PlanGPT: Enhancing Urban Planning with a Tailored Agent Framework

He Zhu[2], **Guanhua Chen**[3*], **Wenjia Zhang**[1,2*]
[1]College of Architecture and Urban Planning, Tongji University
[2]Behavioral and Spatial AI Lab, Tongji University & Peking University
[3]Southern University of Science and Technology
zhuhe@stu.pku.edu.cn, wenjiazhang@tongji.edu.cn

Abstract

In the field of urban planning, general-purpose large language models often struggle to meet the specific needs of planners. Tasks like generating urban planning texts, retrieving related information, and evaluating planning documents pose unique challenges. To enhance the efficiency of urban professionals and overcome these obstacles, we introduce **PlanGPT**, the first specialized AI agent framework tailored for urban and spatial planning. Developed through collaborative efforts with professional urban planners, PlanGPT integrates a customized local database retrieval system, domain-specific knowledge activation capabilities, and advanced tool orchestration mechanisms. Through its comprehensive agent architecture, PlanGPT coordinates multiple specialized components to deliver intelligent assistance precisely tailored to the intricacies of urban planning workflows. Empirical tests demonstrate that PlanGPT framework has achieved advanced performance, providing comprehensive support that significantly enhances professional planning efficiency.

1 Introduction

Due to the impressive reasoning, memory, and comprehension abilities inherent in large language models(OpenAI, 2022, 2023; Touvron et al., 2023; Qwen et al., 2025; Anthropic, 2023; DeepSeek-AI et al., 2025), substantial progress and prospects have arisen in various domains. Particularly in fields like finance(Zhang et al., 2023b), medicine(Wang et al., 2023; Xiong et al., 2023), and law(Cui et al., 2023a), specialized AI systems and agent frameworks tailored to specific verticals have emerged, efficiently tackling challenges commonly associated with general-purpose large models, such as vague responses and hallucinations caused by uniform training data distribution, thereby boosting staff productivity through intelligent task coordination and domain-specific capabilities.

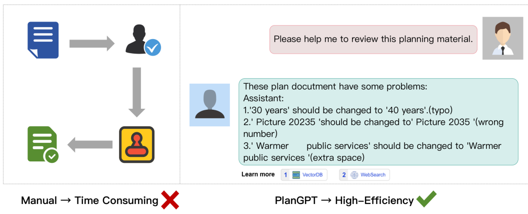

Figure 1: Manual vs. PlanGPT-assisted planning document review workflow, demonstrating improved efficiency through automated issue detection and correction suggestions.

In the field of urban planning, urban planners spend significant time on document management, review, and assessment tasks. These include evaluating planning documents against standard frameworks and assessing them across multiple dimensions like legality, feasibility, and economic viability. Leveraging the robust comprehension and reasoning abilities of LLMs through intelligent agent systems, we posit that the aforementioned processes can be addressed through a comprehensive AI framework that coordinates multiple specialized capabilities, as shown in Figure 1.

However, in practical operations, we have found that developing such an agent system is not straightforward due to the inherent nature of the urban planning industry and the characteristics of urban planning texts: **Government document style:** Linked to government affairs, urban planning documents often employ fixed phrases and structures, creating a challenge for AI systems to balance government style with informative content. The low signal-to-noise ratio (where useful information is obscured by large amounts of standardized text and boilerplate language) in these documents complicates information retrieval and processing. Moreover, heightened attention to data security restricts system design choices. **Interdisciplinary knowledge:**

[*]Corresponding Authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 764–783

Urban and spatial planning texts integrate knowledge from multiple disciplines such as environmental science, ecology, economics, and law. However, current AI systems have not effectively coordinated the activation and application of knowledge across these specialized fields, making it difficult to provide comprehensive planning support. **Timeliness and content heterogeneity:** Urban planning workflows require synchronization with government regulations and involve diverse content types including descriptions, tabular data, and spatial information, necessitating intelligent coordination of specialized tools and real-time information access.

To address the distinctive challenges inherent in urban planning workflows, we introduce **PlanGPT**, the first specialized AI agent framework for urban planning that coordinates multiple intelligent components to address three fundamental challenges in the domain. PlanGPT employs a comprehensive agent architecture that orchestrates specialized capabilities: *PlanRAG*, a domain-aware retrieval system that overcomes distinctive terminology and low signal-to-noise ratio in planning documents through specialized embeddings and hierarchical search strategies; *PlanLLM*, which activates dormant urban planning knowledge through systematic probing and targeted instruction synthesis rather than knowledge injection; and *PlanAgent*, which integrates specialized tools for spatiotemporal analysis, web access, and urban simulations to handle multimodal planning documents while maintaining regulatory compliance. Through intelligent intent recognition and multi-dimensional response scoring, PlanGPT coordinates these components to provide comprehensive assistance that addresses the unique challenges of governmental document style, interdisciplinary knowledge requirements, and content heterogeneity. Experimental evaluations demonstrate that PlanGPT framework shows promising results compared to generic state-of-the-art models across four essential planning tasks, demonstrating its potential as a comprehensive AI assistant framework for urban planning professionals.

2 Related Works

Large Language Models and Domain Applications Large language models (LLMs) have demonstrated versatility across general-purpose and domain-specific applications. General-purpose models (OpenAI, 2023, 2022; Touvron et al., 2023; et al., 2023b; Anthropic, 2023; Mistral-AI, 2023; DeepMind, 2023) showcase broad capabilities, while Chinese language models (DeepSeek-AI et al., 2025; Baichuan, 2023; Du et al., 2022; Qwen et al., 2025; Wei et al., 2023; Cui et al., 2023b) address specific language challenges. Vertical-specific LLMs have emerged across various domains, such as HuaTuo(Wang et al., 2023) and DoctorGLM(Xiong et al., 2023) in medicine, ChatLaw(Cui et al., 2023a) in legal, XuanYuan 2.0(Zhang et al., 2023b) in finance, and MathGPT(Tycho Young, 2023) for mathematics. In urban planning and related fields, specialized models include TrafficGPT(Zhang et al., 2023a) for urban traffic management, NASA's Prithvi(et al., 2023a) for climate and geography predictions, TransGPT(Peng, 2023) for transportation applications, and EarthGPT(Zhang et al., 2024) for remote sensing image interpretation. CityGPT(Feng et al., 2024) and UrbanGPT(Li et al., 2024b) focus on spatial reasoning and urban predictions respectively, but neither fully addresses comprehensive urban planning needs. Currently, no model specifically addresses urban and spatial planning, which motivates our introduction of PlanGPT.

Hallucination Mitigation Techniques Domain-specific models require high levels of factual accuracy and faithfulness. Several approaches have proven effective in mitigating hallucinations. Retrieval-augmented generation (RAG) combines LLMs' parametric knowledge with external information sources (Huang et al., 2023a; Borgeaud et al., 2022; Kim et al., 2023; Cheng et al., 2024). Advanced frameworks like Self-RAG(Asai et al., 2023) introduce specialized tokens to determine document retrieval needs, RA-DIT(Lin et al., 2023) enhances retriever relevance, and HippoRAG(Gutiérrez et al., 2025a,b) combines LLMs, knowledge graphs and PageRank for enhanced knowledge retrieval. Instruction fine-tuning (Wei et al., 2022; Longpre et al., 2023) significantly improves model capabilities and reduces hallucinations through methods by (Li et al., 2023b; Zheng et al., 2024; Lou et al., 2023), with data quality ensured via filtering techniques from (Liu et al., 2024a; Li et al., 2023a; Du et al., 2023). Approaches like self-instruct(Wang et al., 2022), wizardlm(Xu et al., 2023), magpie(Xu et al., 2024) increase training data quality to enhance robustness. Agent-based systems can select appropriate tools including web searches (webglm(Liu et al.,

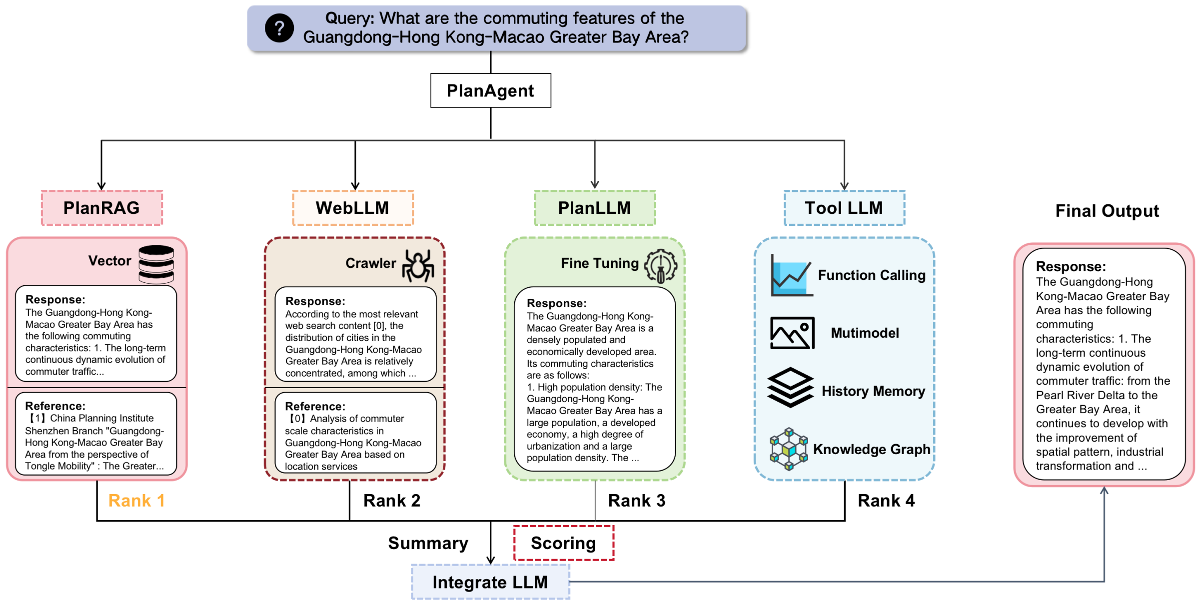

Figure 2: Overview of **PlanGPT**. The framework consists of three key components: *PlanRAG* for domain-specific retrieval, *PlanLLM* for knowledge activation and instruction tuning, and *PlanAgent* for tool integration like (WebLLM, ToolLLM) and regulatory compliance. These components work together to address the unique challenges of urban planning texts while maintaining high accuracy and reliability.

2023), webgpt(Nakano et al., 2021)) or function calls to improve output quality. Drawing on these advances, we propose novel retrieval and instruction labeling methods specifically for urban planning domains, along with PlanAgent to effectively address hallucination issues.

3 PlanGPT Framework

3.1 Overview of PlanGPT Framework

The PlanGPT framework is a comprehensive AI agent system specifically designed for urban planning regulatory environment and professional workflows. As illustrated in Figure 2, the system processes urban planning queries through PlanAgent, which orchestrates four specialized components: PlanRAG for domain-specific retrieval, WebLLM for real-time web search, PlanLLM for knowledge activation and generation, and ToolLLM for professional tool orchestration. While the core methodology is generalizable, the current implementation focuses on Chinese planning practices, incorporating China-specific regulatory frameworks and governmental document styles to support planners across national to local levels.

We detail how this coordinated architecture addresses three critical challenges through specialized components: *PlanAgent* (Section 3.2) orchestrates comprehensive task coordination and tool integration (ToolLLM and WebLLM), *PlanRAG* (Section 3.3) handles specialized terminology and low signal-to-noise ratio through domain-aware retrieval, and *PlanLLM* (Section 3.4) enables knowledge activation through targeted instruction synthesis. These components ensure **accuracy and reliability** in content adherence to governmental standards, **domain expertise** across multiple disciplines, and **timeliness** in processing diverse planning documents.

3.2 Comprehensive Agent Architecture

Intelligent Query Processing and Routing Upon receiving a planning query, PlanAgent analyzes query intent through specialized classifiers to determine optimal routing: domain-specific knowledge retrieval (*PlanRAG*), real-time regulatory information (*WebLLM*), knowledge-activated generation (*PlanLLM*), or specialized analysis tools (*ToolLLM*). The agent employs query rewriting techniques to optimize each component's input while preserving domain-specific terminology and planning context.

Specialized Component Coordination WebLLM handles real-time information access through goal-oriented web search specifically designed for urban planning sources. It employs specialized crawlers targeting governmental websites, planning bureaus, and regulatory databases, maintaining accuracy through domain-specific URL filtering and content validation mechanisms. **Tool-LLM** coordinates professional analysis tools including spatiotemporal analysis systems (Liu and Zhang, 2023; Zhang and Ning, 2023), urban simulations (Zhang et al., 2020), and knowledge graph construction. It handles function calling for specialized computations, maintains history memory for context-aware analysis, and integrates heterogeneous data sources including geographical information, demographic data, and regulatory constraints.

Response Integration and Optimization After collecting responses from active components, PlanAgent applies scoring mechanisms evaluating domain relevance, factual accuracy, regulatory compliance, and response completeness. The agent employs customized reward models trained on planning professional feedback to rank candidate responses. For complex queries requiring multiple perspectives, summarization techniques synthesize information from multiple sources, ensuring coherent final outputs that maintain professional standards while addressing all query aspects (detailed implementation in Appendix A.3).

3.3 Domain-Aware Retrieval Architecture

Urban planning documents exhibit low signal-to-noise ratios and specialized terminology that challenge conventional retrieval systems. To enable effective domain-specific retrieval, we introduce *Plan-Emb* for specialized embeddings and *Plan-HS* for hierarchical search.

Plan-Emb: Specialized Embedding Model We introduce Plan-Emb, an embedding model specialized for urban planning knowledge that addresses two key challenges: specialized terminology (where "regulations" typically means "zoning regulations") and planner's perspective (where "land use" encompasses complex interactions between people, land, and ecosystems). To construct training data, we first extract individual sentences from our urban planning document corpus. For each sentence, we use a language model to generate multiple semantically equivalent paraphrases as positive examples, while randomly sampling other sentences from the corpus as negative examples (Examples are shown in Appendix B.5.1). Plan-Emb employs a two-stage training process with InfoNCE loss augmented by KL divergence regularization to prevent catastrophic forgetting:

$$\text{loss} = -\log \frac{e^{\text{sim}(h^q, h^{a^+})/\tau}}{\sum_{i=0}^{N} e^{\text{sim}(h^q, h^{a_i})/\tau}} + \lambda D_{KL}(P||Q)$$

Plan-HS: Hierarchical Search System To address low signal-to-noise ratio challenges in planning documents, Plan-HS employs a hierarchical approach that combines keyword extraction through a fine-tuned model (detailed in Appendix A.1.1) with semantic similarity scoring. During preprocessing, documents are processed into chunks with extracted keywords stored in hashmaps. The search process recalls relevant documents using both keyword similarity and semantic similarity, then applies exact matching and cross-attention scores for result reranking to enhance accuracy (More details in Appendix A.1 and Section 4.4).

3.4 Knowledge Activation Through Instruction Synthesis

Urban planning requires multi-disciplinary knowledge that general models struggle to coordinate effectively. To activate dormant domain knowledge without extensive retraining, PlanLLM builds upon previous work (Zhou et al., 2024)'s insight that pre-trained models contain dormant knowledge requiring activation rather than injection. Our approach first identifies the urban planning knowledge embedded in the base model, then synthesizes high-quality SFT data to activate this knowledge while minimizing distribution gaps.

In **Stage (1): Knowledge Probing**, we leverage a prompt-based method inspired by GLAN (Li et al., 2024a) to systematically generate a comprehensive knowledge tree of urban planning concepts using the instruction-tuned version of our base model (detailed in Appendix 6). Our approach employs a balanced exploration strategy combining breadth-first and depth-first searches, where leaf nodes capture detailed, fine-grained knowledge points. Through this structured process, we effectively map out the urban planning knowledge that already exists within the base model's parameters.

For **Stage (2): Data synthesis**, we retrieve relevant text segments from high-quality textbook materials indexed in our *PlanRAG* system, using the knowledge points $K = \{k_1, k_2, ..., k_n\}$ identified in the probing stage. We employ a prompt-based Doc2QA transformation function

$f : (k_i, D_i) \to (q_i, a_i)$ that converts each knowledge point k_i and their associated D_i documents into instruction-response pairs to activate dormant knowledge.

In **Stage (3): Filtering and Rewriting**, generated instruction-response pairs undergo multi-dimensional filtering including deduplication, quality evaluation with a reward model (Liu et al., 2024b), complexity assessment (Lu et al., 2023), and diversity enhancement using k-center algorithm (Sener and Savarese, 2017) to ensure high quality. Inspired by (Yang et al., 2024), we employ a fine-tuned model to rewrite responses while preserving semantic meaning, minimizing the distribution gap between synthetic data and the model's internal representations. This approach produces training examples that better align with the model's learned distributions while maintaining the core domain knowledge.

4 Experiment

In this section, we demonstrate the effectiveness of our PlanGPT framework through comprehensive offline and online experiments.

4.1 Experimental Setup

Implementation Details Our training data consists of three main components: (1) knowledge activation data as introduced in Section 3.4, synthesized from study materials, Q&A threads, textbooks, and government documents (see appendix C.2); (2) manually annotated task-specific training data covering the four core tasks shown in Table 2; and (3) general-domain instruction data curated from datasets like ShareGPT and Alpaca-52k, totaling approximately 50k instruction pairs across all three components. We selected GLM3-base[1] as the base models. Implementation used the Transformers framework with AdamW optimizer (5e-5 initial learning rate), DeepSpeed ZeRO-3, and FlashAttention-2.

Evaluation Framework We conduct comprehensive evaluation through two complementary approaches: **offline experiments** using standardized benchmarks for systematic assessment, and **online experiments** for real-world applicability validation.

[1]We also evaluated Qwen2.5-7B as an alternative base model to leverage recent LLM advances while addressing data privacy concerns in urban planning.

(1) Offline Evaluation: We utilize PlanBench (Deng et al., 2025), a comprehensive benchmark for evaluating urban planning capabilities in large language models. PlanBench adopts Bloom's revised taxonomy covering five cognitive levels (Remember, Understand, Apply, Analyze, Evaluate) across urban planning knowledge domains. The benchmark integrates disciplinary knowledge systems from leading institutions and professional qualification examinations across multiple countries, providing systematic assessment through 4 major categories, 24 intermediate classes, and 81 subcategories with Content Validity Index confirmation.

(2) Online Evaluation: We assess practical applicability through two components: (1) Four core urban planning tasks from professional workflows including proposal generation (generating planning proposals and documents), style transfer (adapting planning documents between different formats and styles), information extraction (extracting key planning metrics and requirements), and evaluation (assessing planning documents and proposals) (see Table 2 and detailed task descriptions in Appendix B.2). (2) A two-part knowledge test combining C-Eval(Huang et al., 2023b)'s 418-question urban planning subset (v1) with our curated collection of 3,500 questions from Chinese Registered Urban Planner certification examinations (v2), representing both standardized assessment and real-world professional requirements.

Baselines For offline evaluation, we compare against advanced language models across three categories: Chinese-English bilingual models (Yi-6B, ChatGLM3, Qwen series (Qwen et al., 2025)), English-focused models (Llama3 series (Touvron et al., 2023), Gemma variants (DeepMind, 2023)), and chain-of-thought models (DeepSeek-R1 variants (DeepSeek-AI et al., 2025)) as benchmarked in PlanBench. For online evaluation, we select baseline models including ChatGLM3-6B (Du et al., 2022), Yi-6B, Qwen-7B, GPT-3.5-Turbo, Baichuan2-13B, and GPT4 (OpenAI, 2023), representing diverse architectures and capabilities. Detailed descriptions are provided in Appendix B.3.

4.2 Offline Results: PlanBench Evaluation

Table 1 presents comprehensive results on PlanBench across cognitive abilities. Our PlanGPT framework demonstrates competitive performance among models of comparable scale. Notably,

Models	Cognitive Abilities					Overall
	Remember↑	Understand↑	Apply↑	Analyze↑	Evaluate↑	AVG↑
Chinese-English Bilingual Models						
Yi-6B-Chat	93.8	48.1	75.3	**85.6**	26.2	65.8
ChatGLM3-6B	80.2	37.5	44.4	58.3	21.0	48.3
GLM-4-9B-Chat	91.4	72.8	84.0	79.9	38.3	73.3
Qwen2.5-0.5B-Instruct	65.4	21.0	25.9	69.4	14.8	39.3
Qwen2.5-3B-Instruct	**98.8**	66.7	**92.6**	64.0	29.6	70.3
Qwen2.5-7B-Instruct	98.8	70.4	81.5	65.9	30.9	69.5
English-focused Models						
Meta-Llama-3-8B-Instruct	95.1	58.0	72.8	78.8	**48.1**	70.6
Llama-3.1-Tulu-3-8B	60.5	56.8	30.9	80.8	16.0	49.0
Gemma-7B-it	33.3	6.2	33.3	70.8	6.2	30.0
Gemma-2-2B-it	87.7	44.4	75.3	69.0	28.4	61.0
Gemma-2-9B-it	96.3	**75.3**	90.1	67.3	33.3	72.5
Chain-of-Thought Models						
DeepSeek-R1-Distill-Qwen-7B	96.3	69.1	77.8	73.4	23.5	68.0
DeepSeek-R1-Distill-Llama-8B	93.8	64.2	75.3	78.8	28.4	68.1
Our Models						
PlanGPT (Base: ChatGLM3-6B-Base)	88.9	52.4	68.5	72.1	35.2	63.4
PlanGPT (Base: Qwen2.5-7B)	96.2	74.8	85.3	82.7	42.6	**76.3**

Table 1: Comprehensive Model Performance Comparison across Cognitive Abilities

TASK	#			Metric
	Train	Dev	Test	
Generating	1,089	100	100	Score
Style Transfer	1,181	489	489	Score
Information Extraction	1242	138	138	Acc
Text Evaluation	2345	100	100	Acc, F1

Table 2: Statistics of downstream tasks dataset. "#" indicates the number of samples. The more detailed description of each task is in Appendix B.2.

PlanGPT (Base: Qwen2.5-7B) achieves 76.3 overall score, showing balanced performance across all cognitive levels with particular strength in Apply (85.3) and Analyze (82.7) capabilities crucial for urban planning tasks.

The results reveal important insights about model capabilities in urban planning: (1) **Cognitive Balance**: PlanGPT maintains consistent performance across all levels, essential for comprehensive planning support. (2) **Domain Adaptation**: Compared to the base Qwen2.5-7B-instruct model (69.5), our domain-specific fine-tuning yields significant improvement (+6.8 points), demonstrating the effectiveness of our knowledge activation approach. (3) **Scale Efficiency**: PlanGPT achieves competitive results with smaller parameter counts, highlighting the advantages of domain-specific optimization over general-purpose scaling.

4.3 Online Results: Professional Task Evaluation

Professional Task in Urban Planning To validate our framework's effectiveness in addressing the real-world challenges, we evaluated PlanGPT against leading models across four core capabilities identified through practitioner interviews. We engaged four professional urban planning practitioners for expert assessment, while also utilizing PlanGPT itself as an auxiliary judge to assist in the review process (PlanEval). The detailed evaluation criteria and scoring rubrics are provided in Appendix B.2. Table 3 shows that PlanGPT achieves competitive performance across all essential planning tasks. PlanGPT achieves the highest human evaluation scores in text generation (86.67) and style transfer (80.00), demonstrating strong performance on governmental document styles. The framework also shows advanced capabilities in information extraction (65.18% accuracy) and text evaluation (41.00% accuracy, 35.28 F1). These results indicate that our open-source framework effectively coordinates domain-specific capabilities while achieving performance comparable to larger

[3] Yi-6B only completes 10.8% of our tests, with the majority producing responses that do not meet our requirements.

[3] We utilized ChatGPT & GPT-4 for annotating the test data, therefore we are not reporting this experiment.

Models	Text Generation		Style Transfer		Information Extraction	Text Evaluation	
	PlanEval	Human	PlanEval	Human	Acc	Acc	F1
ChatGLM (Du et al., 2022)	47.67	41.33	63.94	67.00	50.00	26.00	25.67
Yi-6B	16.00	9.00	15.41	12.00	–[2]	20.00	8.33
Baichuan2-13b-Chat(Baichuan, 2023)	62.67	34.00	43.90	39.33	50.32	33.00	17.42
ChatGPT (OpenAI, 2022)	**74.67**	58.0	66.12	70.67	–[3]	31.00	21.30
ChatGLM-2-Shots (Du et al., 2022)	65.33	52.33	**71.10**	63.67	53.81	30.00	21.76
PlanGPT Framework	60.33	**86.67**	66.80	**80.00**	**65.18**	**41.00**	**35.28**

Table 3: Online Task1: Professional Urban Planning Task Performance Evaluation

Models	v1↑	v2↑	Avg↑	δ↑
GPT-4	63.2	55.3	59.3	0.875
ChatGPT	52.2	42.0	47.1	0.805
ChatGLM3-6B	56.5	48.8	52.7	0.864
BlueLM-7B	73.0	27.2	50.1	0.373
Yi-6B	73.1	31.2	52.2	0.427
Baichuan-13b	50.5	24.7	37.6	0.489
PlanLLM	63.0	51.2	57.1	0.812

Table 4: Urban Planning Knowledge Assessment

Method	score@1	score@5	AVG
ChatGLM3-6B	-	-	48.8 (Direct Score)
Raw Search	48.7	48.5	48.6
Raw Search + PlanEmb	49.7	48.8	49.3
PlanRAG (all)	**51.9**	**52.4**	**52.2**

Table 5: Ablation Studies for PlanRAG

proprietary models.

Professional Knowledge in Urban Planning Following the methodology described in Section 3.4, PlanGPT achieved advanced performance among open-source models of comparable scale on our specialized urban planning knowledge benchmark. As shown in Table 4, our approach yielded approximately 5% accuracy improvement over the base model, with performance metrics approaching those of significantly larger proprietary models. The δ value of 0.812 indicates PlanGPT's strong knowledge alignment and reliability for governmental planning applications. This demonstrates the success of our Plan-Annotation framework and capability-focused fine-tuning.

4.4 Component Analysis: Tool Integration Effectiveness

To demonstrate the effectiveness of our framework's specialized components, we conducted ablation studies focusing on PlanRAG's retrieval capabilities and PlanAgent's tool coordination mechanisms in online task scenarios. Table 5 reveals two key findings: First, PlanRAG components show clear effectiveness - Plan-Emb contributes 0.7% improvement through domain-specific semantic understanding, while the full PlanRAG system achieves 52.2% average performance, outperforming raw search by 3.6%. Second, when comparing direct model responses (ChatGLM3-6B: 48.8) with tool-enhanced performance (PlanRAG: 52.2), our results demonstrate that PlanAgent's tool coordination provides substantial benefits over isolated model usage. These results validate our framework's core design: specialized tools like PlanRAG enhance retrieval effectiveness, while PlanAgent's coordination capabilities enable superior performance compared to standalone model responses, effectively addressing the complex requirements of urban planning workflows.

5 Conclusion

We introduced PlanGPT, the first specialized AI agent framework tailored for urban and spatial planning. Through its comprehensive agent architecture integrating a customized local database retrieval system, domain-specific knowledge activation capabilities, and advanced tool orchestration mechanisms, we successfully addressed key challenges faced by urban planners in tasks like generating planning texts, retrieving related information, and evaluating planning documents. Our empirical results demonstrate that PlanGPT achieves advanced performance while providing comprehensive support that significantly enhances professional planning efficiency. **Our system has already been successfully deployed and used in several institutions**. In the future, we will continue to refine and expand PlanGPT's capabilities to further advance intelligent assistance in urban planning workflows.

Ethical Considerations

Deploying PlanGPT in urban planning necessitates addressing several key ethical concerns:

Data Privacy Given the close ties between urban planning and government operations, we prioritize data security and privacy. Our system exclusively utilizes publicly available government documents and officially released planning materials. All training and operational data comes from authorized sources including published urban plans, zoning regulations, and publicly accessible government databases. This ensures compliance with data protection regulations while maintaining transparency in the planning process.

Hallucination Mitigation Given the real-world impact of planning decisions, we implemented: Source-traceable attribution through PlanRAG, confidence scoring for uncertain outputs; and human validation for critical applications.

Bias Considerations We address potential biases through systematic detection mechanisms during training and evaluation, ensuring PlanGPT maintains neutrality across different planning philosophies while accurately representing diverse community needs and regulatory requirements.

6 Limitations

Despite the promising results demonstrated by PlanGPT, several limitations warrant acknowledgment:

Model Selection Our implementation relies on state-of-the-art models from 2024, which we believe possess sufficient capability to handle the complex, interdisciplinary nature of urban planning texts. Nevertheless, the effectiveness of our approach remains constrained by the capabilities of these underlying models.

Evaluation Metrics While our evaluation framework is comprehensive across various dimensions, quantitatively measuring certain qualitative aspects of urban planning work presents inherent challenges that may not be fully captured in our current metrics.

Data Volume and Knowledge Activation Our approach builds upon LIMA's hypothesis that pretrained models contain dormant knowledge requiring activation rather than injection. However, the substantial volume of fine-tuning data employed in our work may challenge this fundamental assumption, raising questions about whether high-volume fine-tuning represents genuine knowledge activation or effectively constitutes knowledge injection.

Acknowledgements

The paper is supported by the Key Project of the Shanghai Municipal Education Commission's AI-Enabled Research Paradigm Reform and Discipline Leap Program (Development of a Domain-Specific Large Language Model in the Field of Urban and Rural Planning for Enhancing Spatial Cognition and Decision-Making Capabilities) and by the Fundamental Research Funds for the Central Universities (22120250239). Guanhua was supported by National Natural Science Foundation of China (No. 62306132).

We would like to thank the support from the Spatial Planning Bureau of the Ministry of Natural Resources of China, the China Land Surveying and Planning Institute, the Planning and Natural Resources Bureau of Shenzhen Municipality, the Planning and Research Center of Guangzhou Municipality, the Shenzhen Marine Development Promotion Research Center, the China Academy of Urban Planning and Design, and Guangzhou Planning Corporation. We would also like to express our sincere gratitude to all members of the BSAI Lab for their invaluable support.

References

Anthropic. 2023. Model card and evaluations for claude models.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw. `https://github.com/PKU-YuanGroup/ChatLaw`.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Google DeepMind. 2023. Gemini. `https://gemini.google.com`.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, and Others. 2025. Deepseek-v3 technical report.

Yijie Deng, He Zhu, Wen Wang, Minxin Chen, Junyou Su, and Wenjia Zhang. 2025. Urban planning bench: A comprehensive benchmark for evaluating urban planning capabilities in large language models. †Equal contribution. *Corresponding author: wenjiazhang@tongji.edu.cn.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jakubik et al. 2023a. Prithvi-100M.

Rohan Anil et al. 2023b. Palm 2 technical report.

Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. 2024. Citygpt: Empowering urban spatial cognition of large language models.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025a. Hipporag: Neurobiologically inspired long-term memory for large language models.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025b. From rag to memory: Non-parametric continual learning for large language models.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. 2023a. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. *arXiv preprint arXiv:2310.14696*.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024a. Synthetic data (almost) from scratch: Generalized instruction tuning for language models.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023a. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *ArXiv*, abs/2308.12032.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024b. Urbangpt: Spatio-temporal large language models.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

C. Liu and W. Zhang. 2023. Social and spatial heterogeneities in covid-19 impacts on individual's metro use: A big-data driven causality inference. *Applied Geography*, 155:102947.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024a. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. 2023. Muffin: Curating multi-faceted instructions for improving instruction following. In *The Twelfth International Conference on Learning Representations*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models.

Tianle Lun, Yicheng Tao, Junyou Su, He Zhu, and Zipei Fan. 2023. Mobilityagent. `https://github.com/XiaoLeGG/mobility-agent`.

Mistral-AI. 2023. mistral. `https://mistral.ai/`.

Yohei Nakajima. Babyagi, 2023. *URL https://github.com/yoheinakajima/babyagi. GitHub repository*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2022. Chatgpt. `https://chat.openai.com`.

OpenAI. 2023. Gpt-4 technical report.

Wang Peng. 2023. Duomo/transgpt.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Q. Shao, W. Zhang, X. Cao, J. Yang, and J. Yin. 2020. Threshold and moderating effects of land use on metro ridership in shenzhen: Implications for tod planning. *Journal of Transport Geography*, 89:102878.

Q. Shao, W. Zhang, X. J. Cao, and J. Yang. 2023. Built environment interventions for emission mitigation: A machine learning analysis of travel-related co2 in a developing city. *Journal of Transport Geography*, 110:103632.

Significant Gravitas. AutoGPT.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Lagent Developer Team. 2023a. Lagent: InternLM a lightweight open-source framework that allows users to efficiently build large language model(llm)-based agents. `https://github.com/InternLM/lagent`.

XAgent Team. 2023b. Xagent: An autonomous agent for complex task solving.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Krish Mangroila Tycho Young, Andy Zhang. 2023. Mathgpt - an exploration into the field of mathematics with large language models.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, et al. 2023. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning.

Siyao Zhang, Daocheng Fu, Zhao Zhang, Bin Yu, and Pinlong Cai. 2023a. Trafficgpt: Viewing, processing and interacting with traffic foundation models.

W. Zhang, C. Fang, L. Zhou, and J. Zhu. 2020. Measuring megaregional structure in the pearl river delta by mobile phone signaling data: A complex network approach. *Cities*, 104:102809.

W. Zhang, D. Lu, Y. Zhao, X. Luo, and J. Yin. 2022. Incorporating polycentric development and neighborhood life-circle planning for reducing driving in beijing: Nonlinear and threshold analysis. *Cities*, 121:103488.

W. Zhang and K. Ning. 2023. Spatiotemporal heterogeneities in the causal effects of mobility intervention policies during the covid-19 outbreak: A spatially interrupted time-series (sits) analysis. *Annals of the American Association of Geographers*, 113(5):1112–1134.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024. Earthgpt: A universal multimodal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822*.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023b. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters.

Tianyu Zheng, Shuyue Guo, Xingwei Qu, Jiawei Guo, Weixu Zhang, Xinrun Du, Chenghua Lin, Wenhao Huang, Wenhu Chen, Jie Fu, et al. 2024. Kun: Answer polishment for chinese self-alignment with instruction back-translation. *arXiv preprint arXiv:2401.06477*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A More Details about Methodology

A.1 PlanHS

A.1.1 KeyModel Construction

KeyModel is a 0.5B lightweight model trained via supervised fine-tuning (SFT) to extract 3-5 keywords from text passages. We use tailored prompt to guide ChatGLM3-6B in generating keyword annotations, followed by manual verification to create high-quality training data. The SFT objective is: $\mathcal{L}_{SFT} = -\sum_{i=1}^{N} \log P(k_i|x;\theta)$ where k_i represents extracted keywords and x is the input passage. This design achieves an effective efficiency-performance trade-off for keyword extraction.

A.1.2 RAG Algorithm Details

PlanHS (Plan Hierarchical Search) is our proposed hierarchical search algorithm that combines keyword-based and semantic-based retrieval methods.

The algorithm consists of two main components: (1) A preprocessing stage that initializes specialized models and builds necessary data structures. (2) A hierarchical search process that leverages both keyword matching and semantic similarity to retrieve relevant documents.

The algorithm first processes the query through both keyword extraction and semantic embedding paths. It then retrieves candidate documents using both methods and combines the results. The final ranking considers both keyword matching scores and semantic relevance through cross-attention, ensuring both lexical and semantic similarity are taken into account.

Algorithm 1 PlanHS: Hierarchical Search

procedure PREPROCESS
  Initialize KeyModel and PlanEmb models
  Build vector database $V : D \to \mathbb{R}^m$ and keyword mapper $H : \{d_i\} \to \{K_i\}$
end procedure
procedure QUERYSEARCH(query)
  Extract query embedding $s \in \mathbb{R}^m$ and keywords K
  Retrieve Top($x/2$) chunks by sim$(K, K_i) \to$ **A**
  Retrieve Top($x/2$) chunks by sim$(s, v_i) \to$ **B**
  Compute keyword score: score$[d] = \sum_{k \in K \cap K_d} 1$
  Re-rank by $\alpha \cdot$ cross-att$(q, d) + \beta \cdot$ score$[d]$
  return ranked document list
end procedure

A.2 PlanLLM

You are an expert urban planner. Based on the following knowledge point, generate a detailed hierarchical knowledge tree that expands this concept into its component parts. ### Knowledge Point: ### Answer:

Table 6: Prompts for Knowledge Tree Generation

A.3 PlanAgent

In the field of urban planning, professionals are required to have a solid grasp of domain-specific knowledge while also being proficient in utilizing tools relevant to the field. Drawing inspiration from previous work involving agents (Team, 2023b; Xie et al., 2023; Team, 2023a; Hong et al., 2023; Nakajima; Significant Gravitas; Wu et al., 2023; Lun et al., 2023), we have designed and developed an agent that aligns closely with the tasks and requirements of urban planning. This agent, coined as the "**PlanAgent**", is intricately tailored to suit the intricacies of urban planning endeavors.

- **Autonomous Todo List Generation**: To assist urban planning professionals in executing complex tasks such as text review, audit, or evaluation, **PlanAgent** autonomously generates and optimizes task lists based on inputs from planners, subsequently executing them in sequence.

- **Orienteering Web Search**: **PlanAgent** utilizes **Web LLM** to access real-time planning regulations and updates. Drawing inspiration from WebGLM's web crawling (Liu et al., 2023), it employs vector queries and URL crawlers to ensure precision. To further enhance search accuracy, we implemented orienting URL crawlers specifically designed to identify information sources related to urban planning.

- **Professional Tool Invocation**: **PlanAgent** proficiently utilizes specialized domain-specific models to execute pivotal tasks integral to urban planning. These tasks include reverse geocoding, knowledge graph construction, and image captioning. Furthermore, PlanAgent integrates advanced tools developed by urban planning researchers for tasks such as spatiotemporal analysis(Liu and Zhang, 2023; Zhang and Ning, 2023), transit-oriented development (TOD) settings(Shao et al., 2020), neighborhood life-circle urban planning(Zhang et al., 2022), integrated land use and transport planning(Shao et al., 2023), urban simulations(Zhang et al., 2020), digital-twin city platforms, and other essential components of smart city initiatives. This holistic approach ensures a scholarly and comprehensive engagement with the intricate challenges inherent in urban planning endeavors.

- **Information Integration and Alignment**: **PlanAgent** autonomously consolidates outputs from diverse LLMs (e.g., Vector LLM (*PlanRAG*), Local LLM (*PlanLLM*)) and specialized models through advanced techniques. It can employ a customized reward model in DPO (Rafailov et al., 2024) or RLHF (Christiano et al., 2017) to select the optimal answer, while also utilizing a summarization model to enhance findings from multiple sources.

The overarching architecture of PlanGPT is depicted as outlined above figure 2, encapsulating its multifaceted capabilities.

B Experimental Setup

B.1 Training corpora

Our training data consists of three main components that together form approximately 50k instruction pairs:

Knowledge Activation Data We curated a specialized urban planning dataset from diverse sources, including study materials, highly-rated

Q&A threads from urban planning forums, high-quality textbooks in related majors, and official documents published by local governments in recent years. Following meticulous selection using **Urban-planning-annotation**, this component provides the foundation for domain-specific knowledge as detailed in Section **??**. Detailed statistics are provided in Appendix C.2.

Task-Specific Training Data For the development of specific capabilities, we employ urban planning data and manual annotation to generate datasets for the four core downstream tasks, as illustrated in Table 2. This component focuses on practical urban planning workflows including document generation, style transfer, information extraction, and evaluation tasks.

General-Domain Instruction Data We incorporate curated general-domain fine-tuning datasets like ShareGPT(Chiang et al., 2023) and Alpaca-52k[4](Taori et al., 2023) to maintain broad language capabilities while enhancing urban planning abilities.

Taking inspiration from LIMA, we demonstrate that even a relatively small amount of fine-tuning data can yield satisfactory results, albeit with some instability.

B.2 Downstream Tasks

. The downstream tasks are described as follows:

Text Generation Large language models offer significant advantages in generating urban planning documentation, including comprehensive land use plans, development proposals, and zoning ordinances. By leveraging these models, urban planning professionals can streamline the process of drafting complex documents, ensuring clarity, coherence, and adherence to legal and regulatory frameworks. To evaluate the quality of the generated content, we created a grading system from 0 to 3, with four levels indicating quality from poor to excellent. Four professional urban planners provided subjective assessments, and their average rating determined the final quality score (Human) of each model, which was then converted to a 100-point scale.

Text Style Transfer Urban planners commonly employ text style transfer techniques in their workflow. Large language models can assist in transforming brief or informal texts into the specific style of urban planning communication, thereby enhancing the efficiency of urban and rural workers. The evaluation method is similarly to **Text Generation**.

Text Information Extraction Large language models can extract key information from various textual sources, including urban planning reports, public comments, and academic studies, to support data-driven decision-making in urban and spatial planning. We self-annotate the top 5 crucial keywords for each test case and calculate accuracy (Acc), which means whether our model can predict the same keywords as we expected within an acceptable range of semantic variation.

Text Evaluation LLMs can aid urban planners in evaluating urban planning proposals by assessing the feasibility, sustainability, and community impact of diverse projects, thereby offering objective evaluations and recommendations. Notably, we simplify the evaluation process by assigning style ratings from 0 to 3 to each paragraph, treating it as a classification task with accuracy (Acc) and F1 scores. Additionally, we utilize the trained model to automatically evaluate two tasks [5] and report the scores(PlanEval).

B.3 Baselines

We select several baseline models for comparison:

- **ChatGLM3-6B**(Du et al., 2022): This is the base model of the ChatGLM3-6B series, known for its smooth dialogue and low deployment threshold.
- **Yi-6B**: Yi-6B is part of the Yi series, trained on a 3T multilingual corpus, showcasing strong language understanding and reasoning capabilities.
- **Qwen-7B**: Qwen-7B is a member of the Qwen series, featuring strong base language models pretrained on up to 2.4 trillion tokens of multilingual data with competitive performance.
- **GPT-3.5-Turbo**: An advanced version of GPT-3, incorporating enhancements in model size, training data, and performance across various language tasks.
- **Baichuan2-13B**: The Baichuan2 series introduces large-scale open-source language models, with Baichuan2-13B trained on a high-

[4]Chinese and English versions

[5]Text Generation, Text Style Transfer

quality corpus containing 2.6 trillion tokens, showcasing top performance.

- **GPT4**(OpenAI, 2023): The latest iteration of the Generative Pre-trained Transformer developed by OpenAI, representing a significant advancement in natural language processing technology.

B.4 Urban and Rural Planner Test V2 Question Samples

Chinese version of the questions:

1. 城市发展与社会关系错误的是____。
   (a) 城市是社会矛盾的集合体
   (b) 城市是社会问题集中发正地
   (c) 城市中旧的社会问题的解决不会带来新的社会问题
   (d) 社会问题的解决是城市发展目标和现实动力

   Answer: c

2. 关于文艺复兴和绝对君权时期，欧洲城市建设特征的表述，正确的是____。
   (a) 文艺复兴时期，具有古典风格的广场，街道是地市的主要特征
   (b) 文艺复兴时期，众多中世纪新建成的城市进行了系统的有机更新
   (c) 绝对君权时期，在欧洲国家首都建设中，伦敦城市改建影响最大
   (d) 绝对君权时期，纵横交错的大道是城市建设的典型特征之一

   Answer: a

3. 根据《市级国土空间总体规划编制指南（试行）》，居住用地规划内容要求不包括____。
   (a) 优化空间结构和功能布局、改善职住关系
   (b) 引导政策性住房优先布局在交通和就业便利地区
   (c) 进一步提升人均居住用地面积
   (d) 严控高层高密度住宅

   Answer: c

English version of the questions (Translated from Chinese version):

1. Which of the following statements about urban development and social relations is incorrect?
   (a) Cities are aggregates of social contradictions
   (b) Cities are places where social problems concentrate
   (c) The resolution of old social problems in cities will not lead to new social problems
   (d) The resolution of social problems is both the goal and realistic driving force of urban development

   Answer: c

2. Regarding the characteristics of European urban construction during the Renaissance and Absolute Monarchy periods, which statement is correct?

   (a) During the Renaissance, squares and streets with classical style were the main features of cities
   (b) During the Renaissance, many medieval newly-built cities underwent systematic organic renewal
   (c) During the Absolute Monarchy period, London's urban renovation had the greatest influence on European capital construction
   (d) During the Absolute Monarchy period, intersecting boulevards were one of the typical features of urban construction

   Answer: a

3. According to the "Guidelines for Municipal Territorial Space Master Planning (Trial)", which of the following is NOT included in residential land planning requirements?

   (a) Optimize spatial structure and functional layout, improve job-housing balance
   (b) Guide priority placement of policy-oriented housing in areas with convenient transportation and employment
   (c) Further increase per capita residential land area
   (d) Strictly control high-rise and high-density residential buildings

   Answer: c

Keyword	Explanation	Rating
煤炭	生物多样性的维护与平衡。	0
水资源开发利用	消防队员正在救火	0
产业名城	产业聚集的城市，以产业为主要经济支柱。	1

Table 7: urban-rural-STS-B-test Samples (Chinese)

Keyword	Explanation	Rating
Coal	Maintenance and balance of biodiversity.	0
Water Resource Development	Firefighters are putting out a fire.	0
Industrial City	A city with industrial clusters, where industry serves as the main economic pillar.	1

Table 8: urban-rural-STS-B-test Samples (English Translation)

Keyword	Explanation	Rating
煤炭	生物多样性的维护与平衡。	0
水资源开发利用	消防队员正在救火	0
产业名城	产业聚集的城市，以产业为主要经济支柱。	1

Table 9: urban-rural-STS-B-test Samples (Chinese)

Keyword	Explanation	Rating
Coal	Maintenance and balance of biodiversity.	0
Water Resource Development	Firefighters are putting out a fire.	0
Industrial City	A city with industrial clusters, where industry serves as the main economic pillar.	1

Table 10: urban-rural-STS-B-test Samples (English Translation)

B.5 urban-rural-STS-B-test Samples

B.5.1 Training Dataset and Test Dataset Examples

C Case Study

In this section, we will discuss relevant tasks in the domain of real-world urban planning and provide potential solutions.

C.1 TASK: Review

Review is the primary task of urban planning institute staff, as extensively discussed in Section 1, which consumes a significant amount of time. By utilizing PlanRAG to identify reference standard to document queries and then conducting reviews using PlanAgent, we believe that LLMs can detect inconsistencies, inaccuracies, or discrepancies within the text, ensuring the integrity and quality of urban planning proposals.

However, in practical work, we have found that despite sophisticated prompting, large models often fail to align with human consciousness, exhibiting extremes by either detecting minor errors that could be overlooked or excessively relaxing standards, resulting in lower recall rates.

Our solution involves employing GPT-4 to randomly introduce partial errors into urban planning text, along with indicating their locations. Our staff then identify error reasons, categorized into three types: 1. factual errors 2. spelling/grammar errors 3. stylistic errors (including harmful language). Initially, we refine the cognitive capabilities of large-scale models to discern the mere presence of errors. Subsequently, we instruct them to identify and flag errors.

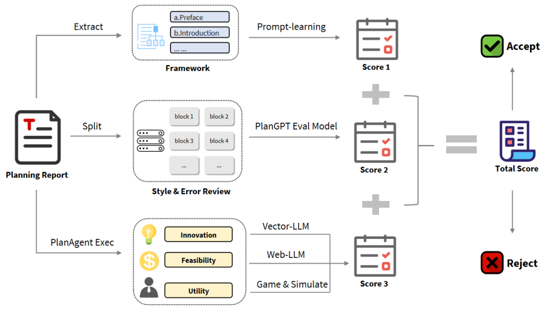

Figure 3: Assessment Task process

C.2 TASK: Evaluation

In the urban planning domain, text evaluation is a complex task, including verifying the framework of the text, reviewing the details and style of the text (as in the aforementioned review steps), and scoring the overall nature of the document. The overall nature of the document includes novelty, feasibility, and utility.

1. **Novelty**: Assessing the differences and connections with historical urban planning.
2. **Feasibility**: Urban planning needs to consider comprehensive conditions such as local economic level, geographical conditions, and interpersonal relationships.
3. **Utility**: Whether the urban planning can solve practical problems.

In actual operations, our solutions are as follows: **Novelty**: We will use PlanRAG to quickly retrieve and match historical urban planning. **Feasibility**: **PlanAgent** integrates network search tools and multimodal capabilities to solve. **Utility**: To evaluate the efficacy of the proposed plan, we will develop a simulation environment where multiple **PlanAgent**s will engage in role-playing activities. Through simulated interactions and scenario analyses, the plan's effectiveness will be assessed across diverse contexts.

Model	Output Answer
PlanGPT	大鹏新区地处深圳东部沿海地区大鹏半岛，是深圳所辖功能区。大鹏新区是深圳生态资源最为集中地区，亦是深圳最早有古人类活动的地区之一，拥有深圳历史遗存的宝贵生态与人文财富。《大鹏新区国土空间总体规划（2021-2035年）》是大鹏新区未来国土空间发展的蓝图和战略部署，是落实新发展理念、融入新发展格局、优化国土空间开发保护格局、提升城市品质竞争力的重要举措，是对深圳市国土空间总体规划、大鹏新区本级规划的引领和协调，是实现大鹏新区高质量发展、高品质生活、高效能治理的法定手段。规划是对区域自然地理条件、资源禀赋、产业基础、治理能力和治理需求的综合评价和总体部署，是未来十年间大鹏新区国土空间开发、保护、利用、修复的总纲领，是编制各专项专项规划、详细规划、村庄规划、相关专项规划的基础。本规划立足区域发展和战略定位，以国家战略为引领，以资源环境承载力为底线，以国土空间开发保护底线、资源约束性空间需求和战略布局为统筹，统筹构建新区国土空间开发保护格局，为深圳建设全球宜居宜业、宜山近海、宜业现代、生态美丽的高品质中国样本提供空间保障。
ChatGLM	大鹏新区位于深圳，是生态资源最为集中的地区之一，也是深圳最早有古人类活动的地区之一。这里拥有深圳历史遗存的宝贵生态与人文财富，是深圳寻求保护与发展相协调的重要试验场。大鹏新区是国家生态文明建设和循环经济发展的重要试验区，同时也是深圳建设全球性生态城市的窗口和平台。其自然环境优美，拥有长达13公里的海岸线，丰富的海岸带生态系统和海洋资源。此地也是深圳重要的生态保护区，保护了大量的海洋生物和栖息地。此外，大鹏新区还拥有丰富的历史文化遗存，如大鹏所城、沙头角炮台、东门古城等，是深圳历史文化的重要代表。为了保护和发展大鹏新区的生态资源，深圳市政府采取了一系列措施，包括加强海岸线管理、保护海洋生态系统、推动海洋经济发展和推广低碳生活方式等。同时，这里也建立了一批国家公园和自然保护区，如大鹏半岛国家地质公园、大鹏湾国家级自然保护区等，对这里的生态环境进行了有效的保护和管理。
Standard Answer	大鹏新区位于深圳东部沿海地区，坐落于大鹏半岛，是深圳市辖下的一个功能区。该区域拥有丰富的生态资源，也是深圳历史上古人类活动的重要区域之一，保留了深圳的历史遗产，融合了宝贵的生态和人文财富。大鹏新区成为深圳寻求生态保护与发展协调发展的重要试验场，同时也是国家生态文明建设示范区和国家"绿水青山就是金山银山"实践创新基地。规划的指导思想基于习近平新时代中国特色社会主义思想，全面贯彻党的二十大精神，立足新的发展阶段，贯彻新发展理念，服务并融入新的发展格局。规划强调高质量发展，坚持生态优先和绿色发展，以人民为中心，促进人与自然的和谐共生。整体谋划了大鹏新区面向2035年中长期发展的空间战略蓝图，旨在为深圳打造"人与自然和谐共生的美丽中国典范"提供有力支撑。此规划的制定考虑了国家和地方层面的相关要求，全面贯彻国土空间规划体系以及深圳市国土空间总体规划的相关内容。它将指导和约束下层级的详细规划编制，以确保市级和区级规划之间的有效传导。在遵循资源环境承载能力的前提下，规划强调了挖掘资源潜力，全面管理和控制所有要素，加强陆海协同发展，同时平衡生态保护和经济发展，旨在探索实现人与自然和谐共生的大鹏绿色发展之路。

Table 11: Planning text generation comparison, Chinese version

Model	Output Answer
PlanGPT	Dapeng New District is located in the Dapeng Peninsula in the eastern coastal area of Shenzhen and is a functional area under Shenzhen's jurisdiction. Dapeng New District is Shenzhen's most concentrated area of ecological resources, and also one of the earliest areas of ancient human activity in Shenzhen, possessing valuable ecological and cultural heritage from Shenzhen's history. "The Dapeng New District Territorial Space Master Plan (2021-2035)" is the blueprint and strategic deployment for the future territorial space development of Dapeng New District. It is an important measure to implement the new development concept, integrate into the new development pattern, optimize the territorial space development and protection pattern, and enhance the city's quality competitiveness. It provides leadership and coordination for Shenzhen's territorial space master plan and Dapeng New District's own plans, and is a legal means to achieve high-quality development, high-quality life, and high-efficiency governance in Dapeng New District. The plan is a comprehensive evaluation and overall deployment of the region's natural geographical conditions, resource endowments, industrial foundation, governance capabilities, and governance needs. It is the general guideline for the development, protection, utilization, and restoration of Dapeng New District's territorial space over the next decade, and is the basis for preparing various special plans, detailed plans, village plans, and related special plans. This plan is based on regional development and strategic positioning, guided by national strategy, with resource and environmental carrying capacity as the bottom line, and territorial space development and protection baseline, resource-constrained spatial demand, and strategic layout as overall considerations. It coordinates the construction of the new district's territorial space development and protection pattern, providing spatial guarantee for Shenzhen to build a high-quality Chinese model that is globally livable and suitable for work, mountains and sea, modern industry, and ecological beauty.
ChatGLM	Dapeng New District is located in Shenzhen and is one of the most concentrated areas of ecological resources, as well as one of the earliest areas of ancient human activity in Shenzhen. It possesses valuable ecological and cultural heritage from Shenzhen's history and is an important experimental field for Shenzhen's pursuit of coordinated protection and development. Dapeng New District is an important experimental zone for national ecological civilization construction and circular economy development, and is also a window and platform for Shenzhen to build a global ecological city. It has a beautiful natural environment, with a 13-kilometer coastline, rich coastal ecosystem, and marine resources. This area is also an important ecological protection zone in Shenzhen, protecting numerous marine organisms and habitats. In addition, Dapeng New District has rich historical and cultural heritage, such as Dapeng Fortress, Shatoujiao Fort, Dongmen Ancient City, etc., which are important representatives of Shenzhen's history and culture. To protect and develop the ecological resources of Dapeng New District, the Shenzhen municipal government has taken a series of measures, including strengthening coastline management, protecting marine ecosystems, promoting marine economic development, and promoting low-carbon lifestyles. At the same time, a number of national parks and nature reserves have been established here, such as Dapeng Peninsula National Geopark and Dapeng Bay National Nature Reserve, effectively protecting and managing the ecological environment.

Table 12: Planning text generation comparison (Part I), English version(Translated by Chinese version)

Model	Output Answer
Standard Answer	Dapeng New District is located in the eastern coastal area of Shenzhen, situated on the Dapeng Peninsula, and is a functional area under Shenzhen's jurisdiction. The area has rich ecological resources and is one of the important areas of ancient human activity in Shenzhen's history, preserving Shenzhen's historical heritage and integrating valuable ecological and cultural wealth. Dapeng New District has become an important experimental field for Shenzhen's pursuit of coordinated ecological protection and development, and is also a national ecological civilization demonstration zone and a national "Green Mountains and Clear Waters are Gold and Silver Mountains" practical innovation base. The guiding ideology of the plan is based on Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, fully implementing the spirit of the 20th Party Congress, standing on the new stage of development, implementing the new development concept, and serving and integrating into the new development pattern. The plan emphasizes high-quality development, adheres to ecological priority and green development, is people-centered, and promotes harmony between humans and nature. It comprehensively plans the spatial strategic blueprint for Dapeng New District's medium and long-term development toward 2035, aiming to provide strong support for Shenzhen to create a "model of beautiful China where humans and nature coexist harmoniously." The formulation of this plan considers relevant requirements at national and local levels, fully implements the territorial space planning system and the relevant content of Shenzhen's territorial space master plan. It will guide and constrain the preparation of detailed plans at lower levels to ensure effective transmission between city and district level plans. While following the carrying capacity of resources and environment, the plan emphasizes tapping resource potential, comprehensively managing and controlling all elements, strengthening land-sea coordinated development, while balancing ecological protection and economic development, aiming to explore the realization of Dapeng's green development path where humans and nature coexist harmoniously.

Table 13: Planning text generation comparison (Part II), English version(Translated by Chinese version)

Data Category	Data Description	Data Volume	Remarks
Provincial Land Spatial Planning	Overall layout and guidance for a specific province, including strategies for the allocation, utilization, and management of various resources such as land, water, minerals, and forests.	Includes 29 provincial land spatial planning texts	Shanghai and Beijing have the latest urban master plans
Municipal Land Spatial Planning	Comprehensive planning for specific cities or municipal administrative regions, providing detailed guidance on the location, area, and use of various types of land.	Includes 337 municipal-level documents	Hong Kong has plans such as Hong Kong 2030+ and Northern Metropolis Area Plan
National Land Spatial Master Plan	Comprehensive planning at the national level, based on the country's development strategy and goals, coordinating and managing the national land spatial freedom.	2820 planning-related case studies	Macau has the Macau 2040 Urban Master Plan
Spatial Planning Manuals	Includes research reports, policy recommendations, and planning proposals related to overall land spatial layout, regional coordinated development, providing decision-making basis for relevant departments.	Over 3000 planning texts at various administrative levels, case studies, and related Q&A	Open source on the internet and compiled by various planning organizations. Planning Cloud website.
Authoritative Textbooks in the Field of Planning	Approximately 200 textbooks covering urban planning, remote sensing control, regional management, and traffic engineering for undergraduate and graduate students. These textbooks encompass the complete education of urban and rural planning at the postgraduate level.	Total of 1GB of text data in PDF version	Source: Baidu Wenku, GitHub, Teaching Syllabus
Some District and County-level Land Spatial Master Plans	Land spatial planning for district and county-level administrative areas, involving resource allocation, infrastructure planning, and past versions of planning documents drafted by relevant government departments at various levels, providing guidance and strategies for local development.	Supplementary documents for county-level planning texts	Source: Spatial Planning Manuals website
Past Provincial, County, and City Land Spatial Planning Texts (2000, 2010)	Including land spatial planning texts for provinces, counties, and cities in the years 2000 and 2010.	Total of 30GB of historical planning text data	Source: Compiled from Zhihu, including municipal, county, and village-level literature