

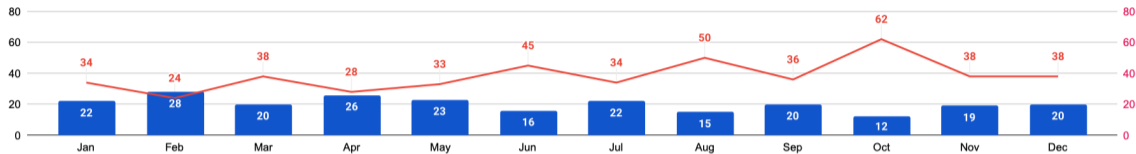
Prompting is not all you need! Or why Multi-LLM Collaboration Matters

Mirella Lapata
School of Informatics
University of Edinburgh



Spoiled for choice

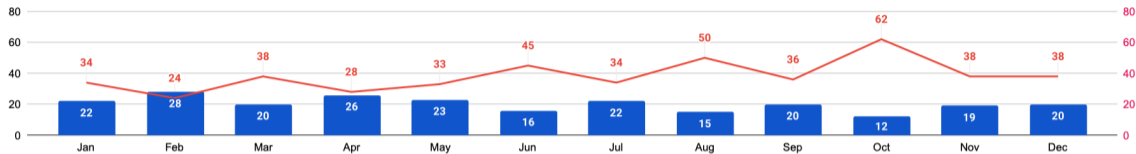
■ Per month — New model released every x hours in 2024



Source: <https://lifaichitect.ai/models/>

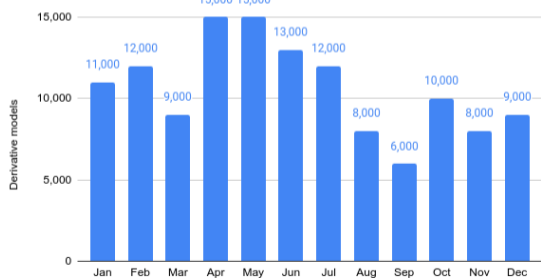
Spoiled for choice

■ Per month — New model released every x hours in 2024



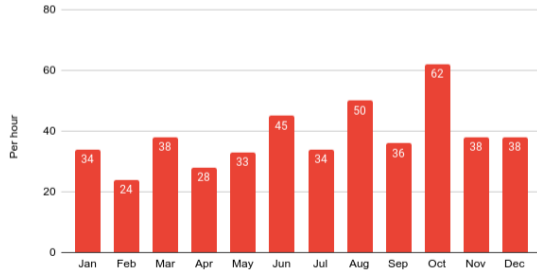
New derivative models released per month on Hugging Face

LifeArchitect.ai/models

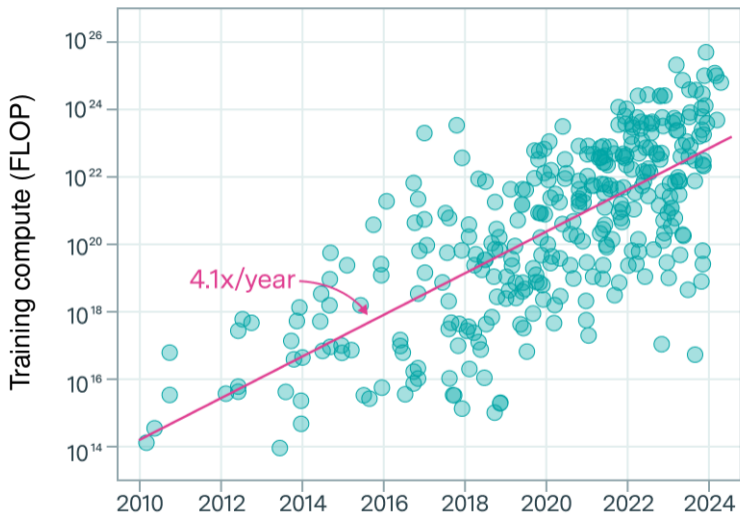


New derivative models released per hour

LifeArchitect.ai/models

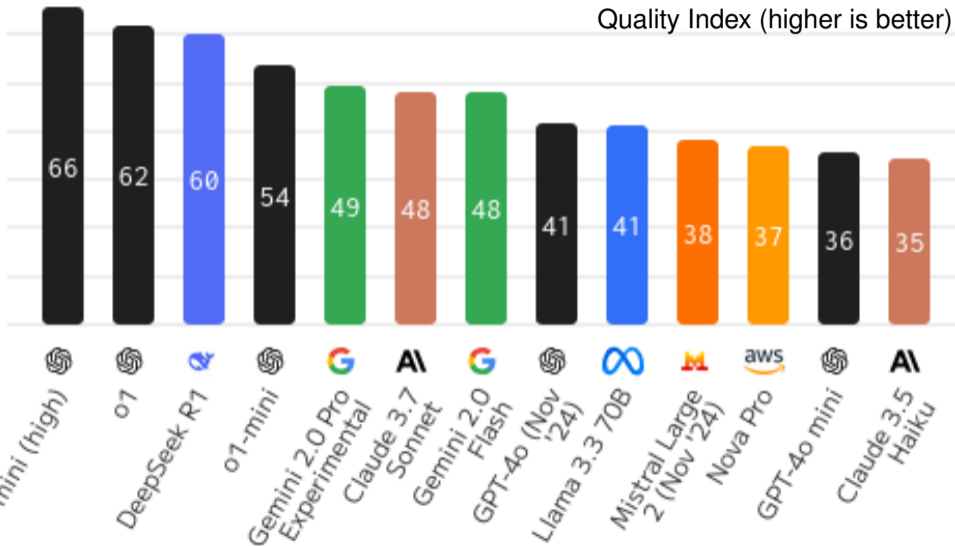


How big are these LLMs?



Source: <https://epoch.ai/>

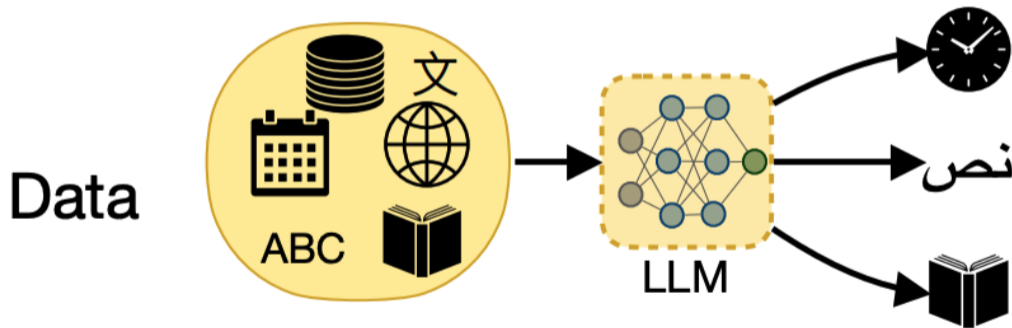
What is the Quality of LLMs?



Source: <https://lifaichitect.ai/models/>

One LLM to Rule them All?

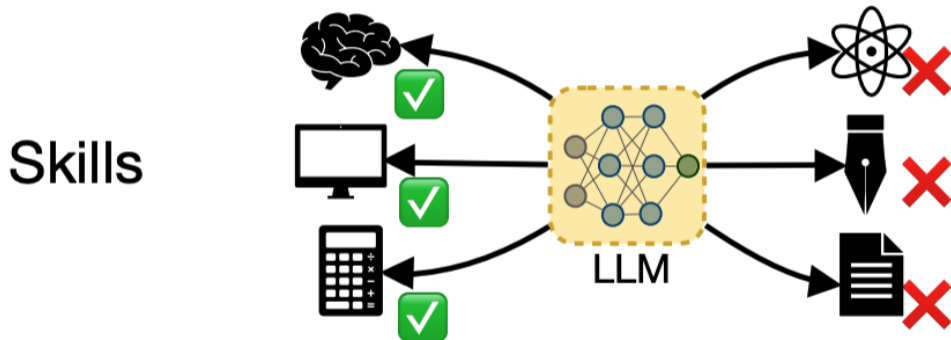
Feng et al. (2025): A single LLM is trained on a **static snapshot** of what is available, there are always elements in the real-world language distributions that are missing.



A single LLM underrepresents **real-world data** distributions (languages, domains, styles).

One LLM to Rule them All?

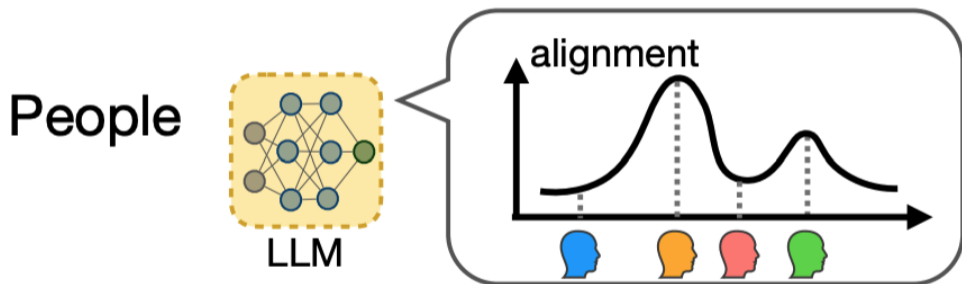
Feng et al. (2025): It is prohibitively expensive (if not impossible) to optimize for a single model that outperforms all other models on all skills; no single LLM is Pareto-optimal.



A single LLM underrepresents **skills** (reasoning, communication, creative, technical).

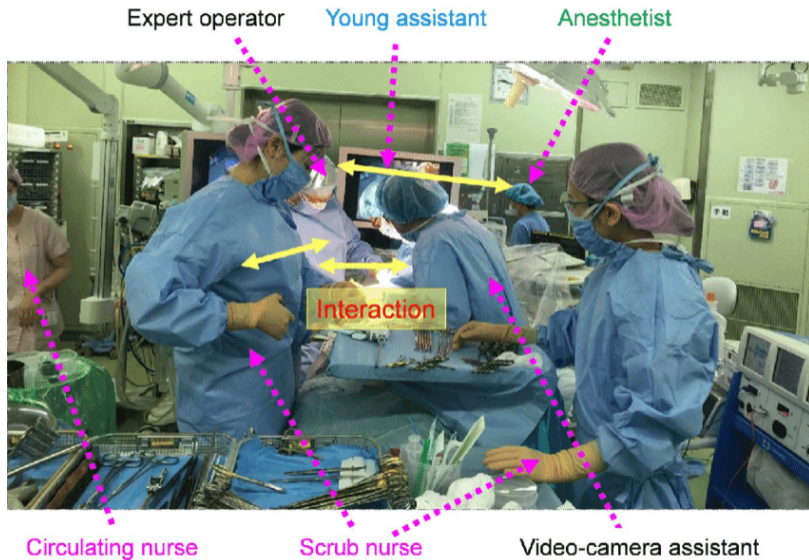
One LLM to Rule them All?

Feng et al. (2025): LLMs are ultimately used by people with diverse needs, pluralistic values, and varying socio-cultural backgrounds. LLM users are not homogeneous.

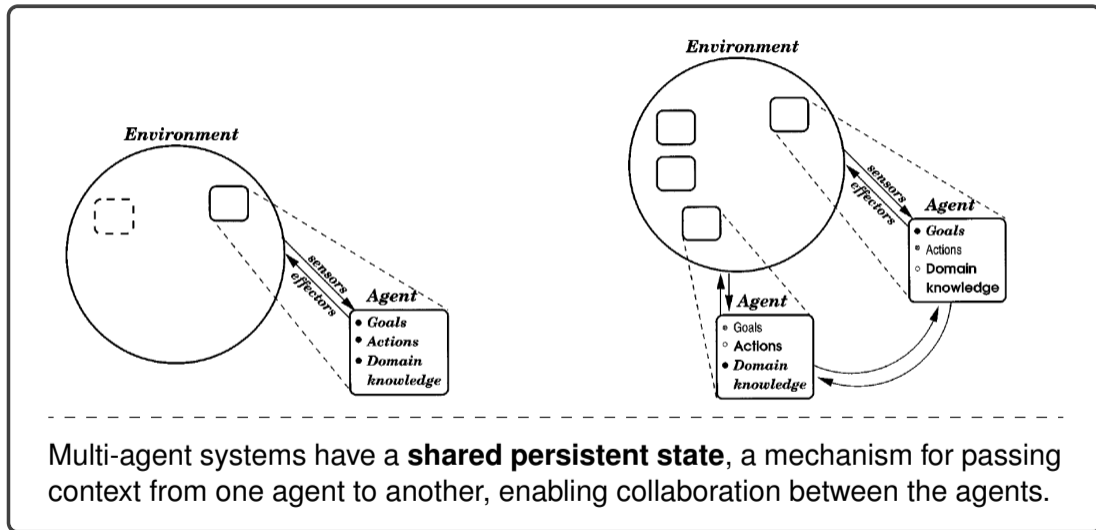


A single LLM underrepresents **people** (opinions, values, cultural norms).

A Real-life Example



Single vs Multi-Agent Systems



Benefits of Multi-Agent Systems

Single-Agent

Independence: operates autonomously without relying on other agents.

Simplicity: fewer components to integrate, easier to design first version; but struggles with new tasks and environments and may be hard to scale.

Learning from external data/feedback: happens once to improve performance over a wide range of tasks, limited by human supervision signals.

Multi-Agent

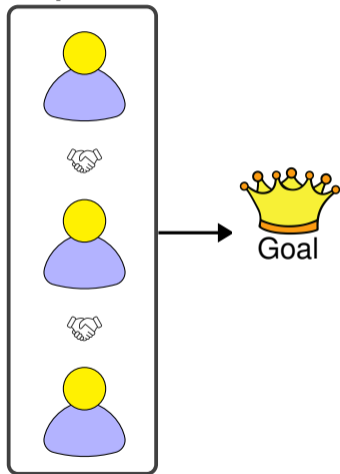
Collaboration: agents interact and collaborate to complete complex tasks.

Robustness: can handle diverse and challenging environments; increased reliability and tolerance due to distributed nature.

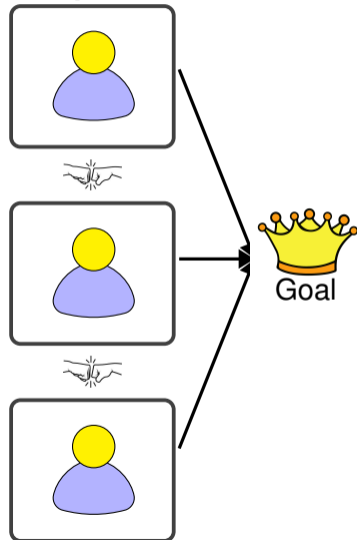
Learning based on interactions and shared experiences: happens multiple times with different task objectives; scalable training

Multi-Agent Systems Come in Many Flavors

Cooperative

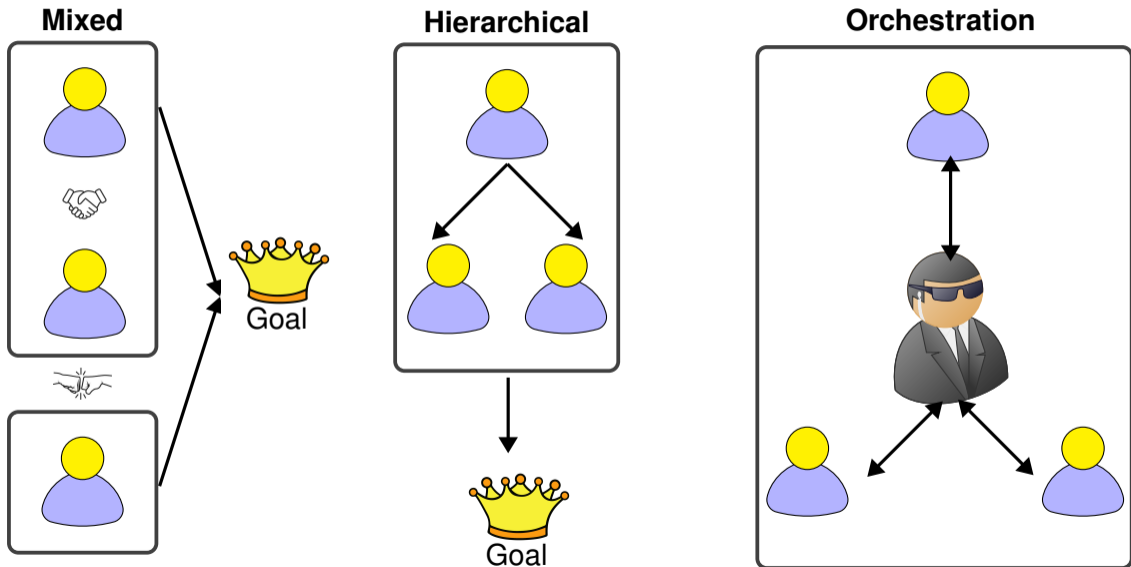


Competitive



Han et al. (2024); Tran et al. (2025); Guo et al. (2024)

Multi-Agent Systems Come in Many Flavors



Research Questions

- Which tasks are suited to multi-agent architectures?
- How are the agents specialized?
- How do they communicate, what is the shared persistent state?
- Are multi-agent systems more explainable?

- Which tasks are suited to multi-agent architectures?
- How are the agents specialized?
- How do they communicate, what is the shared persistent state?
- Are multi-agent systems more explainable?

We will use **collaborative** agents for writing tasks!

Part I

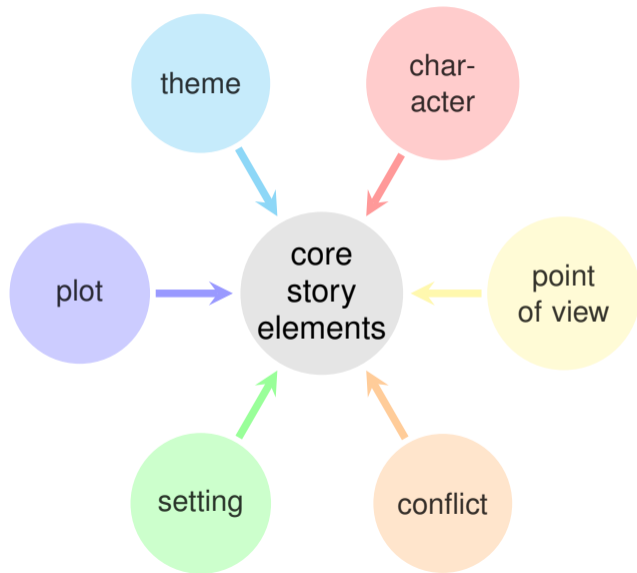
Story Generation

Fiction Writing Task: Tell Me a Story



“Please tell me a story about a puppy who goes to Jupiter on her birthday to solve a mystery and is a mermaid, using an engaging and humorous tone, in approximately a thousand words, in the style of Ernest Hemingway.”

Planning Agents: Elements of Narrative Technique

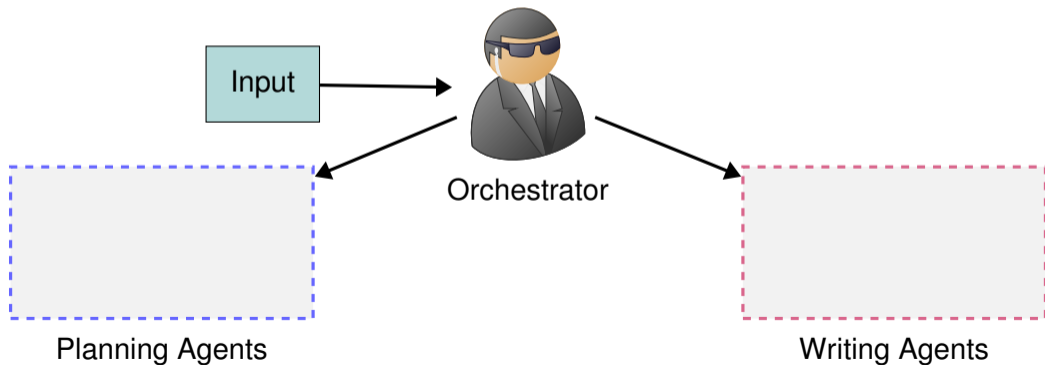


Writing Agents: Freytag's Pyramid

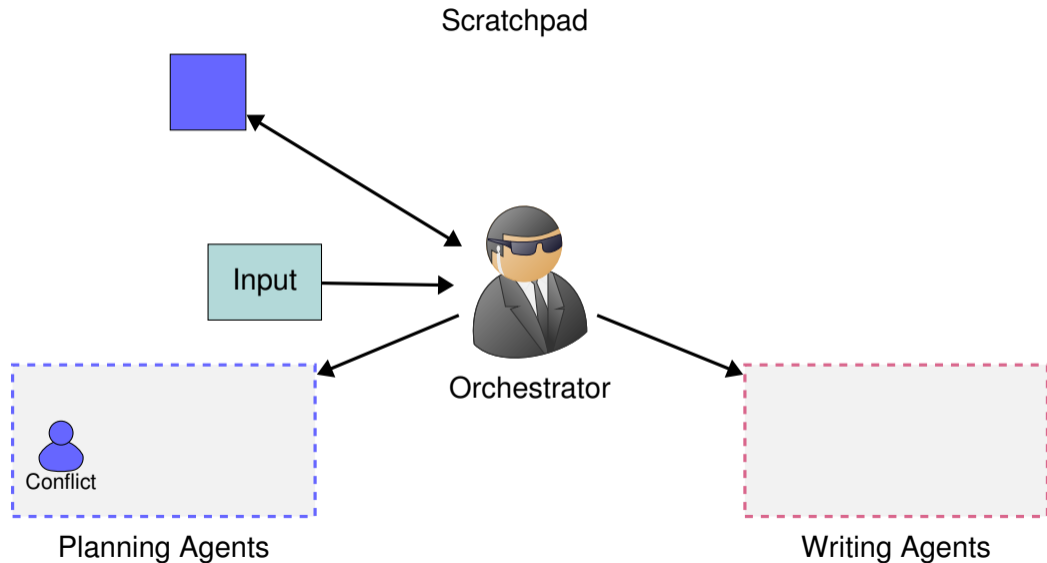


Agents' Room: Central Orchestration of Task Specialists

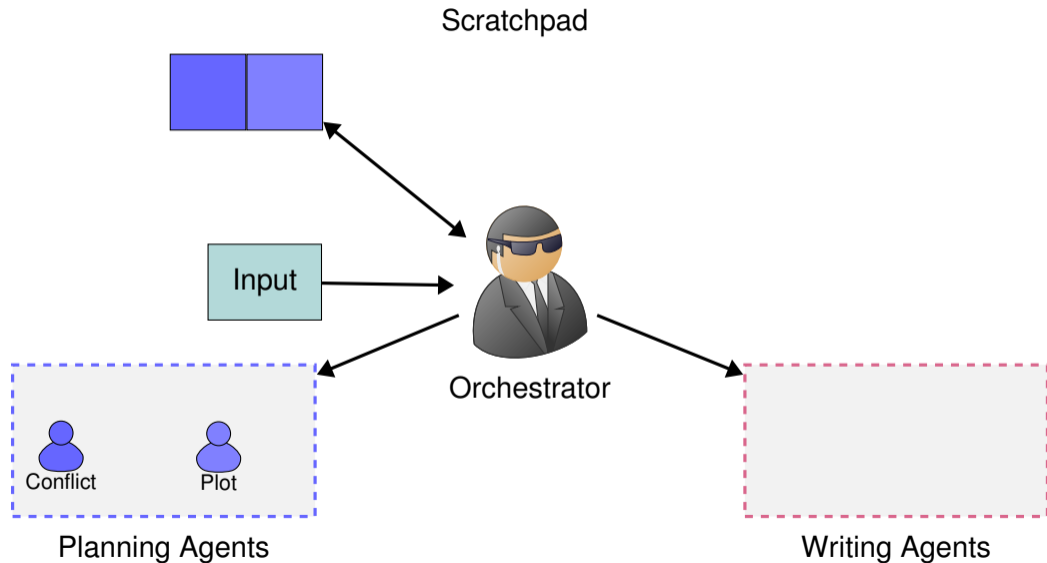
Scratchpad



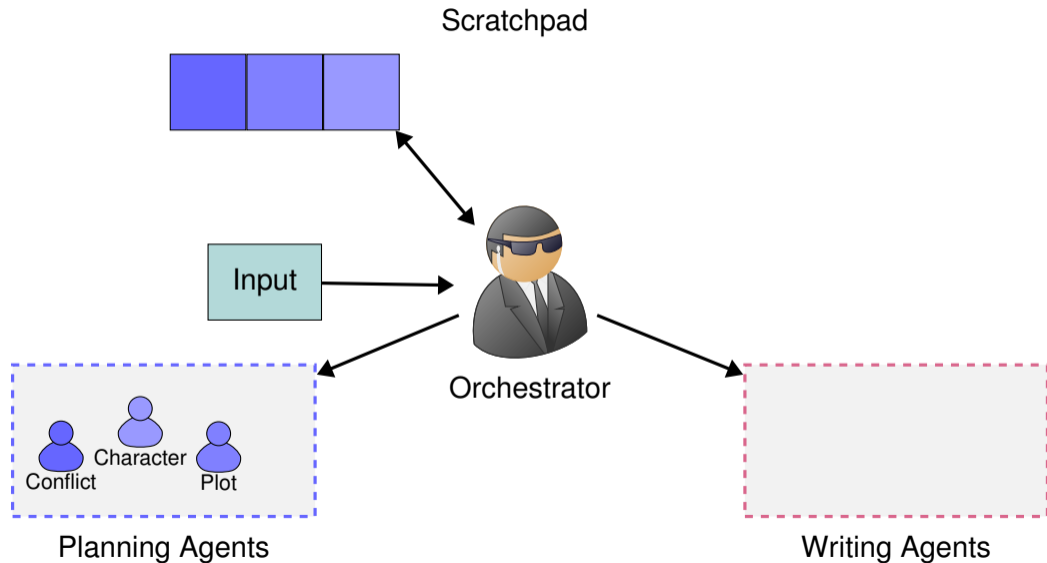
Agents' Room: Central Orchestration of Task Specialists



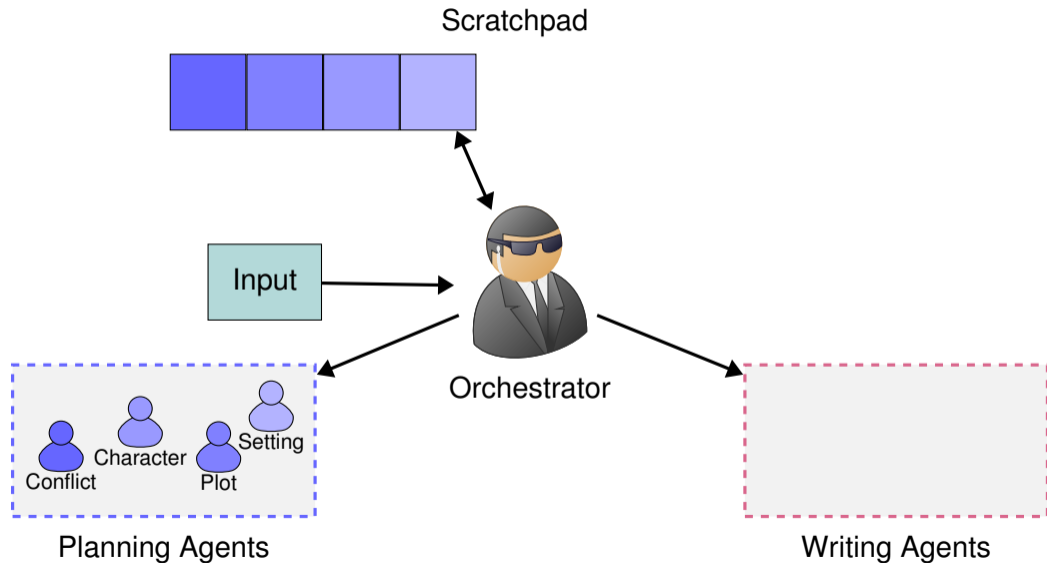
Agents' Room: Central Orchestration of Task Specialists



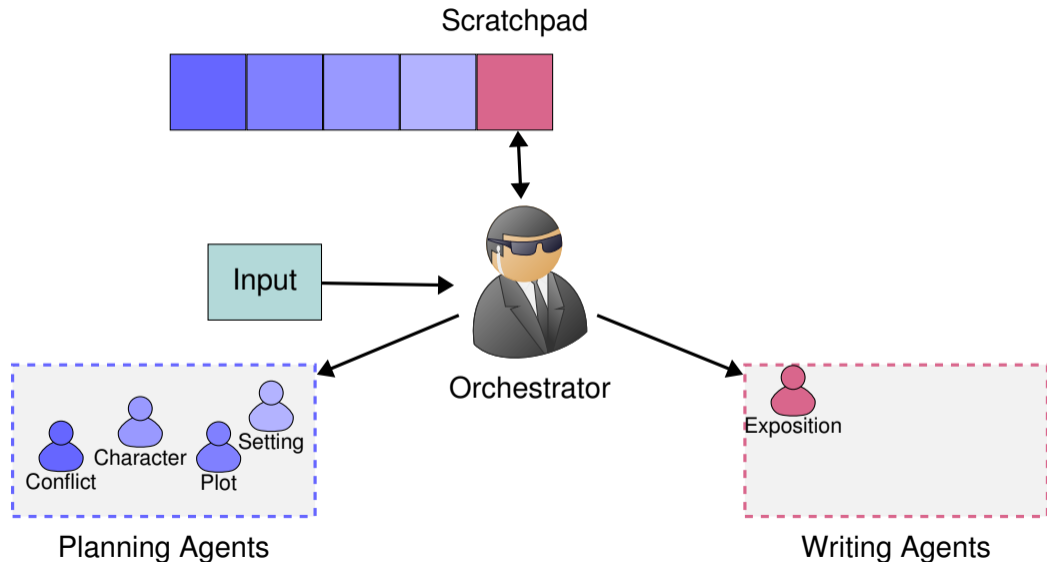
Agents' Room: Central Orchestration of Task Specialists



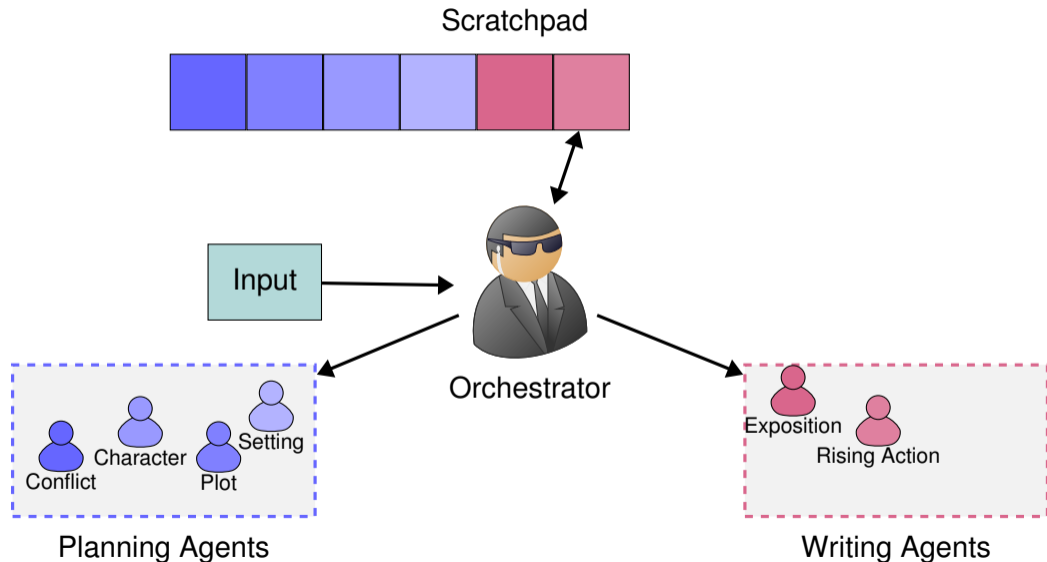
Agents' Room: Central Orchestration of Task Specialists



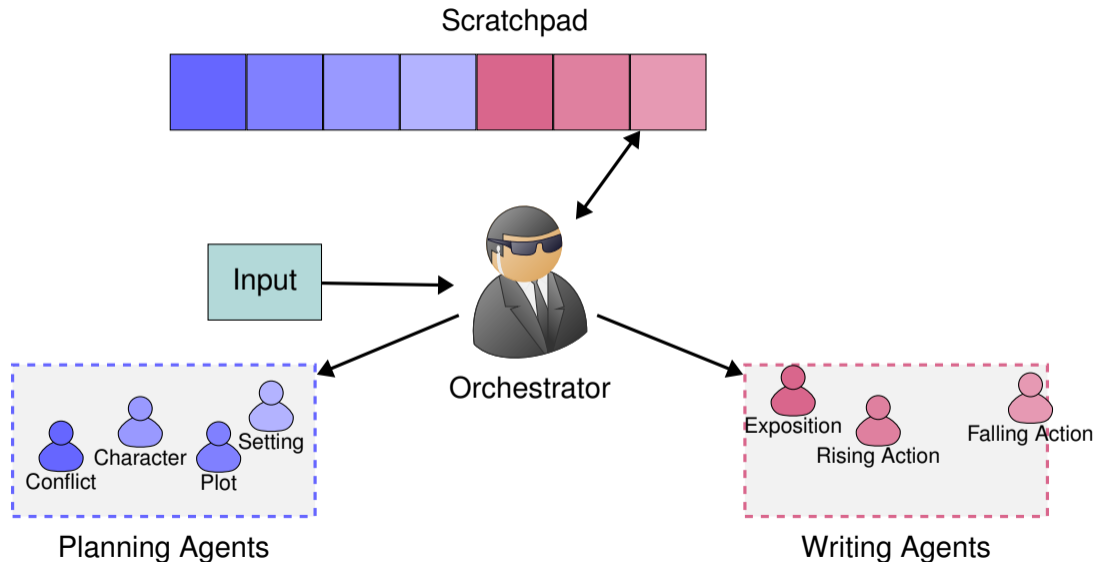
Agents' Room: Central Orchestration of Task Specialists



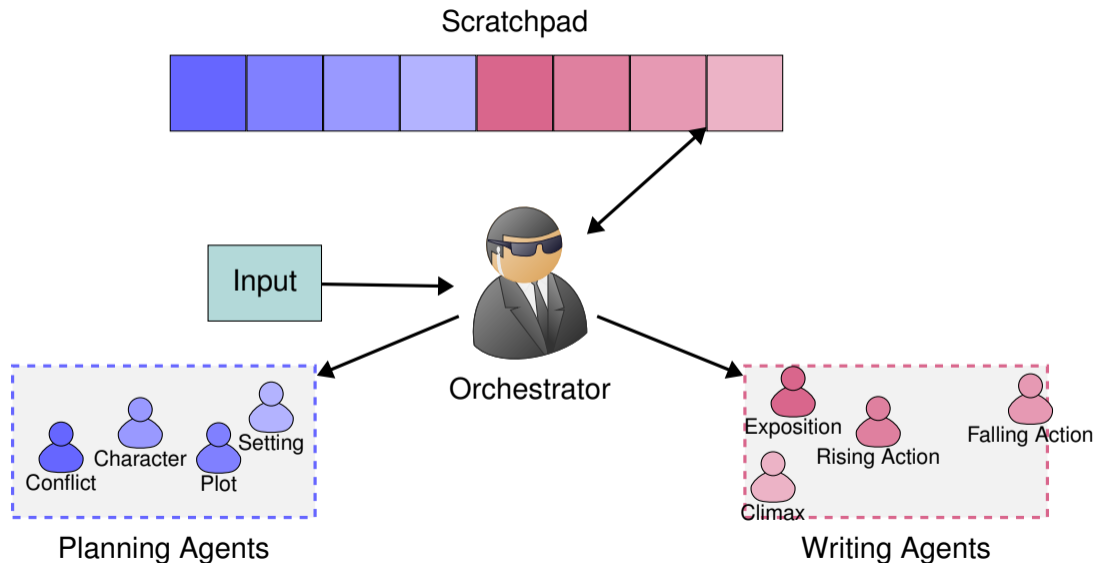
Agents' Room: Central Orchestration of Task Specialists



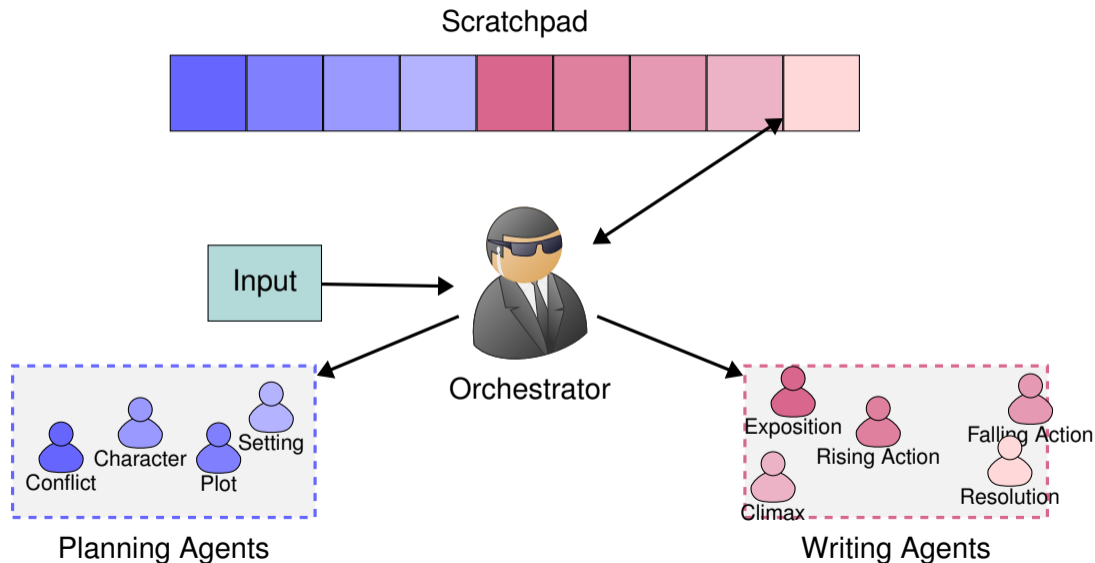
Agents' Room: Central Orchestration of Task Specialists



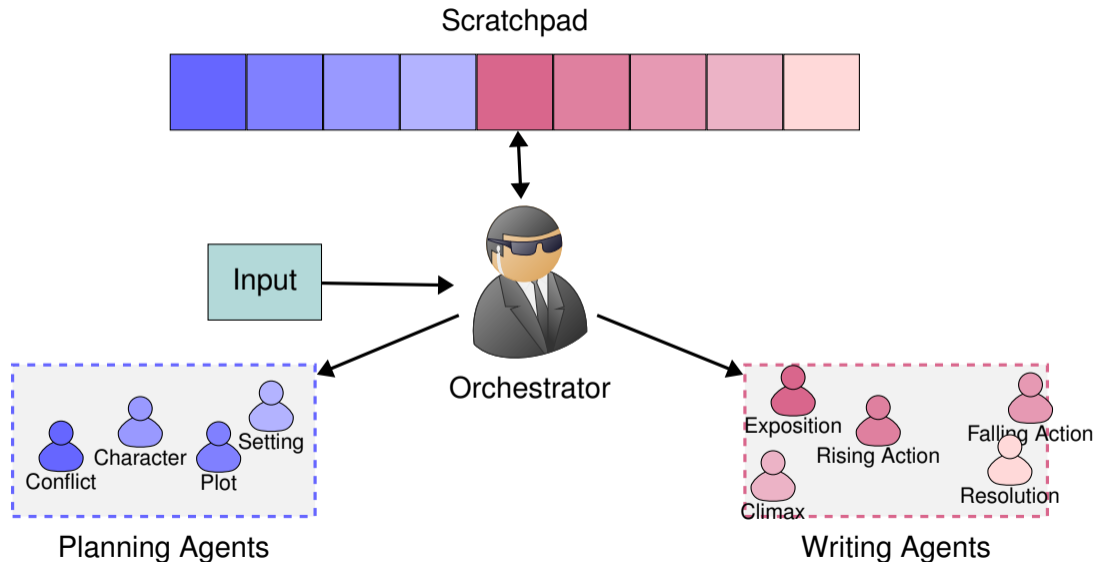
Agents' Room: Central Orchestration of Task Specialists



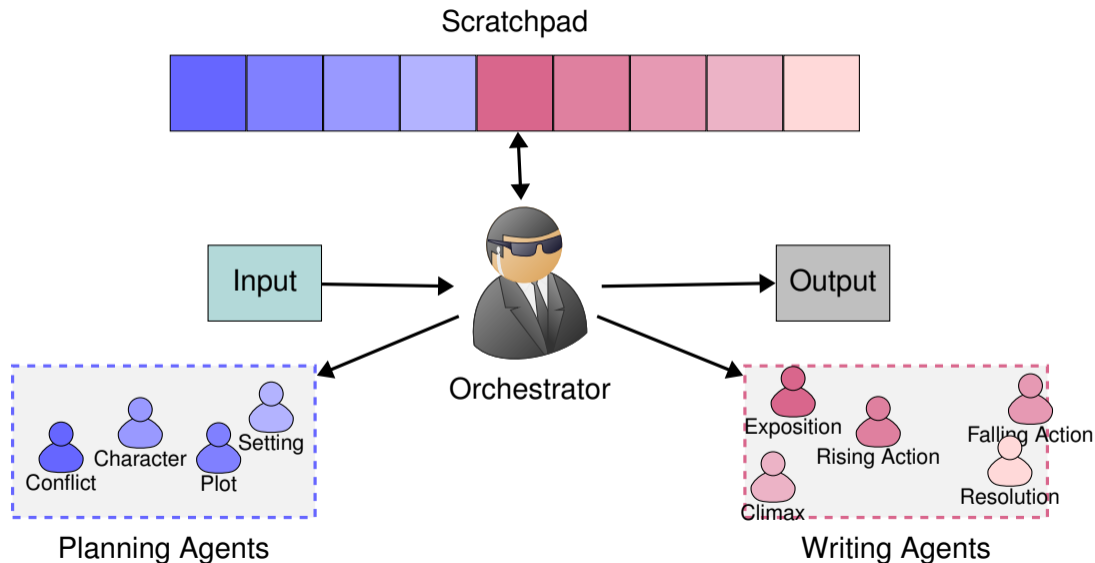
Agents' Room: Central Orchestration of Task Specialists



Agents' Room: Central Orchestration of Task Specialists



Agents' Room: Central Orchestration of Task Specialists



Agents' Room: Central Orchestration of Task Specialists

$s \leftarrow x$

while $o(s, \mathcal{A}) == \text{True}$ and $t < T$ **do**

$a_t = o(s, \mathcal{A})$

$y_t = a_t(s)$

$s \leftarrow (s; (l_t, y_t))$

if $\text{type}(a_t) == \text{writing}$ **then**

$y \leftarrow (y; y_t)$

end if

end while

return y

▷ Initialize the scratchpad

▷ While the orchestrator assigns a next agent

▷ Select an agent given scratchpad

▷ Obtain agent's output

▷ Update scratchpad

▷ If the agent is a writing agent, write to the final output

▷ Return the final output

Each agent $a \in \mathcal{A}$ takes text as input, returns text as output, is specified by unique identifier label l and a mapping $f : \mathcal{V}^* \rightarrow \mathcal{V}^*$; scratchpad $s \in \mathcal{V}^*$ shares information.

Agents' Room: Central Orchestration of Task Specialists

A multi-agent framework for **complex writing tasks** that decomposes the writing problem into smaller sub-tasks each assigned to specialized agents.

- **Communication:** collaborative, via language
- **Agents:** planning and writing, LLM based, zero-shot and fine-tuned
- **Shared Persistent State:** scratchpad
- **Orchestrator:** central deterministic

Experiments on [TellMe a Story dataset](#), elicited from professional writers, up to 1.5K tokens long, all models based on Gemini, Huot et al. (2025, ICLR).

Planning and Writing Agents

Planning Agents



Plot



Setting



Character



Conflict

Writing Agents



Exposition



Rising
Action



Climax



Falling
Action



Resolution

Planning and Writing Agents

Planning Agents



Plot



Setting



Character



Conflict

Planning Agents



Scratchpad

Writing Agents



Exposition



Rising
Action



Climax

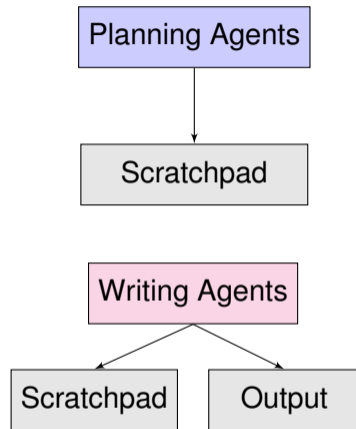
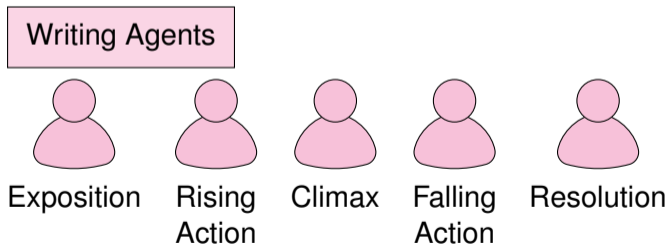
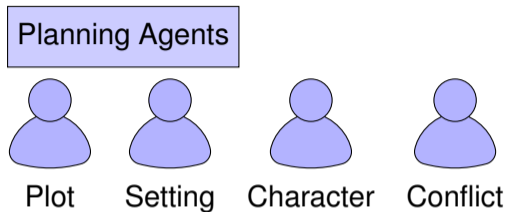


Falling
Action



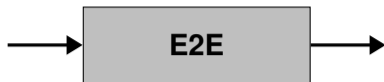
Resolution

Planning and Writing Agents

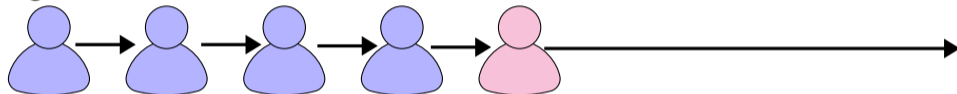


Experimental Setting and Ablations

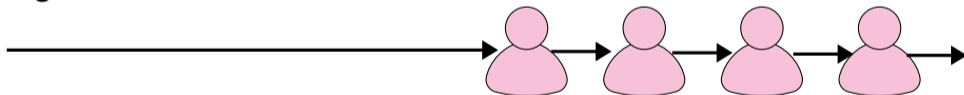
End-to-end Baseline



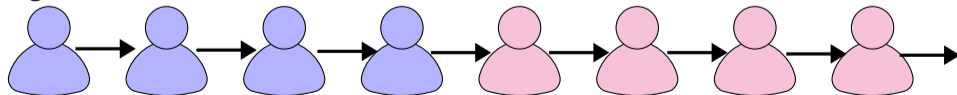
Agent's Room: Plan



Agent's Room: Write



Agents' Room: Plan + Write



Example Story

Writing prompt

Write a short story about a myth or legend being fulfilled in an **unusual or unexpected way**, or even cleverly bypassed. [...] It should take place during the Covid-19 pandemic in Europe. The narrative should involve vampires. The genre should be comedic horror.

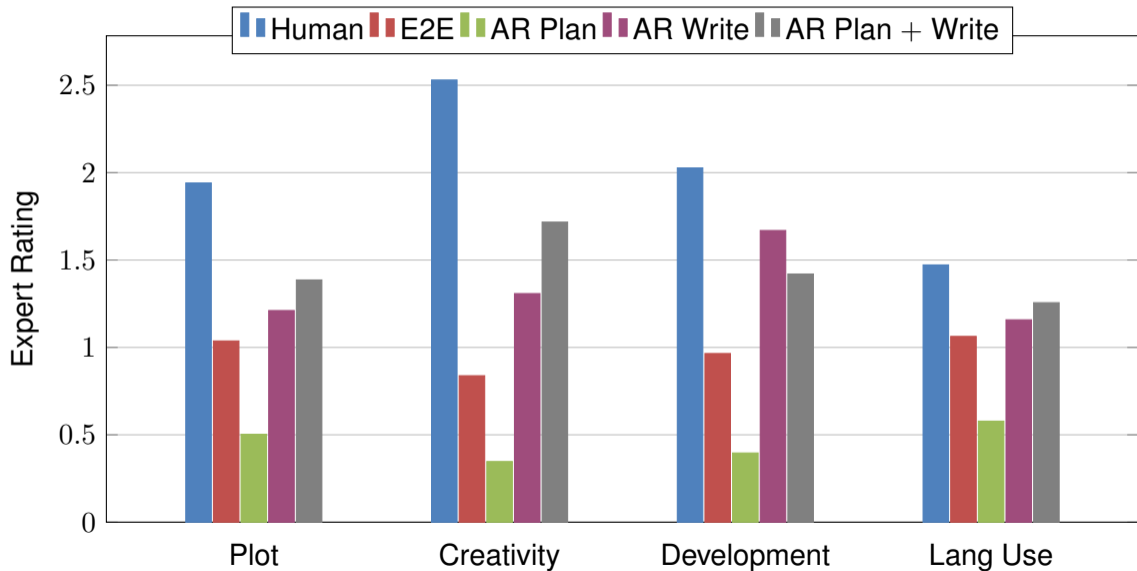
Conflict Agent

[...] Dracula's Creative Solution:
To overcome these obstacles, Dracula cleverly disguises himself as a delivery person for Vlad's Kebab House, a local restaurant. This allows him to move around town undetected and **deliver blood-sucking pizzas to unsuspecting customers**. However, his plan goes awry when he encounters a delivery driver from Pizza Hut, who challenges his territory and exposes his true nature. [...]

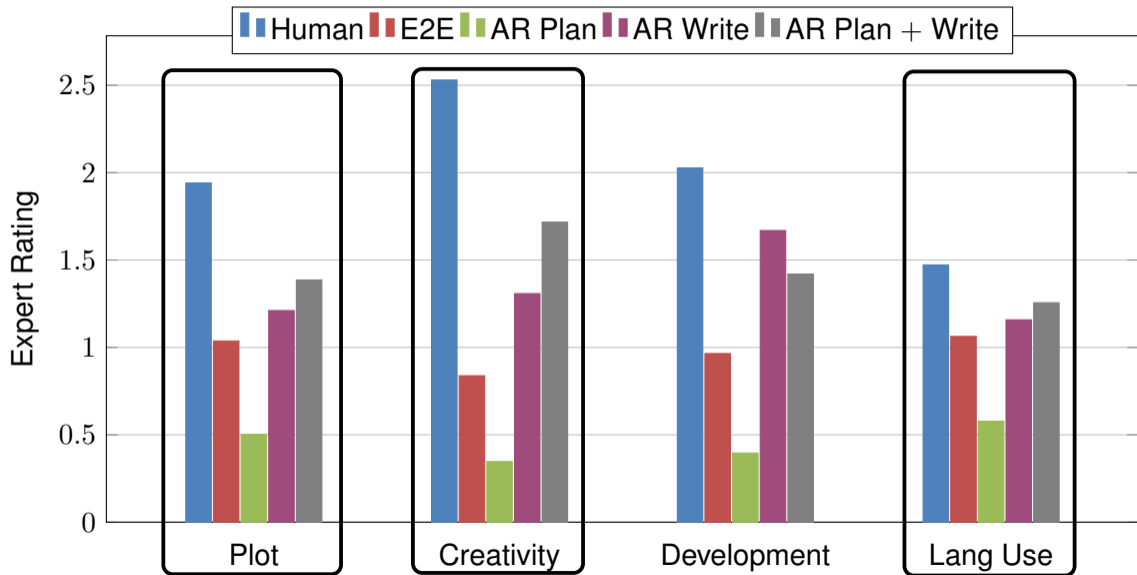
Story Excerpt

[...]
"It is the perfect plan," Dracula said, a sinister smile spreading across his face. "**The people will be so eager to receive their food that they will not suspect a thing.**"
[...]

Results



Results



- Long-form writing is hard!
- Synthesizing thousands of tokens in coherent narratives is somewhat feasible.
- LLMs still struggle with pacing story development, originality, character.
- Agents collaborate through scratchpad, but there is no reasoning involved.

Part II

Next Chapter Prediction

Novel Task: Next Chapter Prediction

Aunt Petunia obviously scented danger, too, because she said quickly, "And we'll buy you another two presents while we're out today. How's that, popkin? Two more presents. Is that all right?" Dudley thought for a moment. It looked like hard work. Finally he said slowly, "So I'll have thirty . . . thirty . . ." "Thirty-nine, sweetums," said Aunt Petunia.

"Oh." Dudley sat down heavily and grabbed the nearest parcel.

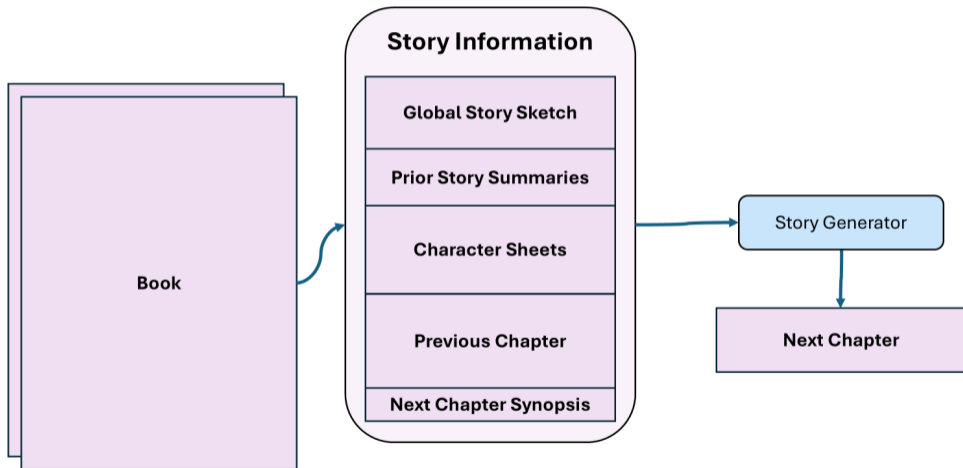
"All right then."

Uncle Vernon chuckled.

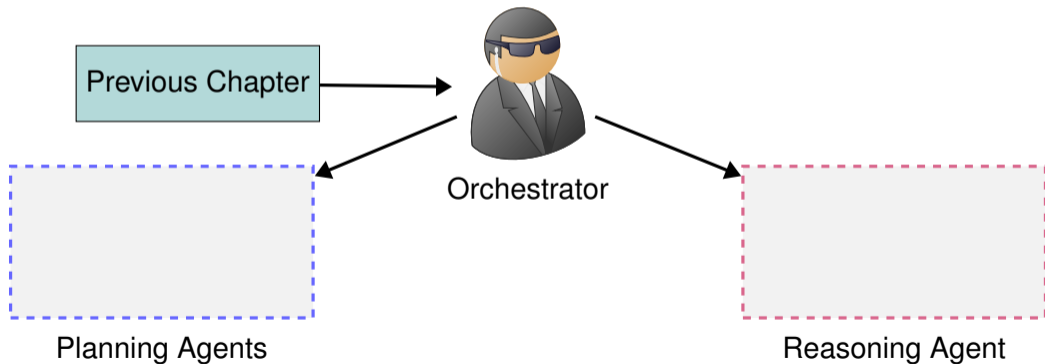
Chapter Two



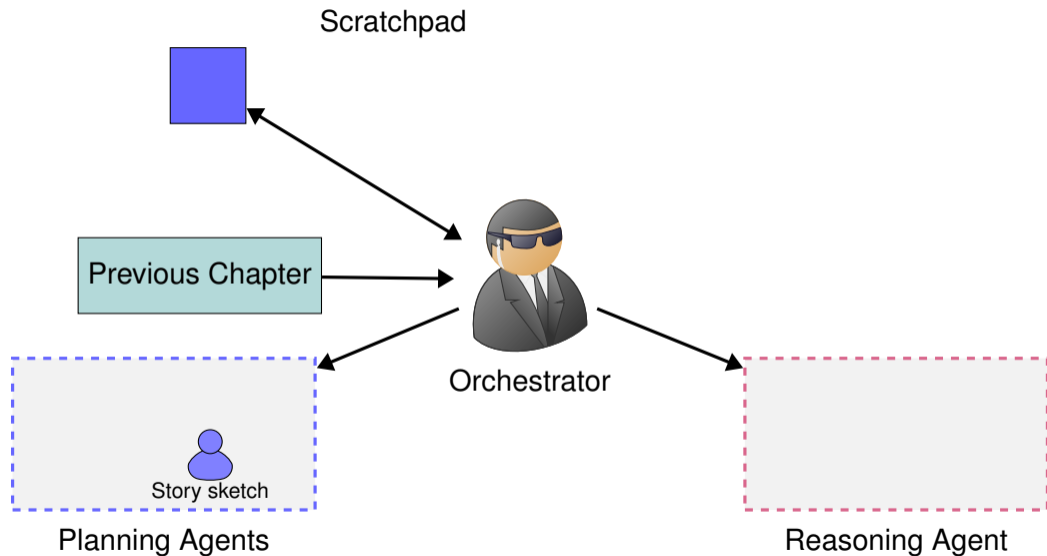
Novel Task: Next Chapter Prediction



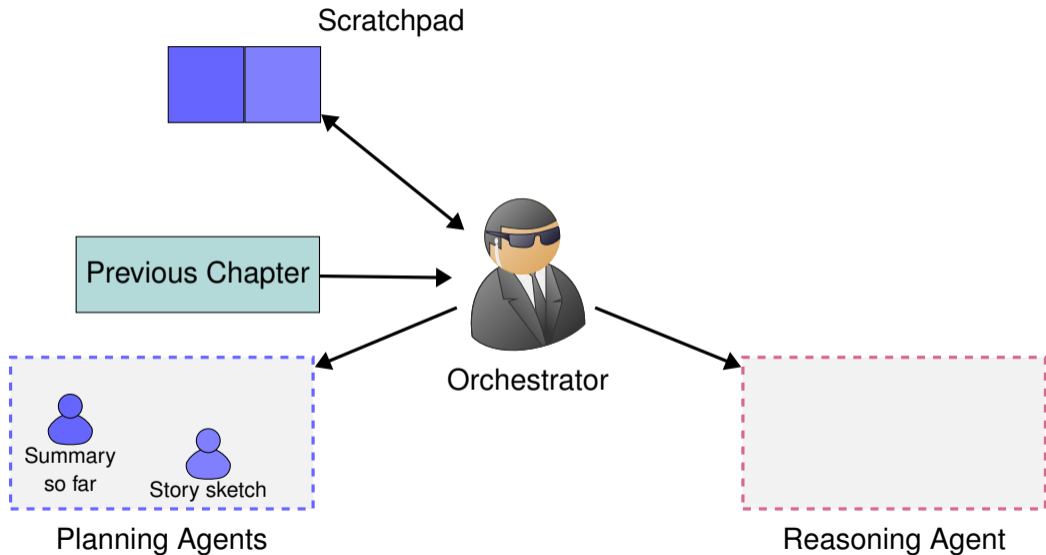
Scratchpad



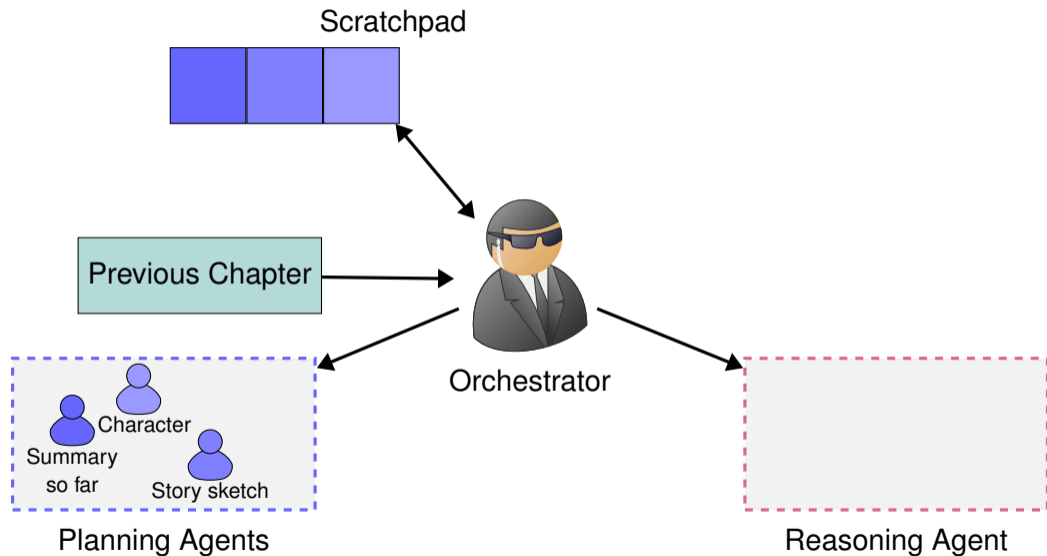
Learning to Reason



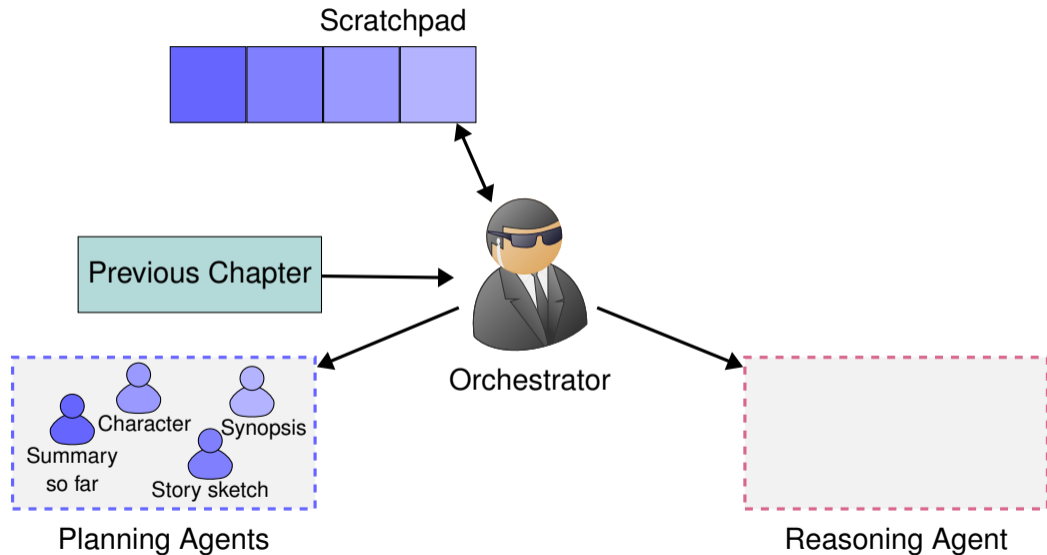
Learning to Reason



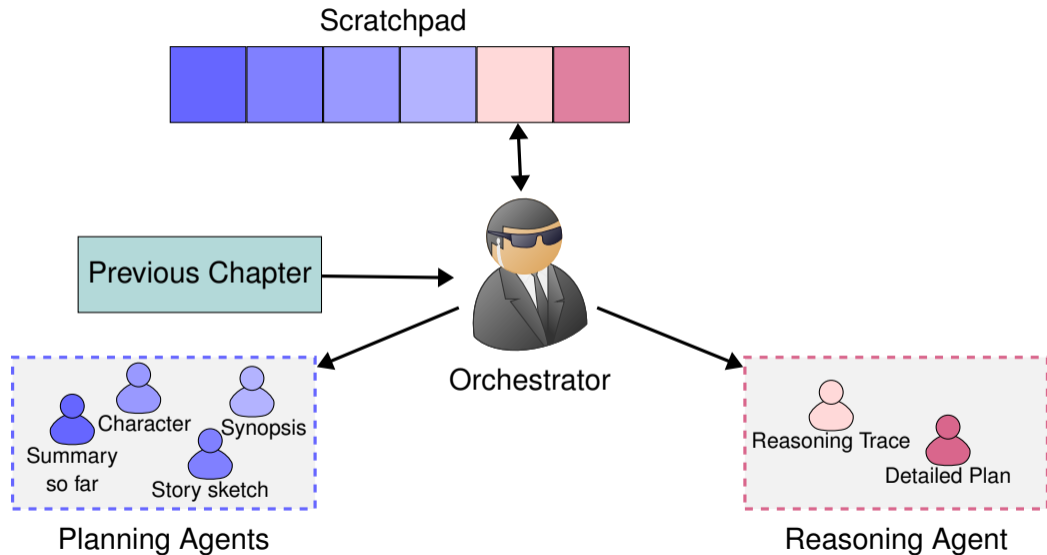
Learning to Reason



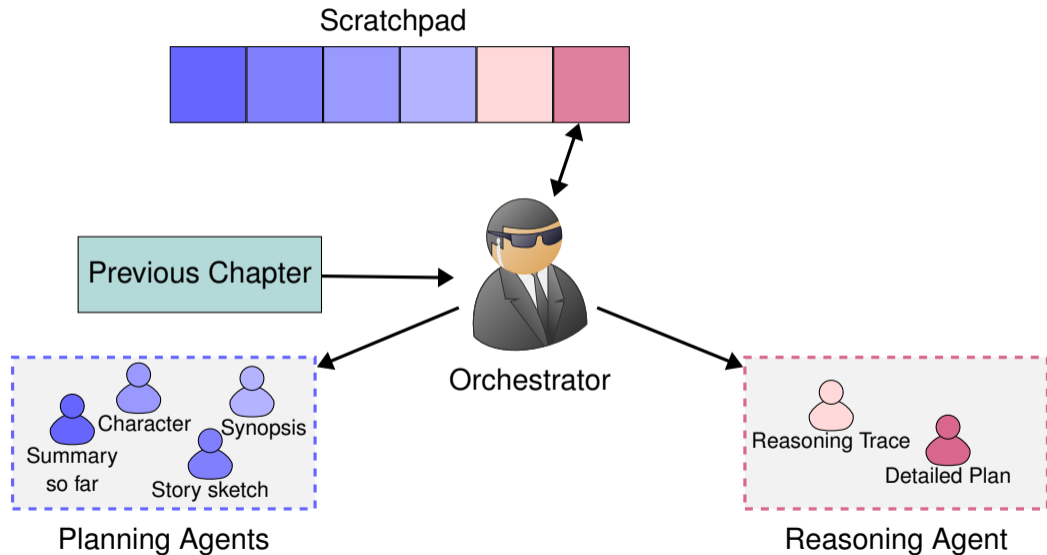
Learning to Reason



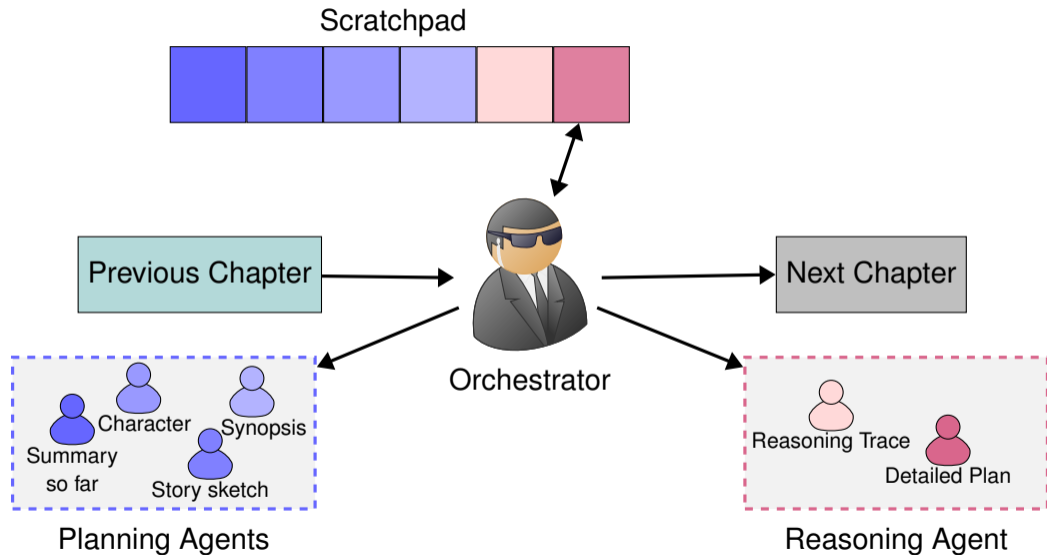
Learning to Reason



Learning to Reason



Learning to Reason



Agent Configuration

Multiple agents work together on a **generation** task. They generate new content, given information about the story and previous chapters.

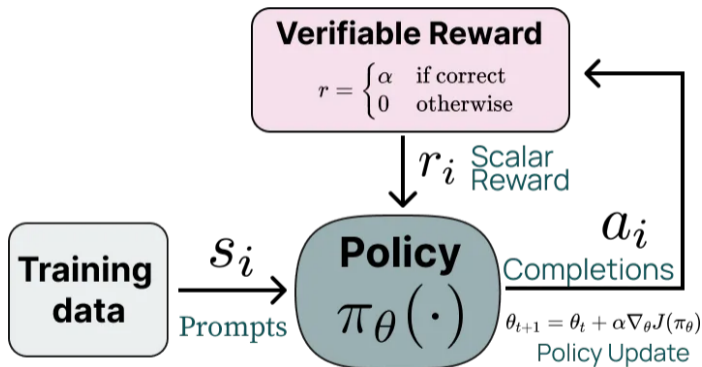
- **Communication:** collaborative, via language
- **Agents:** planning, reasoning, writing
- **Shared Persistent State:** scratchpad
- **Learn to reason** based on scratchpad

Experiments on 30 fiction books published in-or-after 2024 from SuperSummary
Qwen-7B models, avg. next chapter is 2.5K tokens.

Gurung and Lapata (2024, EMNLP), Gurung and Lapata (2025, COLM)

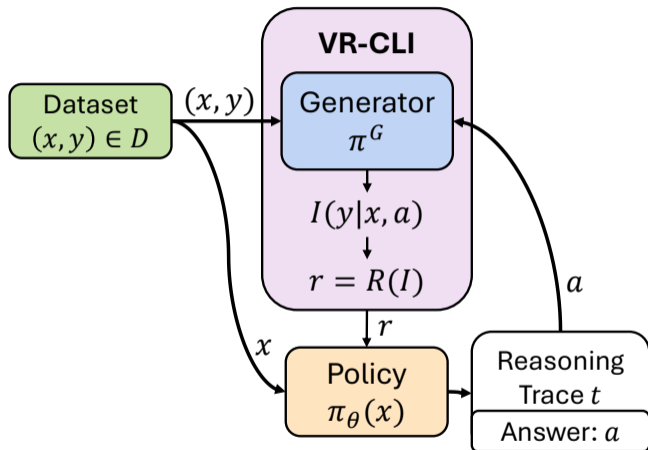
How do we train the Reasoning Agent?

Lambert et al. (2024): Reinforcement Learning from Verifiable Rewards (RLVR).



Model is trained using reinforcement learning algorithm like PPO or GRPO.

Verifiable Rewards via Completion Likelihood Improvement



- Need high-quality dataset: input x , output completions y
- Need generator that should output y
- **Key assumption:** improving the generator's likelihood of producing the provided dataset will improve the quality of its generations.

Improvement: how much does this answer improve the likelihood of y ?

$$I_{\pi_G}(x, y, a) = \left[1 - \frac{PPL_{\pi_G}(y|x, a)}{PPL_{\pi_G}(y|x)} \right] \times 100$$

Reward: some simple formulation based on improvement:

$$R(x, y, a) = \begin{cases} \alpha & I(x, y, a) \leq \omega_0 \\ \beta & \omega_0 < I(x, y, a) \leq \omega_1 \\ \gamma & \omega_1 < I(x, y, a) \end{cases} \quad R(x, y, a) = \max[0, I(x, y, a)]$$

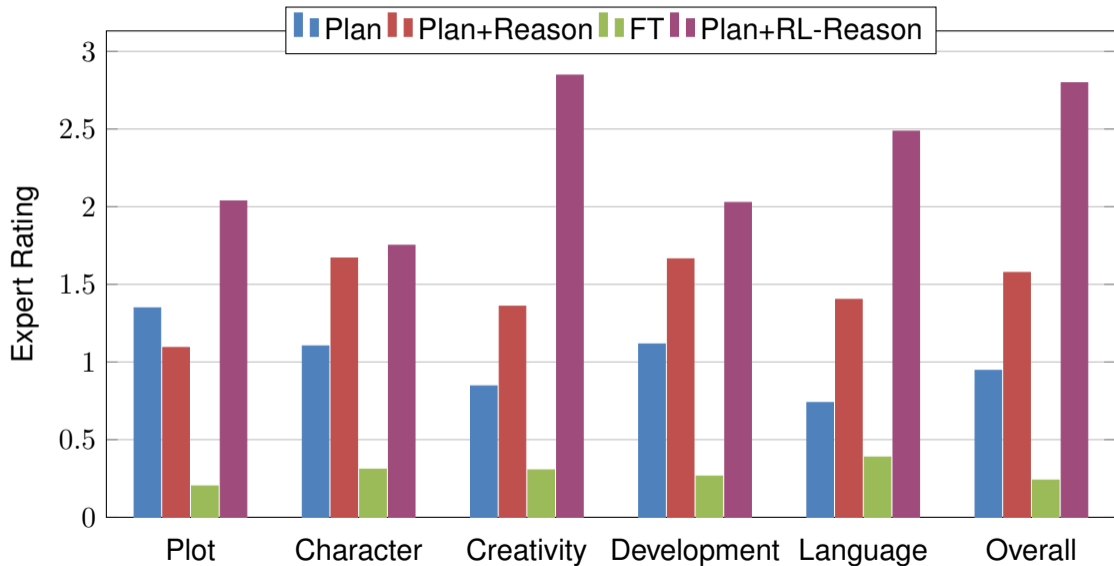
Example Reasoning Trace

Detailed Plan for Next Chapter###

1. **Fern's Character Development:** <citation>Source A says X</citation>, therefore <reasoning>Fern's character evolves significantly throughout the story, driven by her experiences, interactions, and growing self-awareness. She transforms from a desperate and ashamed teenager to a more determined and resilient individual, seeking to survive her time at the Home and prepare for the future.</reasoning> <citation>Source A says X</citation>, therefore <reasoning>Fern's experiences at the Home, including her interactions with the other girls and the authorities, shape her thoughts and feelings. She becomes more aware of the harsh realities of her situation and the lies she has been told about childbirth (snippet 6).</reasoning>
2. **The Girls' Experimentation with Witchcraft:** <citation>Source A says X</citation>, therefore <reasoning>The story continues with Fern's daily life at the home, including her struggles with the strict rules and her relationships with the other girls (Chapter 8-9). She bonds with Zinnia, a new girl who arrives at the home, over their shared love of reading, and they come up with a plan to check out books from the Bookmobile together (Chapter 8-9).</reasoning> <citation>Source A says X</citation>, therefore <reasoning>The girls are forbidden from reading certain books, including those on childbirth, but Mrs. Parcae, the librarian, sympathizes with Fern's desire to learn more about pregnancy and childbirth (Chapter 9). In a significant moment, Mrs. Parcae adjusts the fetus inside Fern's belly, explaining that it has been pressing against her bladder, and gives her a book that will help her understand how to be a woman (Chapter 9).</reasoning>
3. **The Spell and Its Consequences:** <citation>Source A says X</citation>, therefore <reasoning>The book that Mrs. Parcae gave Fern is titled How to Be a Groovy Witch. It contains spells for various things, all of which seem useful to Fern. Because Mrs. Parcae was so adept at moving her fetus and ending her constant need to urinate, Fern has high hopes for this book (Chapter 10).</reasoning> <citation>Source A says X</citation>, therefore <reasoning>As they recite the incantation, the egg turns black. They hope that this means that the spell is working. When they transfer the egg and the other items into a witch bag and hide it in Dr. Vincent's office, he begins vomiting uncontrollably (Chapter 10).</reasoning>

In summary: In the next chapter, Fern's character development is portrayed through her experiences at the Home for unwed mothers. The chapter ends with the girls' experimentation with witchcraft, and the spell they cast to harm Dr. Vincent. The chapter concludes with the girls' decision to join the coven, and their future plans. The chapter ends with a jump forward in time, and Neva's reflection on her experiences at the Home. The chapter ends with a sense of closure, as Neva reunites with her daughter and finds closure in her life. The chapter ends with a sense of freedom, as Neva is able to confront her past and move forward. The chapter ends with a sense of satisfaction, as Neva's life is complete. The chapter ends with a sense of closure, as the story comes to an end. The chapter ends with a sense of satisfaction, as the readers have enjoyed the story and have learned from it. The chapter ends with a sense of closure, as the story has come to an end. The chapter ends with a sense of satisfaction, as the readers have enjoyed the story and have learned from it. The chapter ends with a sense of closure, as the story has come to an end.

Results



What does the future hold?

- **Evaluating multi-LLM collaboration:** no agreed-upon and detailed evaluation methodology.
- **Design of multi-LLM systems:** given a repository of agents, we should learn their configuration for the task at hand, knowing when to call an agent.
- **What are good multi-LLM problems:** designing tasks and datasets where multiple models divide and conquer complex problems.
- **Many levels of LLM collaboration:** we have only looked at inference (what about pre-training and post-training).

References

- Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi, and Yulia Tsvetkov. 2025. When one LLM drools, multi-LLM collaboration rules.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges.
- Alexander Gurung and Mirella Lapata. 2024. CHIRON: Rich character representations in long-form narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8523–8547, Miami, Florida, USA. Association for Computational Linguistics.
- Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. LLM multi-agent systems: Challenges and open problems.
- Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. Agents' room: Narrative generation through multi-step collaboration.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. Publisher: arXiv Version Number: 4.
- Louis Mahon and Mirella Lapata. 2025. A modular approach for multimodal summarization of tv shows.
- Pinelopi Papalampidi and Mirella Lapata. 2023. Hierarchical3D adapters for long video-to-text summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1297–1320, Dubrovnik, Croatia. Association for Computational Linguistics.
- Rohit Saxena and Frank Keller. 2024. MovieSum: An abstractive summarization dataset for movie screenplays. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4043–4050, Bangkok, Thailand. Association for Computational Linguistics.