Language Models as Knowledgeable Agents: A Research Proposal

Min (Henry) Cai
github@HenryCai11
caimin2021@email.szu.edu.cn

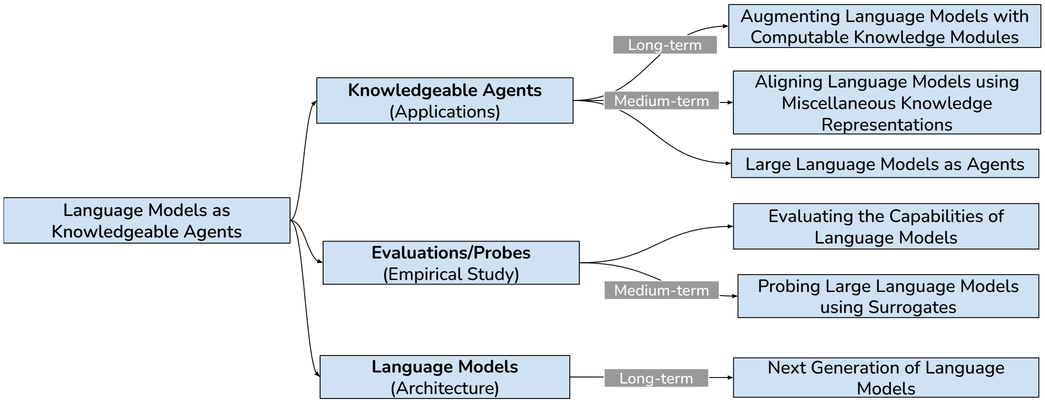
Language Models as Knowledgeable Agents
Knowledgeable Agents (Applications)
Evaluations/Probes (Empirical Study)
Language Models (Architecture)
Long-term
Medium-term
Medium-term
Long-term
Augmenting Language Models with Computable Knowledge Modules
Aligning Language Models using Miscellaneous Knowledge Representations
Large Language Models as Agents
Evaluating the Capabilities of Language Models
Probing Large Language Models using Surrogates
Next Generation of Language Models

Figure 1: Overview of the research plan.

Abstract

The unparalleled surge caused by Large Language Models (LLMs) is changing the field quickly. Some of the traditional sub-tasks in natural language processing (NLP) seemingly become too trivial to evaluate on. Also, with the discovery of LLMs' capability in decision-making, academia gradually resorts to finding more complex and realistic tasks to test their ability. However, even though LLMs have demonstrated such strong power in handling tasks solely with language generation, there also occurs various issues and challenges. One of them is hallucination in factual knowledge, especially when they are asked to generate the latest knowledge, which is not likely to be known due to the limitations in update frequency. My research proposal starts from these issues and challenges, and depicts a detailed plan which aims at filling the research gaps during the future Ph.D study.

1 Introduction

Recent events in academia and industry ranging from ChatGPT to Generative Agents, are driving a paradigm shift not only in NLP communities, but also in the broader field of Artificial Intelligence. Generative models have shown strong capabilities in various traditional NLP tasks, even ones that seemingly do not align well with their generation nature, e.g., tasks in structure prediction [68, 67, 40, 86]. Moreover, recent work also discover that LLM possesses the ability of decision-making, as demonstrated through testing on several benchmarks [76, 39, 75]. However, despite the strong empirical results obtained from the benchmarks, current generative models have inherent flaws that are difficult to eliminate, e.g., hallucination. Additionally, existing benchmarks provide only a narrow view of the capabilities of these generative models. Thus, the community is also calling for comprehensive evaluations. Last but not least, even though most generative models are powerful, the cost to train and maintain them is expensive, especially for large language models (e.g., models with over 7B parameters). Nevertheless it is still inspiring to see that the number of neurons in human brains is much smaller than the number of parameters in a LLM, which means there may possibly be architectures that can perform comparable performance whereas requires fewer parameters [72].

Building on these findings, I propose a detailed research plan, aiming at developing an end-to-end system that uses language models to automatically learn, retrieve and leverage miscellaneous knowledge. This also requires the system to plan, take actions and interact with a world model [3], where all the knowledge comes from. To achieve this, conducting various empirical studies to gain insights into language models will be valuable, as it may inform future research. I divide this research plan into several parts, also along with a rough schedule (also see Figure 1):

1. **Knowledgeable Agents:** For knowledge-enhanced language models, I take it from two perspectives: 1) directly aligning LMs with human preference using appropriate knowledge representations, and 2) designing differentiable knowledge modules that can be grounded on by LMs. Apart from grounding language models with factual knowledge, I also plan to discover knowledge that can improve the LLMs' ability of decision-making.
2. **Evaluations/Probes:** 1) Design comprehensive metrics to evaluate LMs, and 2) Probe LMs to get deeper insights, especially for LLMs for which model checkpoints are not accessible.
3. **Better Language Models:**[1] Explore language models with fewer parameters but comparable performance.

2 Related Work and Research Gap

2.1 Knowledge-Augmented Language Models

Numerous studies have been conducted to inject knowledge into pre-trained language models. Most of these studies focus on incorporating factual knowledge obtained from knowledge graphs, e.g., Freebase, Wikidata [37, 20, 51, 85, 65, 19, 38, 71], while some focus on leveraging task-related knowledge [35, 13, 78, 83]. However, most of these works heavily rely on transformer encoder blocks to obtain contextualized embeddings for knowledge injection, while few employ pure decoder-only backbones [24].

Although a very relevant line of work, Retrieval-Augmented Generation (RAG), alleviates this issue by disentangling the process of encoding and generation, most of the backbones themselves have not actually learned the knowledge; they have only acquired the ability to retrieve it [31, 32, 49, 28, 80, 17, 61, 8]. Although this is seemingly a higher-order of ability, the way of retrieving is still primitive and require further consideration. Another line of work in Knowledge Editing is also relevant [45, 46, 41], where they focus on modifying behaviours of models.

Also, recent improvements in **Reinforcement Learning with Human Feedback** (RLHF) have shown to be very effective in tuning LMs to generate content that we desired. Even though methods in this field are still primitive [2, 60], it is promisingly an effective tool to fine-tune and teach models to acquire various abilities.

I see the goal in this topic as looking for a most suitable way to inject, update knowledge in language models, and empower them to retrieve and leverage all sorts of knowledge, at the lowest cost of degradation of their original capabilities.

2.2 Language Models as Agents

The utility of knowledge and abilities varies depending on the situation. Thus, it is also important to decide which knowledge resource the system should resort to and what abilities best fit the current scenario. All of these can be considered as the problem of decision-making, which requires an AI system to interact with the environment and make proper decisions on the fly.

[1] I will not go into further details for this topic in this proposal. But here is a reference link that may be of interest to you and for your information. https://babylm.github.io/

Recent studies have demonstrated that large language models (LLMs) possess strong abilities in memorization, planning, tool use, and even communication with other agents [48, 74]. These discoveries are rapidly changing the game in the NLP community, where some of the traditional tasks have been reformed and can be well solved simply with prompt engineering [82].

These findings and improvements can serve as building blocks for implementing a system that can self-improve through interaction with a *world model* [3, 6], using language models as backbones.

2.3 Evaluations and Probes on Language Models

Empirical studies of language models are becoming increasingly important, as the insights gained can guide refinements in LM development. There are various kinds of materials that we can study, from simply the outputs of LMs [55, 63], to embeddings in specific layers (geometry) [73]. Techniques used in the field are also various, from prompt engineering [55, 25], to probing geometric features of embeddings [73].

However, the probing and evaluation methods for cloud-based large language models, for which parameters are not accessible, remain primitive. These prompt-based approaches require intensive manual effort, hindering more efficient exploration and deeper understanding of these LLMs. Inspired by works on affecting LLMs using surrogates [77], I deem that we can affect LLMs to behave in the manners that we desired, and then probe the surrogate models, whose parameters are accessible, to see what cause the LLMs to behave in such ways.

Additionally, it would be insightful to examine how different methods of knowledge injection affect the overall capabilities of LMs, beyond just knowledge-intensive tasks.

3 Research Plan

In this section, I display a more detailed research plan that includes all the ingredients that I need and the rough timeline for a single project.

3.1 Knowledgeable Agents

Ingredients	Familiar	Literature
Knowledgeable Agents, ∼4-6 months/project		
Knowledge-Augmented Language Models	High	[37, 15, 19, 20, 24, 26, 38, 51, 59, 65, 70, 71, 79, 85, 81]
Retrieval-Augmented Generation	High	[31, 32, 49, 28, 80, 17, 61, 8]
Language Models as Agents	Medium	[76, 39, 75, 27, 54, 29, 58, 21, 57, 47, 10, 56]
Knowledge Editing	Low	[45, 46, 41, 9, 30, 22, 5, 44, 53, 43, 7, 18]
Reinforcement Learning with Human Feedback	Low	[2, 60, 48, 74, 6, 14, 1, 88, 36, 64, 34, 11, 69, 42]

Table 1: Table of a plan on **Knowledgeable Agents** with the estimated time duration for a single project on this topic, its ingredients, related literature, and other resources.

Table 1 compiles all the relevant literature for each topic. As for the estimated time duration of a single project, I put a detailed timeline/to-dos in the appendix. I've broken it down into actionable items that can serve as a template for quickly initiating the project. Since I have already conducted most of the survey work in advance, it is not included in the template. Additionally, in Table 1, I've indicated the extent to which I am familiar with each topic. This can serve as a guide for areas where further surveying may be needed during the project.

It's worth noting that this project may require significant computational resources, which could make setting up the experimental environment time-consuming. This includes hardware configuration and engineering work to accelerate experiments.

3.2 Evaluation and Probe

The project on **Evaluation and Probe** differs significantly from **Knowledgeable Agents** in that the methodologies and ideas are quite diverse and inconsistent. This requires substantial effort in

Ingredients	Familiar	Literature
Evaluation/Probe, ∼3-10 months/project		
Pre-trained Language Models	Medium	[52, 66, 62, 50, 55, 87, 73, 25, 12, 23, 33]
Large Language Models	Medium	[16, 4, 63, 84]
Reinforcement Learning with Human Feedback	Low	[2, 60, 48, 74, 6, 14, 1, 88, 36, 64, 34, 11, 69, 42]

Table 2: Table of a plan on **Knowledgeable Agents** with the estimated time duration for a single project on this topic, its ingredients, related literature, and other resources.

conducting surveys before initiating each project. The survey should include not only literature on methodologies from NLP but also consider interdisciplinary studies when relevant, e.g., literature from psychology and linguistics. Relevant literature can also be found in Table 2

For instance, factors such as crowd-sourcing for data collection and cleaning, diachronic studies [63] that may take a considerable amount of time, and questionnaires for user studies should also be considered. Therefore, the estimated time duration may vary depending on the specific research topic.

A template for a quick start is also shown in the appendix, where I list all the possible factors that may be considered in future studies.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023.

[4] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms, 2023.

[5] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models, 2021.

[6] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug, 2023.

[7] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers, 2022.

[8] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay-Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. Case-based reasoning for natural language queries over knowledge bases, 2021.

[9] Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. Calibrating factual knowledge in pretrained language models, 2022.

[10] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[11] Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. Compositional semantic parsing with large language models, 2022.

[12] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models, 2020.

[13] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*, 2020.

[14] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation, 2023.

[15] Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. Entities as experts: Sparse memory access with entity supervision. *arXiv preprint arXiv:2004.07202*, 2020.

[16] Michael C Frank. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, pages 1–2, 2023.

[17] Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5770–5793, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[18] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, 2022.

[19] Ziniu Hu, Yizhou Sun, and Kai-Wei Chang. Relation-guided pre-training for open-domain question answering. *arXiv preprint arXiv:2109.10346*, 2021.

[20] Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. Empowering language models with knowledge graph reasoning for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9581, 2022.

[21] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022.

[22] Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. Transformer-patcher: One mistake worth one neuron, 2023.

[23] Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, 2022.

[24] Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. Language generation with multi-hop reasoning on commonsense knowledge graph. *arXiv preprint arXiv:2009.11692*, 2020.

[25] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know?, 2020.

[26] Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. *arXiv preprint arXiv:2106.10502*, 2021.

[27] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models, 2023.

[28] Tian Lan, Deng Cai, Yan Wang, Heyan Huang, and Xian-Ling Mao. Copy is all you need, 2023.

[29] Jack Lanchantin, Shubham Toshniwal, Jason Weston, Arthur Szlam, and Sainbayar Sukhbaatar. Learning to reason and memorize with self-notes, 2023.

[30] Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. Plug-and-play adaptation for continuously-updated QA. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 438–447, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[31] 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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[32] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation, 2022.

[33] Jiaxi Li and Wei Lu. Contextual distortion reveals constituency: Masked language models are implicit parsers, 2023.

[34] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[35] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*, 2019.

[36] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback, 2023.

[37] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of AAAI 2020*, 2020.

[38] Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425, 2021.

[39] Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents, 2023.

[40] Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. Prompt for extraction? paie: Prompting argument interaction for event argument extraction, 2022.

[41] Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve gpt-3 after deployment, 2023.

[42] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes, 2023.

[43] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

[44] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale, 2022.

[45] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale, 2022.

[46] Shikhar Murty, Christopher D. Manning, Scott Lundberg, and Marco Tulio Ribeiro. Fixing model bugs with natural language patches, 2022.

[47] Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling, 2023.

[48] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[49] Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. Retrieval augmented code generation and summarization, 2021.

[50] Christian S. Perone, Roberto Silveira, and Thomas S. Paula. Evaluation of sentence embeddings in downstream and linguistic probing tasks, 2018.

[51] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019.

[52] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[53] Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules, 2021.

[54] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models, 2023.

[55] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020.

[56] Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[57] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models, 2023.

[58] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents, 2023.

[59] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. Colake: Contextualized language and knowledge embedding. *arXiv preprint arXiv:2010.00309*, 2020.

[60] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision, 2023.

[61] Dung Thai, Srinivas Ravishankar, Ibrahim Abdelaziz, Mudit Chaudhary, Nandana Mihindukulasooriya, Tahira Naseem, Rajarshi Das, Pavan Kapanipathi, Achille Fokoue, and Andrew McCallum. Cbr-ikb: A case-based reasoning approach for question answering over incomplete knowledge bases, 2022.

[62] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022.

[63] Shangqing Tu, Chunyang Li, Jifan Yu, Xiaozhi Wang, Lei Hou, and Juanzi Li. Chatlog: Recording and analyzing chatgpt across time, 2023.

[64] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022.

[65] Pat Verga, Haitian Sun, Livio Baldini Soares, and William W Cohen. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. *arXiv preprint arXiv:2007.00849*, 2020.

[66] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics, 2020.

[67] Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Revisiting relation extraction in the era of large language models, 2023.

[68] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models, 2023.

[69] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023.

[70] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

[71] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

[72] Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus, 2023.

[73] Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting bert, 2021.

[74] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf, 2023.

[75] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions, 2023.

[76] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.

[77] Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Retroformer: Retrospective large language agents with policy gradient optimization, 2023.

[78] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*, 2021.

[79] Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. Jaket: Joint pre-training of knowledge graph and language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11630–11638, 2022.

[80] Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

[81] Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. Dkplm: decomposable knowledge-enhanced pre-trained language model for natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11703–11711, 2022.

[82] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check, 2023.

[83] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. Greaselm: Graph reasoning enhanced language models. In *International conference on learning representations*, 2021.

[84] Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. Beyond positive scaling: How negation impacts scaling trends of language models, 2023.

[85] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[86] Zixuan Zhang and Heng Ji. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online, June 2021. Association for Computational Linguistics.

[87] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. Evaluating commonsense in pre-trained language models, 2021.

[88] Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023.

A Timeline/To-dos

The detailed templates are shown in the following pages.

Timeline/To-dos

Knowledgeable Agents

- □ Set Up Environment
  - □ Hard-ware Configuration
  - □ Experiment Code ~1-4 weeks in total
    - □ Implement Baseline Code
    - □ Add Test-cases
  - □ Find Proper Benchmarks ~1 week in total
  - □ Data Preparation

~2-5 weeks in total

- □ Experiment
  - □ Run Basic Experiment ~1 week
  - □ Implementation ~2-4 weeks
  - □ Further Experiments ~2-4 weeks

Implementation and Further Experiments may repeat multiple times (say, twice). Then the estimated time will be:

~9-17 weeks in total

- □ Empirical Study ~1-2 weeks
  - Case Study
  - Statistical Analysis
  - Theoretical Analysis

~1-2 weeks in total

Total estimated time: ~12-24 weeks (Provided that paper writing can be done simultaneously)

Evaluations/Probes

- □ Further Survey
  - □ Gather Tricks ~1-2 weeks
  - □ Literature Review ~1-2 weeks

~2-4 weeks in total

- □ Set Up Environment
  - □ Hard-ware Configuration
  - □ Experiment Code ~1-4 weeks in total
    - □ Find Baseline Models
    - □ Add Test-cases

~1-4 weeks in total

- □ Data Preparation (depend on the concrete method) ~1-4 weeks in total
  - Crowd-source Data
  - Automatic Data Generation
  - Use Existed Benchmarks

~1-4 weeks in total

- □ Experiment
  - □ Implementation ~1-4 weeks in total
  - □ Refinement ~1-4 weeks in total

Implementation and Refinement can also repeat several times (say, twice). Then the estimated time will be:

~4-16 weeks in total

□ Empirical Study • Case Study • Statistical Analysis • Theoretical Analysis • User Study • Diachronic Study • Miscellaneous…	~1-12 weeks in total

~1-12 weeks in total (This can vary a lot based on the concrete method)

Total estimated time: ~9-40 weeks (Provided that paper writing can be done simultaneously)