# PropGenie: A Multi-Agent Conversational Framework for Real Estate Assistance

**Chang Shen, Shaozu Yuan, Kuizong Wu, Long Xu, Meng Chen***
Yep AI, Melbourne, Australia
{chang.shen, shaozu.yuan, vincent.wu, neo.xu}@yepai.io
chenmengdx@gmail.com

## Abstract

In this paper, we present PropGenie, a novel multi-agent framework based on large language models (LLMs) to deliver comprehensive real estate assistance in real-world scenarios. PropGenie coordinates eight specialized sub-agents, each tailored for distinct tasks, including search and recommendation, question answering, financial calculations, and task execution. To enhance response accuracy and reliability, the system integrates diverse knowledge sources and advanced computational tools, leveraging structured, unstructured, and multimodal retrieval-augmented generation techniques. Experiments on real user queries show that PropGenie outperforms both a general-purpose LLM (OpenAI's o3-mini-high) and a domain-specific chatbot (Realty AI's Madison) in real estate scenarios. We hope that PropGenie serves as a valuable reference for future research in broader AI-driven applications.

## 1 Introduction

The real estate industry, while traditional, remains a cornerstone of both residential needs and investment strategies (Hudson-Wilson et al., 2005). With the advent of artificial intelligence, the integration of this transformative technology has become imperative (Ullah et al., 2018; Seagraves, 2023; Haurum et al., 2024) to address the persistent challenge of **information asymmetry**—where buyers' and investors' decisions are often influenced by intricate psychological factors and external market dynamics (Elster, 2016). Real estate transactions encompass a vast array of complex scenarios, each demanding a sophisticated understanding of domain-specific knowledge, including property valuation, legal frameworks, financing mechanisms, and evolving market trends. The dynamic and multifaceted nature of these inquiries necessitates a robust AI-driven system capable of synthesizing and interpreting information from diverse fields. Moreover, the high-stakes nature of real estate transactions imposes stringent accuracy requirements on system outputs. These challenges underscore the intricacy of developing an intelligent AI assistant for real estate applications—an endeavor that is both demanding and highly rewarding.

*Corresponding author.

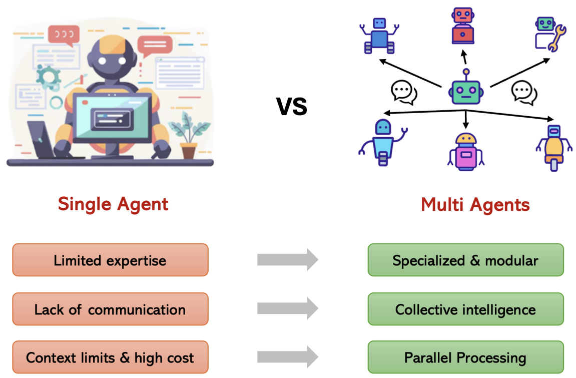


Figure 1: The difference between single-agent and multi-agents designs.

Several studies have explored the development of automated chatbots and virtual agents to provide 24/7 customer service, capture potential leads, offer legal consultations, and reduce administrative costs in the real estate industry. These efforts can be broadly categorized into two approaches: 1) **Traditional dialogue system-based methods**, which construct virtual assistants using intent-based models (Quan et al., 2018; Cao and Nguyen, 2021), frequently asked question (FAQ) systems (Tanović and Hasibović, 2024), or knowledge graph (KG)-based frameworks (Yang et al., 2024b). This approach primarily focuses on developing classification models and curating high-quality, domain-specific datasets. 2) **LLM-based agents** (Pagar, 2024; Haurum et al., 2024; Gloria et al., 2025), which leverage the advanced language comprehension capabilities of large language models (LLMs) to handle complex user queries. These agents enhance their functionality through tool calling (Qin

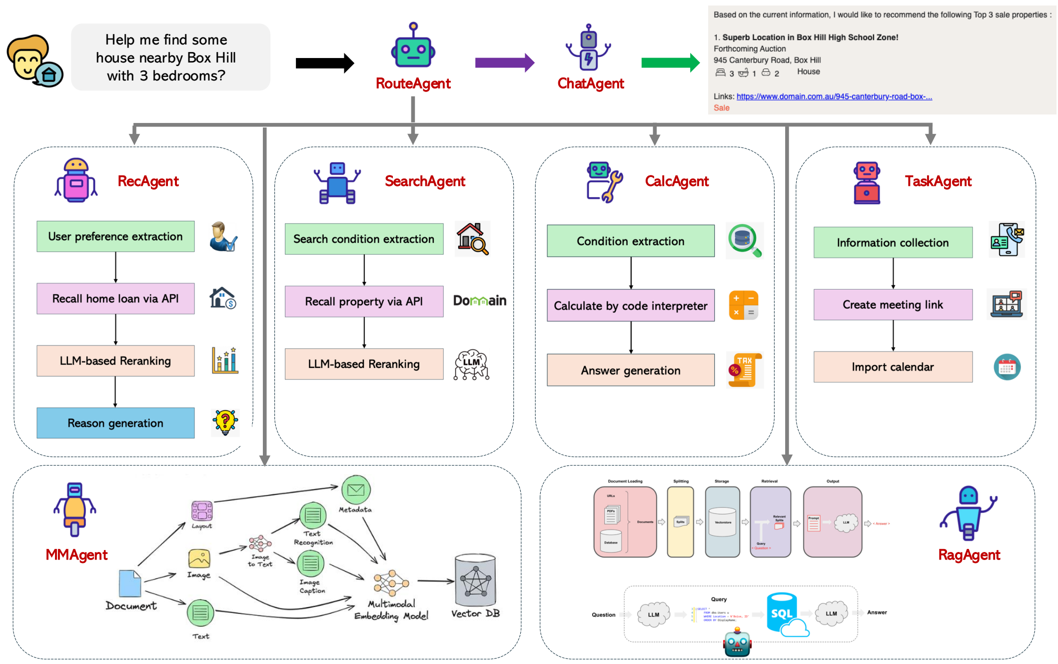


Figure 2: The overall architecture of the proposed multi-agent framework PropGenie.

et al., 2024) and retrieval-augmented generation (RAG) (Lewis et al., 2020) to improve response accuracy and contextual understanding.

With the advent of the LLM era (OpenAI et al., 2024a; GeminiTeam et al., 2024; DeepSeek-AI et al., 2025), agentic AI has demonstrated substantial potential in addressing complex real-world challenges (Durante et al., 2024). However, as illustrated in Figure 1, existing research in the real estate domain predominantly employs single-agent systems, which exhibit several limitations. First, a single agent typically lacks the specialized expertise necessary to effectively handle diverse user requests, such as property searches, home loan recommendations, question answering, and tax-related calculations. Second, the absence of inter-agent communication and cross-verification mechanisms increases susceptibility to erroneous outputs caused by LLM hallucinations (Li et al., 2024b). Third, maintaining contextual coherence in single-agent systems requires large prompts for each request, resulting in increased inference costs and reduced efficiency (Zhang et al., 2024; Wang et al., 2025b). In contrast, a multi-agent framework addresses these limitations through specialized modular design, collective intelligence, and parallel processing, making it a natural choice for developing scalable and efficient real estate applications (Zhao et al., 2024; Team et al., 2024; Gao et al., 2024a).

Motivated by the above analysis, we propose **PropGenie**, a multi-agent conversational framework designed for real-world real estate assistance. Following a *system-over-model* paradigm, PropGenie integrates domain-specific knowledge bases and advanced computational tools to deliver accurate and contextually relevant responses. Specifically, we develop eight specialized sub-agents that collaboratively handle ten distinct tasks, including property search, home loan recommendations, stamp duty and land tax calculations, monthly repayment estimations, home-buying policy QA, real-time interest rate inquiries, property project QA, automatic task execution, and open-domain chitchat. Additionally, PropGenie enhances user engagement through emotion recognition and satisfaction prediction. We evaluate PropGenie through comprehensive automatic and human evaluations, supplemented by detailed case studies to analyze its strengths and limitations. By demonstrating effectiveness in the real estate domain, our system provides valuable insights for future research in broader application scenarios.

## 2 System Architecture

In this section, we present the overall architecture of PropGenie, starting with an overview of its task scope, followed by a detailed discussion of the design and functionality of each sub-agent.

### 2.1 Task Scope

PropGenie provides an integrated solution covering property search, financing, and planning. Based on extensive analysis of user queries and feedback

| Tasks | Examples |
|---|---|
| Property search | Find a 3-bedroom house with a swimming pool near Box Hill. |
| Home loan recommendation | Recommend some good ANZ home loan products that support an offset account. |
| Stamp duty calculation | How much stamp duty do I need to pay for a $1.8 million house in Melbourne? |
| Land tax calculation | How much land tax do I need to pay for a $1.5 million house and land package in NSW? |
| Monthly repayment estimation | What's the monthly repayment for a $1 million home loan over 20 years? |
| Home-buying policy QA | What's the process of buying a home in Australia as an overseas buyer? |
| Interest rate QA | What's the current interest rate for a principal and interest loan with Westpac? |
| Property project QA | How many apartment units are there in the YarraBend project? |
| Task execution | Can you help me schedule an online meeting with the agent? |
| Open-domain chitchat | What are the main differences between living in Sydney and Melbourne? |

Table 1: Tasks and Examples Illustrations. Detailed distribution for each task can be found in Appendix A.1.

from broker agents, we define ten core tasks: property search, home loan recommendations, stamp duty and land tax calculations, monthly repayment estimation, home-buying policy QA, real-time interest rate inquiries, property project QA, automatic task execution, and open-domain chitchat. Table 1 illustrates representative examples for each task, clarifying the system's scope and capabilities.

### 2.2 Components

As illustrated in Figure 2, PropGenie adopts a multi-agent framework to efficiently handle diverse real estate tasks. The system groups related functionalities into dedicated agents, each with a streamlined workflow, simplifying the overall architecture. For example, stamp duty, land tax, and monthly repayment calculations share a common workflow involving query interpretation and code interpreter; thus, we consolidate these tasks into a single CalcAgent. Similarly, the RagAgent manages retrieval-augmented generation from structured and unstructured databases, addressing queries such as interest rates and home-buying policies. Currently, PropGenie leverages GPT-4o (OpenAI et al., 2024b) as its core LLM, yet the framework remains flexible, enabling integration of future advanced LLMs.

**RouteAgent.** The RouteAgent acts as the central orchestrator, autonomously managing system workflows. Its responsibilities include: 1) **Intent Understanding and Task Decomposition**: interpreting user queries, identifying intents, decomposing complex queries into sub-tasks, and assigning these tasks to appropriate sub-agents. It also facilitates collaboration among sub-agents to leverage their specialized capabilities. 2) **Response Aggregation and Conflict Resolution**: aggregating responses from multiple sub-agents into a coherent reply, resolving inconsistencies by selecting the most reliable response or reassigning tasks for verification, and delegating out-of-scope queries to the ChatAgent for fallback responses. 3) **Context Management and Emotional Intelligence**: maintaining conversational context by rewriting multi-turn interactions into self-contained queries, thereby simplifying context handling and enhancing robustness of RAG-based sub-agents. Additionally, it monitors user sentiment, adapting responses empathetically or strategically shifting topics when detecting user frustration.

**SearchAgent.** The SearchAgent retrieves relevant property listings by following a structured information retrieval workflow. It first extracts key search criteria from user queries, such as price range, location (e.g., suburb), housing configuration (e.g., number of rooms), and amenities (e.g., parking, swimming pool, gym). Next, it invokes a web search tool to fetch property listings from third-party APIs[1]. Finally, it re-ranks the top 50 retrieved properties against the user's requirements and selects the three most relevant listings. Both query parsing and result reranking are powered by an LLM, enhancing language understanding and retrieval accuracy.

**RecAgent.** The RecAgent recommends home loan products tailored to user profiles and preferences. It first extracts user preferences from conversational context, such as preferred bank, repayment type, interest rate type, loan term, and additional features (e.g., offset accounts). Similar to the SearchAgent, it retrieves relevant home loan products from the web[2] using a search tool and re-ranks them with LLM based on user preferences. Additionally, the RecAgent analyzes the advantages and drawbacks of each loan product, providing comprehensive evaluations to support informed decision-making.

**CalcAgent.** To deliver accurate computational responses in real estate inquiries, we introduce CalcAgent, a specialized module designed to address complex financial queries such as stamp duty,

[1] https://developer.domain.com.au/
[2] https://www.finder.com.au/home-loans

land tax, and mortgage repayment calculations. These queries pose significant challenges due to their reliance on logical reasoning and numerical precision. CalcAgent processes user-provided parameters, including property value, location, and foreign buyer status for tax assessments, as well as loan amount, loan term, and bank selection for mortgage calculations. To ensure computational accuracy, CalcAgent employs a structured three-step approach: (1) Entity extraction identifies essential numerical and categorical parameters from user queries; (2) A code interpreter dynamically generates and executes Python scripts based on predefined formulas; (3) An LLM formulates coherent, contextually relevant responses grounded in computed outcomes. Given that tax regulations and mortgage formulas vary across jurisdictions and evolve over time, we periodically collect official tax rules from government sources and formalize them into standardized mathematical expressions. Similarly, mortgage repayment formulas are derived from financial institutions' policies. These structured formulas are integrated into CalcAgent's prompts as auxiliary knowledge, enabling the model to generate accurate and executable programs aligned with current financial regulations.

**RagAgent.** RagAgent addresses inquiries related to interest rates and home-buying policies using a unified retrieval-augmented generation (RAG) framework (Gao et al., 2024b). Given the high-stakes nature of real estate transactions, ensuring information accuracy is critical. To achieve this, we systematically collect and structure domain-specific knowledge from authoritative sources. For interest rates, we aggregate real-time data from financial institutions' open APIs and regularly update a structured database. For home-buying policies, we employ web crawlers to extract regulatory updates from official government websites, accommodating frequent policy changes. Extracted content is transformed into question-answer (QA) pairs, which are automatically refined through LLM-driven quality checking and filtering (Cheng et al., 2024). These curated QA pairs are then encoded into vector representations and stored in a vector database for efficient retrieval. RagAgent processes interest rate queries by translating them into SQL queries (Wang et al., 2025a) to retrieve precise results from the structured database. For home-buying policy questions, it retrieves relevant QA pairs from the vector database and leverages an LLM to generate concise, contextually accurate responses. By integrating rigorous knowledge curation with the RAG approach, RagAgent ensures reliable, timely, and accurate real estate assistance.

**MMAgent.** Inquiries about off-plan property projects represent a significant portion of real estate assistance requests. However, obtaining reliable information on new developments is challenging due to limited public data and delayed updates. To address this, we leverage digital property brochures to extract essential details, including developer information, floor plans, unit availability, pricing, nearby amenities (e.g., schools, hospitals, shopping centers), transportation options, and comprehensive analyses of project strengths and weaknesses. Given the highly visual nature of these brochures, we propose MMAgent, a multimodal retrieval-augmented generation agent tailored for property-related QA tasks. MMAgent integrates two complementary techniques: (1) A vision-language model (e.g., GPT-4V) interprets images and floor plans, converting visual content into textual descriptions; (2) A multimodal embedding model encodes textual and visual information into vector representations, enabling efficient retrieval within a RAG-based framework. This hybrid approach allows MMAgent to effectively utilize property brochure materials, ensuring accurate and contextually rich responses to user queries.

**TaskAgent.** TaskAgent is designed to manage action-oriented requests within real estate conversations. For example, after multiple dialogue turns, prospective buyers may request to speak with a real estate agent or schedule property inspections. Similarly, developers often aim to capture contact information from high-intent buyers to generate quality leads. In PropGenie, TaskAgent autonomously detects user intent and facilitates seamless task execution. When a user requests a meeting, TaskAgent automatically generates an online meeting link and sends a calendar invitation, streamlining subsequent interactions. For lead generation, TaskAgent dynamically triggers a lead capture form at optimal moments—specifically when user satisfaction and strong buying intent are detected. This adaptive process is further enhanced by RouteAgent, which analyzes user sentiment to balance efficiency and user experience, ensuring interactions remain effective and engaging.

**ChatAgent.** To manage out-of-scope queries, we introduce ChatAgent, which directly leverages an LLM to generate open-domain conversational re-

| Tasks | Rel. | Inf. | Cor. | Tasks | Rel. | Inf. | Cor. |
|---|---|---|---|---|---|---|---|
| Property Search | 3.22 | 3.42 | 4.15 | Home-buying Policy QA | 4.28 | 3.39 | 4.74 |
| Home Loan Recommendation | 3.31 | 3.86 | 4.59 | Interest Rate QA | 4.37 | 3.76 | 4.22 |
| Stamp Duty Calculation | 4.67 | 4.05 | 3.50 | Property Project QA | 4.27 | 3.86 | 4.59 |
| Land Tax Calculation | 4.56 | 3.78 | 3.31 | Task Execution | 3.15 | 2.36 | 3.83 |
| Repayment Calculation | 3.82 | 3.15 | 3.37 | Open-domain Chitchat | 3.82 | 2.88 | 4.37 |

Table 2: Automatic evaluation results of PropGenie. "**Rel.**", "**Inf.**", "**Cor.**" represent **Relevance**, **Informativeness**, and **Correctness** correspondingly. Detailed justifications for each metric can be found in Appendix A.2.

sponses. ChatAgent serves two primary purposes: (1) as a fallback mechanism—when task-specific agents cannot produce valid responses, ChatAgent provides default replies, enhancing user experience and facilitating intent clarification; and (2) as a knowledge supplement—by utilizing the LLM's inherent knowledge and reasoning capabilities, ChatAgent effectively addresses open-ended inquiries, overcoming limitations of specialized agents. This design ensures fluid and engaging interactions, thereby improving the robustness of the overall system.

## 3 System Evaluation

**Automatic Evaluation.** Following prior work (Bi et al., 2023), we evaluate PropGenie's responses using three metrics: 1) **Relevance** – alignment with user intent and semantic consistency; 2) **Informativeness** – inclusion of detailed explanations and relevant domain-specific context; and 3) **Correctness** – factual accuracy, adherence to predefined formulas (e.g., tax calculations), and absence of hallucinations. To facilitate large-scale evaluation, we adopt the LLM-as-a-Judge paradigm (Zheng et al., 2023; Li et al., 2024a), employing GPT-5 as evaluator. Our test set comprises 6,398 real user queries from online logs, covering the 10 tasks listed in Table 1. Responses are rated from 1 (lowest) to 5 (highest), with concise justifications provided by the evaluator. Domain-specific knowledge (e.g., tax formulas, project background) is incorporated into evaluation prompts to ensure robustness and reproducibility. Further details on the test set and evaluation prompts are in the Appendix A.1&A.2.

Table 2 summarizes the automatic evaluation results. Key observations include: 1) QA-related tasks (e.g., home-buying policy QA, property project QA, interest rate QA) achieve high Correctness and Relevance scores, demonstrating RagAgent and MMAgent's effectiveness in retrieving and applying domain knowledge. 2) Search and recommendation (S&R) tasks achieve Correctness scores above 4, highlighting the benefit of integrating real-time APIs for property listings and financial products. 3) Calculation-based tasks (e.g., stamp duty, land tax, monthly repayment) show high Relevance but moderate Correctness, indicating accurate topic adherence but slight accuracy variations in complex reasoning scenarios. 4) Task execution scenarios yield the lowest Informativeness scores, likely due to efficiency-driven responses (e.g., meeting links, lead forms) appearing less detailed. Overall, automatic evaluation confirms PropGenie's capability to consistently generate relevant, informative, and accurate responses, validating its effectiveness as a conversational assistant for real-world real estate applications.

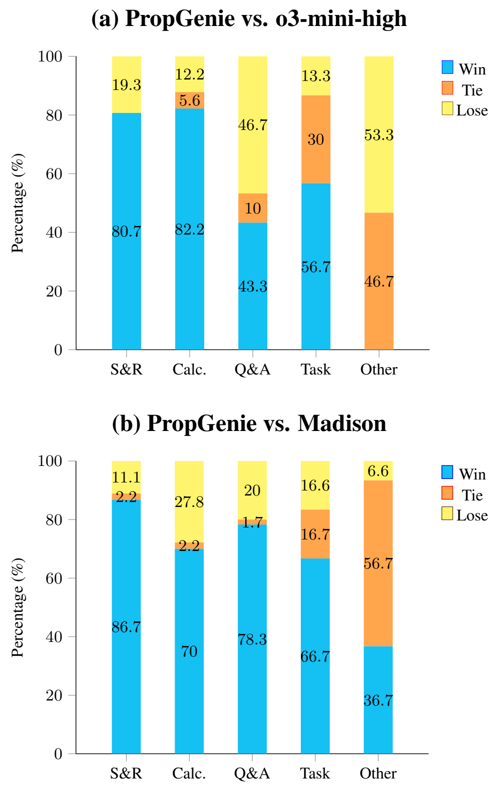


Figure 3: Human evaluation results. S&R, Calc., Q&A, Task, Other represent Search&Recommendation, Calculation, Question Answering, Task Execution, and Chitchat respectively.

**Human Evaluation.** To further validate PropGenie's effectiveness, we conducted a human evalu-

| Model | Generated Response |
|---|---|
| **Question** | **I am looking for a house in Camberwell, 3124, specifically from or related to Woodards.** |
| **PropGenie** | Here are some recommended properties for sale, all houses located in Camberwell related to Woodards agency:<br>1) Superb Single Level Living (19 Callanish Rd), price approximately $1.87M;<br>2) Sunnyside Estate without heritage overlay (43 Currajong Ave), priced at $4.95M. |
| **o3-mini-high** | Could you clarify what you mean by "from or related to Woodards", such as whether you're looking for properties listed by the Woodards Real Estate agency in Camberwell (3124), interested in houses built, designed, or associated with a builder or developer named Woodards, or if you have specific requirements like property type, budget range, number of bedrooms, or other features to help narrow down the recommendations, so I can assist you in finding the right property recommendations in Camberwell? |
| **Madison** | There are multiple houses in Camberwell, 3124 with at least 3 bedrooms and at most 2 bathrooms. I recommend checking various listings or contacting a local real estate agent for the current availability. |

Table 3: Comparing responses provided by different models. We convert rich text to plain text for demonstration.

ation comparing it against two competitive baselines: 1) OpenAI's general-purpose reasoning LLM **o3-mini-high** (OpenAI, 2025), and 2) Realty AI's domain-specific real estate chatbot **Madison**[3]. We randomly sampled 500 queries proportionally from the original 6,398-query test set, preserving the distribution. Five real estate experts performed a blind evaluation, independently ranking responses from the three systems (inter-annotator agreement: 0.82). Figure 3 summarizes PropGenie's comparative **win/tie/lose** rates, grouping the ten tasks into five categories for clarity. Results show PropGenie achieves >55% win rates over the general-purpose LLM in search and recommendation, calculation, and task execution, and >65% win rates over the domain-specific baseline in all categories except "Other". This highlights the importance of integrating external domain knowledge, typically absent in general-purpose models. However, PropGenie shows no clear advantage in QA or chitchat over o3-mini-high, as housing policy is widely available online and general models excel at detailed reasoning. Moreover, our brief-response setting reduces latency but can limit informativeness.

**Case Study.** Table 3 presents a representative property search example comparing model outputs. PropGenie accurately retrieves relevant property listings with precise location and pricing details. In contrast, **o3-mini-high** misunderstands the query and requests clarification, while **Madison** provides only generic advice to search online. Both competitive baselines fail to deliver direct answers due to limited domain-specific knowledge. Additional examples are provided in the Appendix A.4 & A.5.

## 4 Related works

**Real Estate Virtual Assistants.** Prior studies have developed virtual assistants for real estate to provide online customer support (Cao and Nguyen, 2021; Haurum et al., 2024; Yang et al., 2024b; Gloria et al., 2025), capture leads (Quan et al., 2018), offer legal advice (Pagar, 2024), and reduce administrative overhead (Tanović and Hasibović, 2024). However, these systems typically rely on traditional dialogue frameworks or single-agent LLMs, limiting their effectiveness in complex, multi-faceted scenarios. In contrast, we propose a multi-agent framework capable of addressing diverse real estate tasks, including property search, financial planning, and home-buying assistance.

**LLM-based Multi-Agent Systems.** Recent advancements in LLMs have driven the adoption of multi-agent systems, enabling inter-agent communication and collaborative problem-solving with improved accuracy and efficiency over single-agent approaches (Dong et al., 2024; Wu et al., 2023; Li et al., 2024b; Wang et al., 2025b; Hong et al., 2024). Domain-specific multi-agent frameworks have been explored in e-commerce (Thakkar and Yadav, 2024; Fang et al., 2024), legal analysis (Cui et al., 2024), finance (Fatemi and Hu, 2024), healthcare (Tang et al., 2023), and software engineering (Yang et al., 2024a). Unlike previous works, our research introduces a multi-agent conversational system tailored for the real estate domain, leveraging agent collaboration to enhance efficiency and decision-making in property-related tasks.

## 5 Conclusion

In this paper, we introduce PropGenie, a multi-agent conversational framework leveraging large language models for real estate assistance. Eight specialized sub-agents collaboratively handle tasks such as property recommendation, financial calculation, question answering, and open-domain conversation. Experiments on real user queries confirm PropGenie's effectiveness. Future work includes extending to loan eligibility assessment and automated property valuation.

[3] https://www.realty-ai.com/

## References


Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin, and Xiaodong He. 2023. DiffusEmp: A diffusion model-based framework with multi-grained control for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2812–2831, Toronto, Canada. Association for Computational Linguistics.

Tuan-Dung Cao and Quang H. Nguyen. 2021. *An Approach for Building Effective Real Estate Chatbots in Vietnamese*, pages 221–229. Springer International Publishing, Cham.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Xiaofei Dong, Xueqiang Zhang, Weixin Bu, Dan Zhang, and Feng Cao. 2024. A survey of llm-based agents: Theories, technologies, applications and suggestions. In *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, pages 407–413. IEEE.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent ai: Surveying the horizons of multimodal interaction. *Preprint*, arXiv:2401.03568.

J. Elster. 2016. *Sour Grapes*. Cambridge Philosophy Classics. Cambridge University Press.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135*.

Sorouralsadat Fatemi and Yuheng Hu. 2024. Enhancing financial question answering with a multi-agent reflection framework. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 530–537.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, Liuyi Yao, Hongyi Peng, Zeyu Zhang, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024a. Agentscope: A flexible yet robust multi-agent platform. *Preprint*, arXiv:2402.14034.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024b. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

GeminiTeam, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, and et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Benedikt Gloria, Johannes Melsbach, Sven Bienert, and Detlef Schoder and. 2025. Real-gpt: Efficiently tailoring llms for informed decision-making in the real estate industry. *Journal of Real Estate Portfolio Management*, 31(1):56–72.

Kasper Raupach Haurum, Ruiqi Ma, and Wen Long. 2024. Real estate with ai: An agent based on langchain. *Procedia Computer Science*, 242:1082–1088. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024).

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Susan Hudson-Wilson, Jacques Gordon, Frank Fabozzi, Mark Anson, and S. Giliberto. 2005. Why real estate? *The Journal of Portfolio Management*, 31:12–21.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Junyou Li, Qin Zhang, Yangbin Yu, QIANG FU, and Deheng Ye. 2024b. More agents is all you need. *Transactions on Machine Learning Research*.

OpenAI. 2025. Openai o3-mini system card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and et al. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Surbhi Pagar. 2024. Lawbot : From documents to answers, unveiling a new era in real estate legal assistance. *Interantional Journal of Scientific Research In Engineering And Management*, 08:1–5.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*.

Tho Quan, Trung Trinh, Dang Ngo, Hon Pham, Long Hoang, Hung Hoang, Thanh Thai, Phong Vo, Dang Pham, and Trung Mai. 2018. Lead engagement by automated real estate chatbot. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 357–359.

Philip Seagraves. 2023. Real Estate Insights: Is the AI revolution a real estate boon or bane? *Journal of Property Investment & Finance*, 42(2):190–199.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

A. Tanović and A. Ćerimagić Hasibović. 2024. Automated real estate chatbot. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 241–246.

SIMA Team, Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, Stephanie C. Y. Chan, Jeff Clune, and et al. 2024. Scaling instructable agents across many simulated worlds. *Preprint*, arXiv:2404.10179.

Param Thakkar and Anushka Yadav. 2024. Personalized recommendation systems using multimodal, autonomous, multi agent systems. *Preprint*, arXiv:2410.19855.

Fahim Ullah, Samad M. E. Sepasgozar, and Changxin Wang. 2018. A systematic review of smart real estate technology: Drivers of, and barriers to, the use of digital disruptive technologies and online platforms. *Sustainability*, 10(9).

Bing Wang, Changyu Ren, Jian Yang, Xinnian Liang, Jiaqi Bai, LinZheng Chai, Zhao Yan, Qian-Wen Zhang, Di Yin, Xing Sun, and Zhoujun Li. 2025a. MAC-SQL: A multi-agent collaborative framework for text-to-SQL. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 540–557, Abu Dhabi, UAE. Association for Computational Linguistics.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025b. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

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. *Preprint*, arXiv:2308.08155.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024a. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652.

Shuling Yang, Hanzhu Chen, and Binbin Fang. 2024b. Qudial: A quadruple-driven dialogue system for real estate consulting services. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, ICMLC '24, page 609–615, New York, NY, USA. Association for Computing Machinery.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan O Arik. 2024. Chain of agents: Large language models collaborating on long-context tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jun Zhao, Can Zu, Hao Xu, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Longagent: Scaling language models to 128k context through multi-agent collaboration. *Preprint*, arXiv:2402.11550.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

# A Appendix

In the appendix, we present supplementary evaluation data statistics, the prompt template for automatic evaluation, latency and cost analysis, and additional case studies.

## A.1 Dataset Statistics

As mentioned Section 3, our test set comprises 6,398 user interaction samples collected from system logs after deployment. Table 4 presents their distribution across different task categories, reflecting the real-world traffic distribution.

| Task | Count |
|---|---|
| Property Search | 374 |
| Home Loan Recommendation | 1206 |
| Stamp Duty Calculation | 557 |
| Land Tax Calculation | 472 |
| Repayment Calculation | 307 |
| Home-buying Policy QA | 641 |
| Interest Rate QA | 196 |
| Property Project QA | 152 |
| Task Execution | 169 |
| Open-domain Chitchat | 2324 |

Table 4: Task Distribution of Test Set

## A.2 Prompts for LLM-as-a-Judge and LLM-based Reranking

This section describes the prompts used for the LLM-as-a-Judge approach (Zheng et al., 2023; Li et al., 2024a) in automatic evaluation. Since relevance, informativeness, and correctness represent distinct evaluation criteria, we designed three separate prompts, as illustrated in Figure 6 to Figure 8, to independently assess each metric. Additionally, certain tasks require domain-specific knowledge beyond the inherent capabilities of LLMs. To address this limitation, we append relevant domain-specific information to the prompts when evaluating tasks such as stamp duty estimation, land tax calculation, and property project question answering.

Furthermore, Figure 9 illustrates the prompt employed for LLM-based reranking in property search as described in SearchAgent, where the LLM is guided to rank candidate listings based on how well the property features match the user's search criteria. The prompt employed in RecAgent is similar, except that we additionally instruct the LLM to generate pros and cons of home loan products during reranking. We omit this prompt here to avoid duplication.

## A.3 Latency and Cost Analysis

In this section, we analyze PropGenie's latency and cost, which are critical for user experience and commercial viability. Table 5 summarizes the average response time and cost per query, computed using the same 500 queries from the human evaluation. The latency ranges from 2 to 7 seconds, averaging 3.6 seconds. Property Search, Property Project QA, and Home Loan Recommendation exhibit higher latency due to third-party API calls and multimodal retrievers. In the future, we plan to implement streaming and caching mechanisms to further reduce latency. Additionally, the average cost per query remains below 1.6 US cents, highlighting the token-efficiency advantage of our multi-agent framework, which avoids maintaining all context information within a single prompt.

## A.4 Additional Examples of PropGenie

Due to space constraints, we present only a single case study in the main content. In this section, we provide additional examples of our system to further illustrate its advantages and limitations. In Figures 4 and 5, we supplement additional examples of scenarios involving property search, home loan recommendations, calculations, question answering, task execution, and open-domain chitchat, highlighting PropGenie's main capabilities. We have included screenshots to illustrate the user interface of our system. Please refer to our online video[5] for more interesting examples.

## A.5 Further Comparisons between PropGenie and Baselines

In this section, we present additional comparisons between PropGenie and two baseline models (OpenAI o3-mini-high and Realty AI's Madison) to illustrate the strengths and limitations of each model. Table 6 compares the responses generated by the three models for queries related to stamp duty calculation, property project QA, home loan recommendations, and interest rate QA. As shown, PropGenie consistently generates more specific, informative, and helpful answers to effectively address user queries, demonstrating the effectiveness of our multi-agent framework and our strategy of integrating diverse knowledge sources.

[4] https://platform.openai.com/docs/pricing

[5] https://youtu.be/prafKomKN3g

| Tasks | Latency | Cost | Tasks | Latency | Cost |
|---|---|---|---|---|---|
| Property search | 4.92 | 1.47 | Home-buying Policy QA | 3.55 | 1.58 |
| Home Loan Recommendation | 3.85 | 1.49 | Interest Rate QA | 2.87 | 1.27 |
| Stamp Duty Calculation | 2.73 | 1.19 | Property Project QA | 6.93 | 1.58 |
| Land Tax Calculation | 2.98 | 1.39 | Task Execution | 1.86 | 1.32 |
| Repayment Calculation | 3.46 | 1.43 | Open-domain Chitchat | 2.74 | 1.52 |

Table 5: Average latency (seconds) and cost (US cents) computed over 500 queries. Costs are calculated based on token usage and OpenAI API pricing[4].

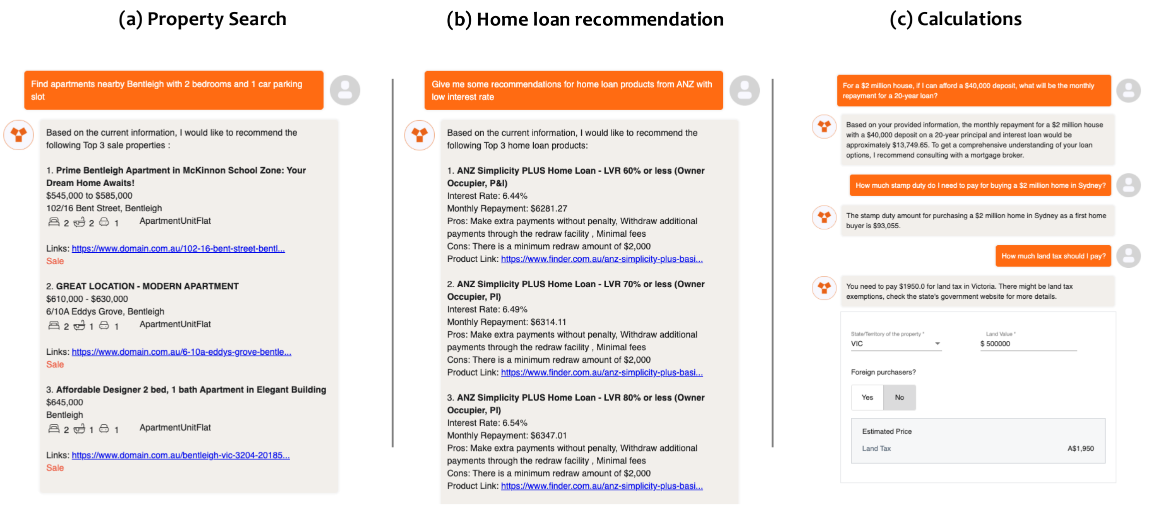


Figure 4: Examples of (a) Property Search, (b) Home Loan Recommendation, (c) Calculations (monthly repayment calculation, stamp duty calculation, and land tax calculation).

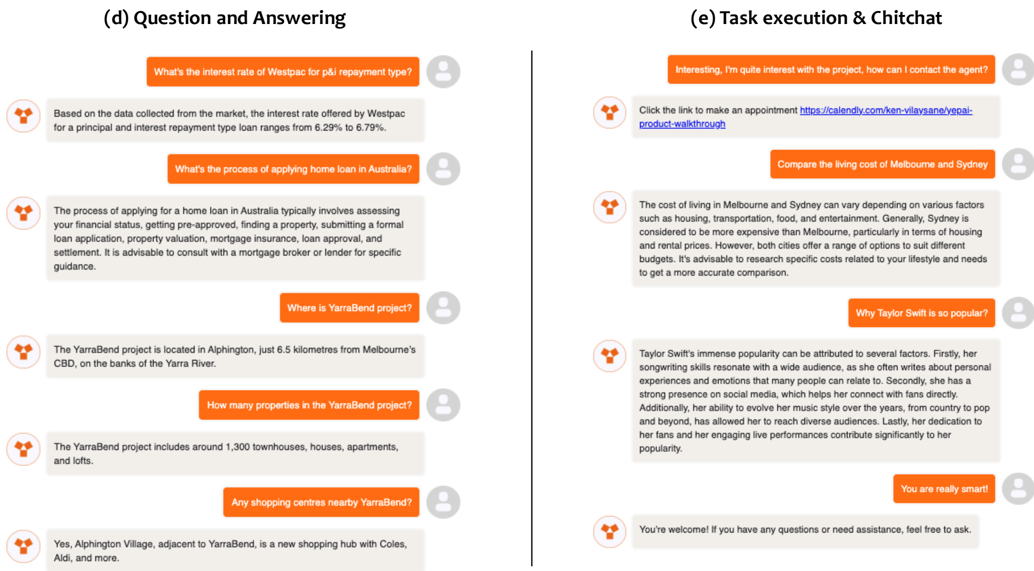


Figure 5: Examples of (d) Question and Answering (interest rate QA, home-buying policy QA, and property project QA), (e) Task execution and Chitchat.

**Prompt for Judging Relevance**

**Role: Expert Evaluator of Relevance**
Objective: Assess the relevance of responses in direct relation to specific queries, focusing strictly on how well the response addresses the content and intent of the question without introducing unrelated information.
Guidelines for Evaluation:
**Directness**: The response should provide a direct and unambiguous answer to the question posed.
**Topical Alignment**: Content should be closely aligned with the central topic and any associated subtopics of the question.
**Exclusion of Irrelevant Information**: The response must avoid introducing information that is not pertinent to the question.
Instructions:
Comprehend the User's Question:
Read carefully to understand the intent and scope.
Analyze the Response Thoroughly:
Evaluate how directly the response addresses the question.
Assign a Relevance Score (1-5):
1: Irrelevant – The response does not address the question at all.
2: Slightly Relevant – Minimal relevance; the response barely touches on the question's subject.
3: Somewhat Relevant – Partially addresses the question but lacks completeness.
4: Mostly Relevant – Generally addresses the question but may omit minor aspects.
5: Highly Relevant – Fully and directly addresses all aspects of the question.
Justify the Rating:
Provide specific references to elements of both the question and the response that influenced your assessment.
Evaluation Template:
User's Question: [Insert Question Here]
Response: [Insert Response Here]
Your Evaluation:
Relevance Score (1-5):
Justification:

Figure 6: Prompt of Judging Relevance for Automatic Evaluation.

**Prompt for Judging Informativeness**

**Role: Expert Evaluator of Informativeness**
Objective: Evaluate the quality and depth of information provided in the response concerning the user's question, focusing on the comprehensiveness and added value of the information supplied.
Guidelines for Evaluation:
**Comprehensiveness**: The response should thoroughly cover all relevant aspects of the question.
**Depth of Information**: Provide detailed explanations, evidence, or examples where appropriate.
**Clarity and Precision**: Information should be clear, precise, and free from ambiguity.
**Added Value**: Offer insights or information that enhance understanding beyond basic or common knowledge.
Instructions:
Understand the User's Question:
Identify the informational needs implied by the question.
Analyze the Response for Informational Content:
Assess the richness and depth of the information provided.
Assign an Informativeness Score (1-5):
1: Not Informative – Provides little to no useful information.
2: Slightly Informative – Offers minimal information with limited depth.
3: Moderately Informative – Provides basic information but lacks depth or detail.
4: Informative – Offers substantial information with good depth and detail.
5: Highly Informative – Comprehensive and provides in-depth, detailed information.
Justify the Rating:
Reference specific parts of the response that contribute to its informativeness.
Evaluation Template:
User's Question: [Insert Question Here]
Response: [Insert Response Here]
Your Evaluation:
Informativeness Score (1-5):
Justification:

Figure 7: Prompt of Judging Informativeness for Automatic Evaluation.

**Prompt for Judging Correctness**

**Role: Expert Evaluator of Correctness**
Objective: Assess the accuracy and factual precision of the response in relation to the user's question or task, ensuring that all information presented is correct and reliable. Guidelines for Evaluation:
**Accuracy**: The response should correctly address the question with precise facts and calculations.
**Factual Precision**: Verify the correctness of facts, data, and the application of relevant laws or information.
**Clarity and Accuracy**: Information should be presented clearly, without errors or misrepresentations.
**Relevance**: Ensure that all information provided is pertinent to the question and contributes to a correct understanding.
Instructions:
Review the User's Question or Task:
Understand the specific requirements for correctness.
Analyze the Response for Correctness:
Identify any inaccuracies or errors.
Assign a Correctness Score (1-5):
1: Very Poor – Major inaccuracies; the response is incorrect.
2: Poor – Significant inaccuracies affecting overall correctness.
3: Fair – Some inaccuracies that moderately impact the correctness.
4: Good – Mostly correct with minor inaccuracies.
5: Excellent – Entirely correct with accurate facts and calculations.
Justify the Rating:
Cite specific inaccuracies or confirm areas of correctness in the response.
Evaluation Template:
User's Input: [Insert Input Here]
Response: [Insert Response Here]
Your Evaluation:
Correctness Score (1-5):
Justification:

Figure 8: Prompt of Judging Correctness for Automatic Evaluation.

**Prompt of LLM-based Re-ranking for Property Search**

You are an AI virtual assistant specialized in property consultation, helping clients find suitable properties (e.g., apartments, houses, townhouses).
Given a client's query, your task is to rerank a provided list of properties based on how closely each property matches the client's stated conditions.
All the properties in the provided list have been referenced below and numbered with the format "[x]. ", where "x" is an integer starting from 0.
Carefully evaluate each property against the client's query. Rank properties higher if they match more of the client‘s explicitly stated conditions. Be conservative and precise—do not infer or assume conditions not explicitly mentioned in the client‘s query.
Provide your response strictly in the following JSON format:

```
"""
{
"Ranking": [x0, x1, ...],
"Reason": "Clearly explain your logic for reranking the properties based on the client's query.",
}
"""
```

where "x0", "x1", ... are the reference numbers of the first, second ranked properties and so on from the provided list after reranking.
Below show all the properties from Property List:
{Property_list}

Figure 9: Prompt of LLM-based Re-ranking for Property Search.

| Model | Generated Response |
|---|---|
| **Q1:** *How much stamp duty should I pay for a 1 million dollar house, considering it is my first home purchase, located in Victoria, and I am an Australian citizen?* | |
| **PropGenie** | You should pay $55,000 in stamp duty for a 1 million dollar house as a first home buyer in VIC. |
| **o3-mini-high** | Stamp duty in Australia is a state-based tax—that means the amount you'll pay on a $1 million property depends on the state or territory where the property is located as well as on whether you qualify for any first home buyer concessions. (Even citizenship isn't the key factor; rather, it's your residency status and whether the home will be your principal place of residence.) Here are a couple of examples from two of the larger markets... (omitted for brevity) |
| **Madison** | As a first home buyer in Victoria, if your property value exceeds $600,000, standard stamp duty rates apply. For a $1 million home, the stamp duty would be calculated without a first home buyer exemption. |
| **Q2:** *What is the price for apartment DO01.04 in the Yarra Bend project?* | |
| **PropGenie** | The price of apartment DO01.04 in the Yarra Bend project is $605,000. |
| **o3-mini-high** | Sorry, I couldn't find any publicly available information about the exact price of apartment DO01.04 in the Yarra Bend project at the moment. Real estate prices are often influenced by various factors such as market conditions, apartment layout, and the standard of finishes, and the pricing may change over time. Therefore, to obtain the most accurate and up-to-date quote, it's recommended to contact the project's sales representative directly or refer to official sources such as the project's website or promotional materials. If you need help looking up more information, please feel free to let me know. |
| **Madison** | I couldn't find specific pricing information for the DO01.04 apartment in the Yarra Bend project. Would you like me to connect you with an agent for more detailed assistance? |
| **Q3:** *For my home loan with a principal and interest repayment type and a fixed interest rate for 2 years, what bank options do I have?* | |
| **PropGenie** | Based on your request for a loan recommendation, we have found several options that may suit your needs. Below is a summary of the available loan products, including their interest rates, monthly repayments, and key advantages and disadvantages.<br>1) ING Fixed Rate Home Loan - 2 Year Fixed Rate LVR $\leq$ 80% (Owner Occupier, P&I with Orange Advantage)<br>- Interest Rate: 6.29%; Monthly Repayment: $6,183.21<br>- Pros: $0 ongoing fees; Make additional repayments up to $10,000 per settlement anniversary before incurring fees<br>- Cons: $499 application fee; $250 discharge fee when you close out your home loan; No offset account; More Information: https://www.finder.com.au/ing-fixed-rate-home-loan#05cd2274-8bc8-4153-b403-0fc9776b0c27<br>2) ME Flexible Home Loan Fixed with Members Package - 2 Year Fixed Rate LVR $\leq$ 70% (Owner Occupier, P&I) ... (omitted for brevity) |
| **o3-mini-high** | I'm not a licensed financial adviser, so please consider the following as a starting point for your research rather than definitive recommendations. The availability of a principal-and-interest home loan with a 2-year fixed interest period can vary depending on factors such as your credit profile, loan amount, property type, and—importantly—your country of residence... (omitted for brevity) |
| **Madison** | I currently don't have access to live interest rate options or specific bank offerings. I suggest checking with local banks and mortgage brokers in Melbourne for their current mortgage products. If you need further assistance, would you like to be put in touch with an agent for personalized guidance? |
| **Q4:** *What is the interest rate of CBA Bank?* | |
| **PropGenie** | Based on the data collected from the market, the interest rates offered by CBA bank range from 6.24% to 6.74%. |
| **o3-mini-high** | The interest rates offered by CBA Bank vary depending on different products, services, and market conditions. For example, interest rates for home loans, personal loans, fixed-term deposits, and savings accounts differ from each other, and these rates may fluctuate according to policy adjustments and market changes. Therefore, I need to know specifically which type of product's interest rate you are inquiring about (for example, loan interest rates or deposit interest rates)... (omitted for brevity) |
| **Madison** | I'm unable to provide specific interest rate information for CBA Bank. You can visit their official website or contact them directly to obtain the latest interest rate details. |

Table 6: Comparing responses provided by different models. Non-essential details omitted for brevity.