Dealing with opinionated requests and information inputs.

Aleksei Kolesnikov
Staff Software Engineer

Uncaptured Moments with the RAG Family

Issue:

In my earlier post on the RAG family, accessible here I purposefully left out a potential problem with context.

What happens if the question or context is biased?
What if the information sources reiterate misleading data?

Under such circumstances, the performance of the RAG family might falter.

So, what's our workaround?

The answer is simple:

1. Filter the information!
2. Rate the response!

System 2 Attention (S2A) by Meta

Idea:

Use specially tuned language models to rewrite the context. S2A does this mainly by eliminating irrelevant text.

As a result, it allows the language models to make precise decisions about what parts of the input to home in on before churning out a response.

Why:

The built-in attention mechanism isn't perfect and can sometimes pull in needless info into the context. This gets particularly tricky when a specific entity pops up several times in the context.

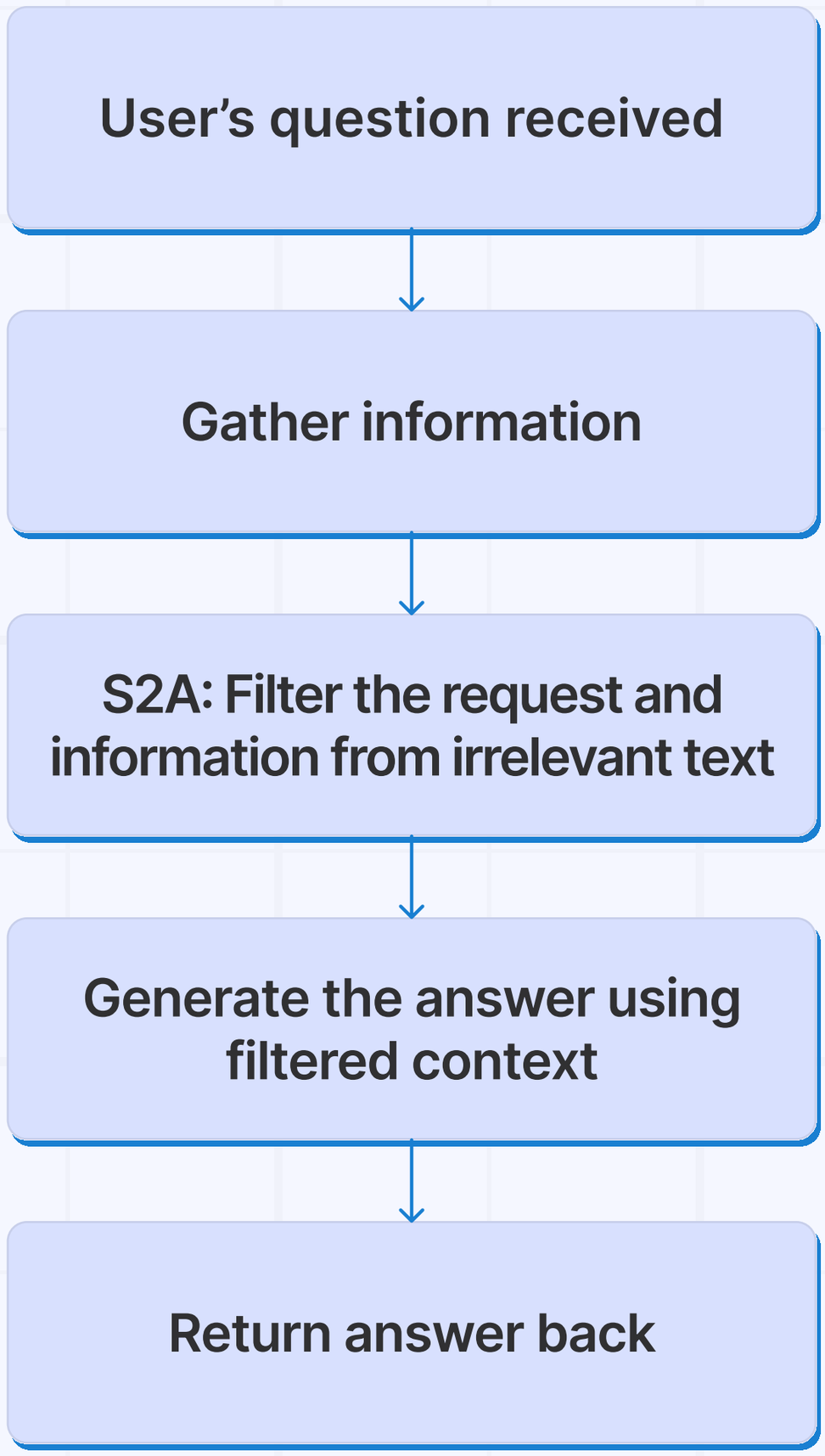
Want to dive deeper into this?

Feel free to check out these useful links for a more detailed scoop:

1. Academic Paper
2. GitHub Repository

Aleksei Kolesnikov
Staff Software Engineer

System 2 Attention (S2A) by Meta



System 2 Attention (S2A) by Meta

Advantages:

→ Improved performance

→ Better answer correlation

→ Huge improvements in case of biased or opinionated questions and requests

Additional details

Adding additional clean up steps will not be beneficial according to research

*for more information you can visit:
https://arxiv.org/pdf/2311.11829.pdf

Reinforcement Learning from Human Feedback (RLHF)

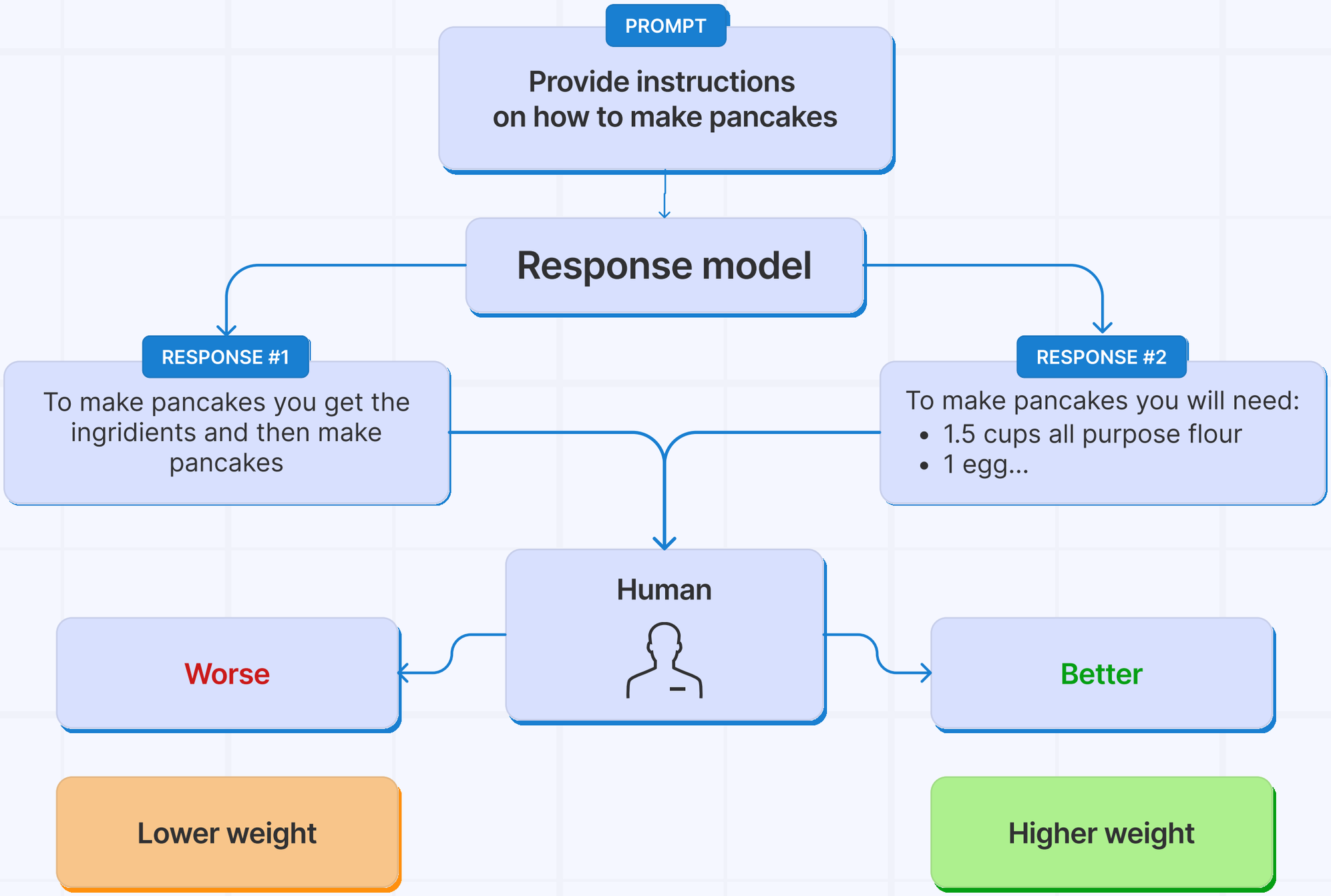
Idea:

RLHF, used in AI assistants like ChatGPT, counts on human feedback for training via a Preference Model (PM).

This allows us to enhance answers and evaluate used sources based on user feedback.

For a deeper dive, you may visit the following repo:
https://github.com/Glareone/awesome-RLHF-GenAI

Reinforcement Learning from Human Feedback (RLHF)




Aleksei Kolesnikov
Staff Software Engineer

Reinforcement Learning from Human Feedback (RLHF)

Advantages:

→ Ethical alignment

→ Context recognition: Human evaluators can offer feedback specific to context

Limitations:

→ Human bias

→ Scalability issues: The requirement for human input can limit RLHF's scalability

Aleksei Kolesnikov
Staff Software Engineer

RLAIF Reinforcement Learning from AI Feedback

Concept:

RLAIF's utilizes another AI Feedback Model's feedback rather than direct human input.

Anthropic's RLAIF method has one AI model rectifying another based on a principles set, termed as "Constitutional AI". The application of these principles to AI judgments enables models to learn how to make better choices.

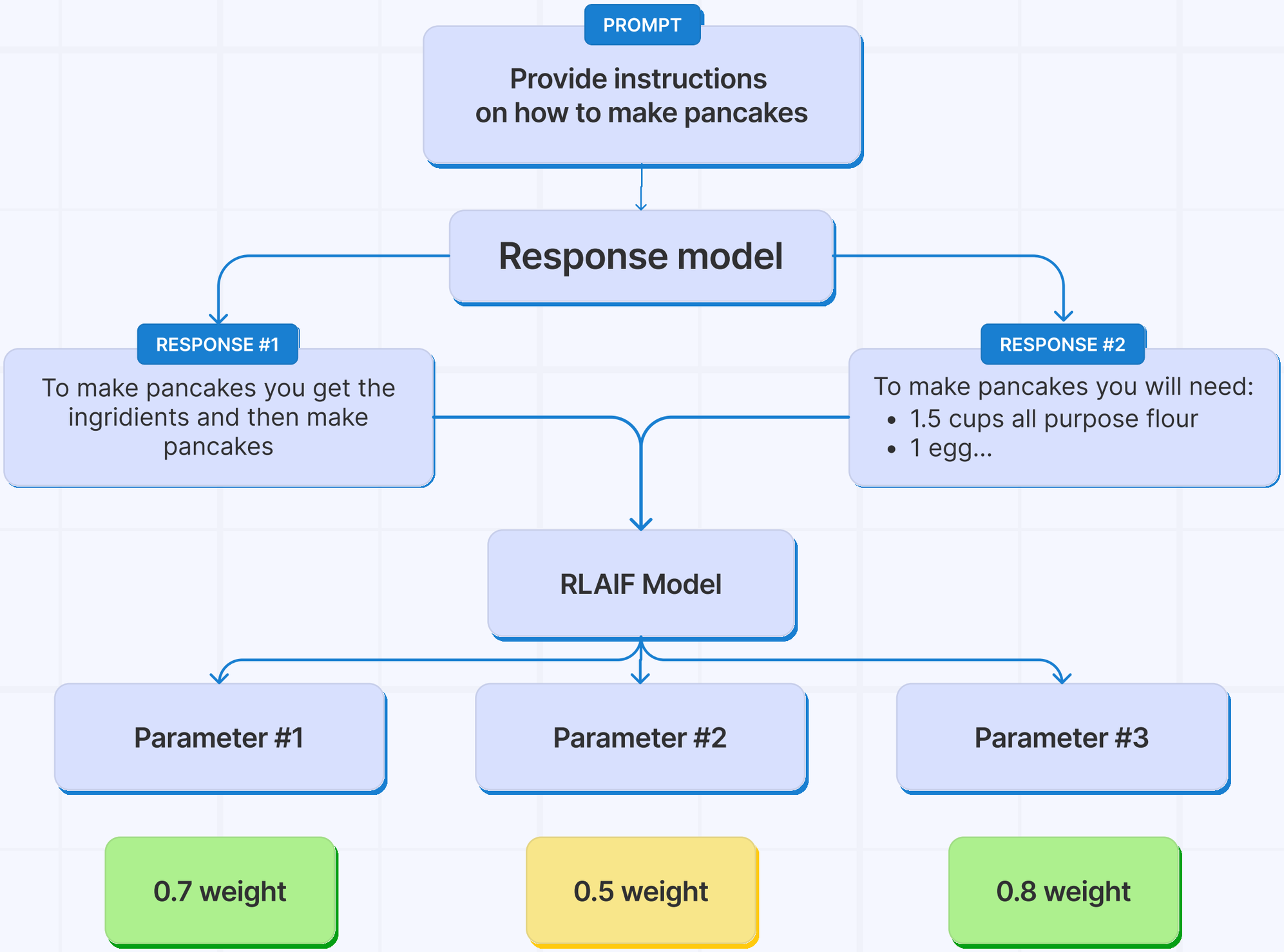
Where could be useful:

→ Traffic Management: AI systems can enhance traffic movement and lessen jams.

→ Environmental Monitoring: AI-backed environmental tracking systems can process colossal data amounts to spot trends and give early hazard warnings.

→ Ethnic and Compliance

RLAIF Reinforcement Learning from AI Feedback




Aleksei Kolesnikov
Staff Software Engineer

RLAIF Reinforcement Learning from AI Feedback

Advantages:

→ Scalability: RLAIF surpasses RLHF in effective scalability

→ Efficiency: RLAIF fosters quicker learning and adaptation in AI systems

Limitations:

→ Prerequisite for robust principles: RLAIF necessitates a broad set of rules

→ Partiality: AI-generated feedback may lack ethical examinations and nuances compared to human inputs.

For a deeper dive, you may visit the following repo:
https://github.com/Glareone/awesome-RLHF-GenAI

Aleksei Kolesnikov
Staff Software Engineer