncategorized category c_0. After each categorization cycle, occurring at a defined time interval T_{period}, the taxonomy T is updated to reflect the inclusion of newly categorized papers, forming an updated taxonomy T' as follows:

$$T' = T \cup \{(p', c_i^*) \mid p' \in D', c_i^* = \underset{c_i \in C \cup \{c_0\}}{\arg\max}\ g(p', c_i)\}$$

This evolving taxonomy model thus continuously integrates new research data, adapting to emerging topics and ensuring that the classification system remains relevant and current.

Experimental Setup

We aim to identify the most effective method for accurately categorizing research papers by applying the single- and multi- label setups across different machine-learning models. Additionally, we implemented a **human-augmented** method that combines human expertise with the initial categorization provided by the best-performing automated classification model, further refining the categorization results.

To support this goal, we developed an *Automated Paper Extraction Tool* that continuously retrieves and filters papers from the ArXiv scholarly database. The system, implemented with T_{period} = 14 days, queries the ArXiv database for recent publications in the fields of LLMs and planning. Papers are filtered using predefined keywords and stored in a database for deduplication and categorization. This pipeline ensures continuous, up-to-date retrieval of relevant research documents, resulting in an evolving dataset D' ready for categorization.

Dataset: The dataset includes a labeled set of previously categorized papers D_1 (126 papers) obtained from Pallagani et al. (2024) and a new set D_2 (47 papers), which is extracted using the automated extraction tool that requires categorization. Titles and abstracts are combined to create a comprehensive representation and features are extracted using *TF-IDF vectorization*, with a maximum of 5,000 features and stop word filtering on the NLTK's (Bird, Klein, and Loper 2009) `English` word set.

Classification Models: We implemented two main approaches that chose some of the state-of-the-art classification models (Olson et al. 2018):

- **Traditional ML**: Decision Trees (Wu et al. 2008), Random Forests (Ho 1995), and Support Vector Machines (SVM) (Cortes and Vapnik 1995), adapted for multi-label classification using a One-vs-Rest (OVR) approach, as implemented in *scikit-learn* (Pedregosa et al. 2011), effectively adapting these classifiers for both single-label and multi-label setups.
- **Transformer-based**: BERT (Kenton and Toutanova 2019) and SciBERT (Beltagy, Lo, and Cohan 2019) models are employed with frozen pre-trained weights and a trainable classification head, leveraging domain-specific scientific language understanding. The transformer architecture's capability to handle semantic nuances and contextual embeddings makes it a powerful tool for categorizing scientific literature.
- **Human-augmented**: Using the best-performing automated classification model (DT) as a foundation, this method involves two expert annotators who independently reviewed and refined the model's categorization output. Each annotator reviewed all documents categorized by the model and re-categorized papers where necessary. A consensus mechanism resolved discrepancies between annotators, ensuring high-quality final labels. This combined approach leverages human expertise to improve classification accuracy, especially for ambiguous or nuanced papers that require a new category absent in C.

Classifier Name	Single-Label Setup		Multi-Label Setup	
	D_1	D_2	D_1	D_2
SVM	**0.222**	**0.346**	0.123	0.280
DT	0.124	0.258	**0.233**	0.349
RF	0.117	0.213	0.044	0.215
BERT	0.049	0.043	0.102	0.069
SciBERT	0.000	0.013	0.102	0.150
Human-augmented	-	-	-	**0.83**

Table 1: F1 Macro Scores for single-label and multi-label setups across different classifiers on datasets D_1 and D_2. The best scores across each setup are shown in bold.

Experimental Results

The F1 macro scores across single-label and multi-label setups are shown in Table 1. In the single-label setup, SVM achieved the highest scores for both datasets, with 0.222 on D_1 and 0.346 on D_2, suggesting that SVM effectively distinguishes primary categories in this setup. The Decision Tree (DT) followed closely, with 0.258 on D_2, indicating it may also be well-suited for single-label tasks with straightforward category distinctions. In the multi-label setup, DT performed best on both datasets, scoring 0.233 on D_1 and 0.349 on D_2, demonstrating its ability to manage overlapping categories more effectively than other models. SVM

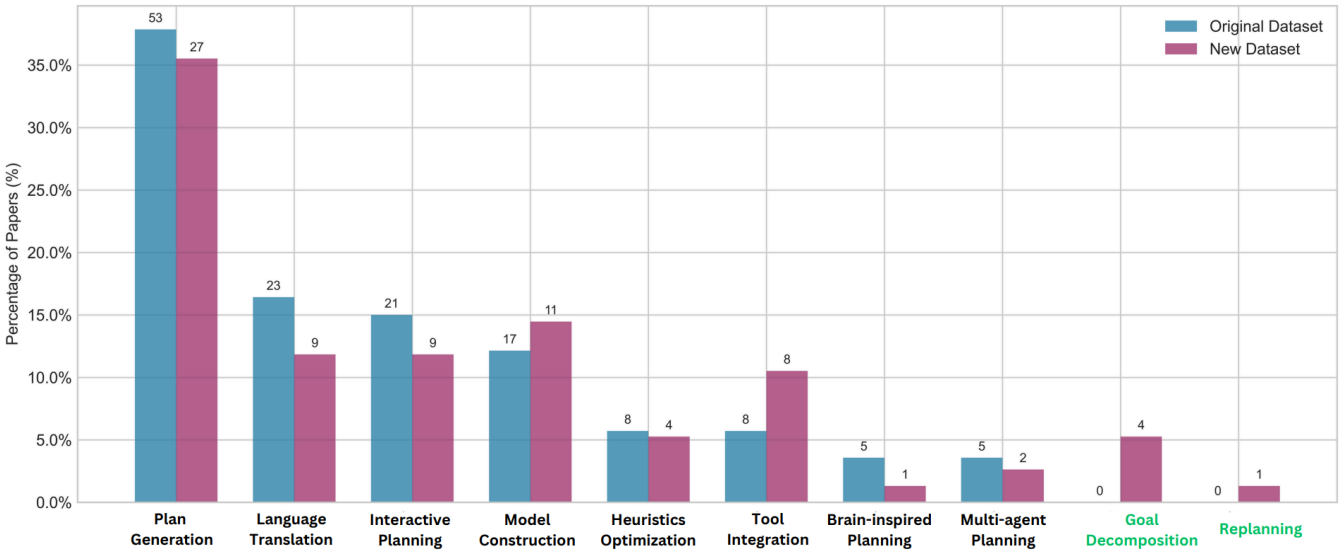

Figure 1: Comparison of category distributions between D_1 and D_2. The y-axis shows the percentage of papers, x-axis shows the categories, and the numbers above each bar represent the actual paper counts. New categories identified in D_2, *Goal Decomposition* and *Replanning*, indicate emerging research directions in using LLMs for planning.

also showed strong performance in this setup, particularly on D_2 with an F1 score of 0.280. These results suggest that DT and SVM may better capture the complexity of multi-label classification, where documents often span multiple categories.

Transformer-based models, BERT and SciBERT, showed lower performance across both setups, particularly with a minimal score on D_2 in the single-label setup (SciBERT: 0.013; BERT: 0.043). Their relatively low scores could be due to limited training data, which typically affects transformer models that require large datasets to capture fine-grained distinctions. While BERT and SciBERT are designed for general language understanding, they may lack the specific adaptations needed for categorizing niche research topics in AI and planning domains.

The human-augmented method achieved a macro F1-score of 0.83, outperforming both traditional machine learning and transformer-based models in multi-label setup on D_2. The inter-annotator agreement, measured using Cohen's Kappa, was 0.78, indicating substantial agreement between annotators and supporting the consistency of human categorization decisions. This higher performance can be attributed to the human annotators' ability to capture subtle distinctions within overlapping categories and to identify new, emerging research areas that automated models were unable to detect in the current setup. Consequently, this human-augmented approach offers a significant advantage in scenarios requiring high classification accuracy and adaptability to nuanced and evolving categories.

Analysis of Category Distribution

In our analysis, we compared the distribution of research papers across categories in two datasets: D_1 and D_2. D_1, established in prior work (Pallagani et al. 2024), organized 126 papers into eight distinct categories, represented as the first eight categories in the bar plot. These categories include **Plan Generation**, **Language Translation**, **Interactive Planning**, **Model Construction**, **Heuristics Optimization**, **Tool Integration**, **Brain-Inspired Planning**, and **Multi-agent Planning**.

As research in LLMs and planning has progressed, we applied our automated categorization tool to classify the papers in D_2. This new set of research papers not only shifted the distribution of papers across the existing eight categories but also introduced two new categories: **Goal Decomposition** and **Replanning**.

Figure 1 visually illustrates these changes in category distribution. The y-axis represents the percentage of papers, while the numbers displayed above each bar indicate the actual count of papers in each category. The x-axis represents the existing eight categories along with the two newly identified categories. Key observations include that **Plan Generation** remains the most dominant category, although its decreased percentage reflects a shift in research focus as new studies emphasize specialized, task-oriented applications of LLMs. This change is driven by a growing belief that LLMs face fundamental challenges in generating fully executable plans independently, as highlighted in recent research on their planning limitations. Similarly, the reduced representation in **Language Translation** and **Interactive Planning** categories may be attributed to researchers recognizing that while LLMs perform well in isolated translation and interaction tasks, they face difficulties when these capabilities must be applied within complex, end-to-end planning frameworks. **Model Construction** has become the second-highest category by percentage of papers in D_2, reflecting increased efforts to leverage the parametric knowledge of LLMs for building planning domain models.This shift indi-

cates a trend toward leveraging LLMs to automate the labor-intensive process of generating structured planning domain representations, traditionally one of the most manually demanding tasks in planning. Furthermore, the stable presence of **Heuristics Optimization** in both datasets suggests sustained interest in optimizing planning strategies within the constraints of current LLM capabilities, with a focus on improving efficiency rather than achieving autonomous planning.

The growth in **Tool Integration**, which holds the third-highest percentage in D_2, reflects a shift toward augmenting LLMs with external systems that can supplement their planning capabilities. Conversely, the decrease in **Brain-Inspired Planning** and **Multi-agent Planning** suggests a shift in focus away from mimicking high-level cognitive processes and multi-agent dynamics solely within LLMs. In the case of brain-inspired planning, researchers appear to be moving towards neurosymbolic systems, which integrate symbolic reasoning with neural networks rather than purely emulating cognitive processes within LLMs. This approach has shown to be more effective in combining structured logic with data-driven insights, especially for complex planning tasks requiring precise execution. Similarly, the reduced interest in multi-agent planning reflect the challenges LLMs face in coordinating interactions across multiple agents autonomously, particularly when real-time or high-fidelity coordination is required. The newly identified categories, **Goal Decomposition** and **Replanning**, are notable additions with 4 and 1 papers, respectively, in D_2. These categories point to emerging research directions focused on task structuring and adaptability, as researchers increasingly explore LLMs potential to support modular planning and real-time adjustments. Their emergence reflects the growing interest in creating flexible LLM-based systems suited for dynamic, real-world applications.

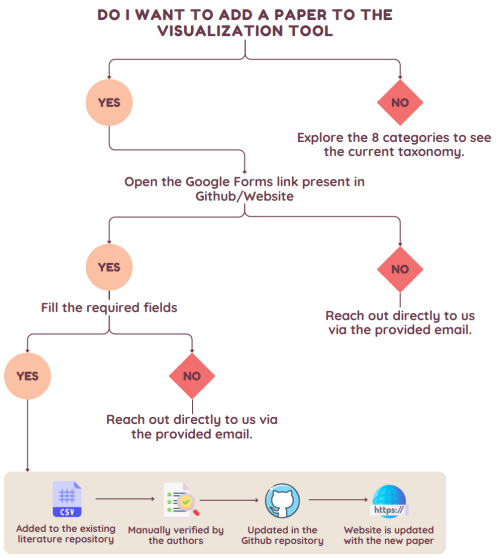

Figure 2: Decision flowchart for adding research papers to the visualization tool, detailing the submission, verification, and integration process.

This distribution shift reflects broader conversations in the field regarding the role and limitations of LLMs in planning tasks. For instance, recent works argue that "LLMs can't plan by themselves but can aid planning within a framework that integrates external verifiers" (Kambhampati et al. 2024), while others emphasize that direct planning through LLMs often fails upon execution, despite their strength in commonsense reasoning (Li et al. 2024). Furthermore, there is a call to develop foundation models specifically tailored for planning-like tasks, as current LLMs lack the specialized training needed to meet the intricate demands of these domains (Srivastava and Pallagani 2024).

Testbed and Tools

To encourage innovation and rigor in automated research paper categorization, we will open-source our tool for automatic extraction of relevant literature and release a human-annotated dataset containing categorizations of the updated dataset of papers on LLMs and planning. Additionally, a leaderboard will be established to motivate contributions of novel automated categorization methods. This leaderboard will feature baseline methods, as detailed in the previous section, allowing for transparent benchmarking and fostering a collaborative environment for advancements in automated categorization techniques. Additionally, we have developed an interactive visualization tool that allows users to explore research papers systematically categorized across various categories. This tool not only facilitates streamlined access to pertinent literature within each category but also provides real-time insights on trending topics, category shifts, and emerging areas of study. Leveraging our automated paper extraction and categorization pipeline, new publications are dynamically incorporated into the visualization. Furthermore, a manual submission feature enables users to add their own research papers, with the submission workflow outlined in Figure 2[1].

Conclusion

This study presents a scalable, adaptive framework for tracking literature evolution in LLM and planning research, combining automated categorization with human-augmented refinement to address category drift and emerging themes. Our approach reports on category drift from the previous taxonomy revealing a decline in the percentage of papers for six categories, an increase in two, and the emergence of two new categories—Goal Decomposition and Replanning—that extend the prior taxonomy. The automated system and leaderboard streamline the integration of new research while fostering collaborative innovation in classification methods. By enabling a dynamic understanding of the field's trajectory, this approach helps researchers efficiently navigate and contribute to the complex, rapidly advancing intersection of LLMs and planning. This work is representative of an emerging trend to automate review of literature and generate insights (TAMA-Review 2024), and one can

[1]GitHub with the tool, dataset, and necessary code will be made public post review phase

extend it in future to support other disciplines beyond AI planning.

References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20(3): 273–297.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L. P.; and Murthy, A. B. 2024. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Forty-first International Conference on Machine Learning*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.

Li, H.; Chen, Z.; Zhang, J.; and Liu, F. 2024. LASP: Surveying the State-of-the-Art in Large Language Model-Assisted AI Planning. *arXiv preprint arXiv:2409.01806*.

Olson, R. S.; Cava, W. L.; Mustahsan, Z.; Varik, A.; and Moore, J. H. 2018. Data-driven advice for applying machine learning to bioinformatics problems. In *Pacific symposium on biocomputing 2018: Proceedings of the pacific symposium*, 192–203. World Scientific.

Pallagani, V.; Muppasani, B. C.; Roy, K.; Fabiano, F.; Loreggia, A.; Murugesan, K.; Srivastava, B.; Rossi, F.; Horesh, L.; and Sheth, A. 2024. On the prospects of incorporating large language models (llms) in automated planning and scheduling (aps). In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, 432–444.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Srivastava, B.; and Pallagani, V. 2024. The Case for Developing a Foundation Model for Planning-like Tasks from Scratch. *arXiv preprint arXiv:2404.04540*.

TAMA-Review. 2024. Selected AI-Based Literature Review Tools. *https://tamu.libguides.com/c.php?g=1289555*.

Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Philip, S. Y.; et al. 2008. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1): 1–37.