[CHAT-5.0] REASONING MODELS

Chain-of-Thought (CoT) prompting

- Give the model time to think, instructing it to work out its own solution before rushing to a conclusion.
- CoT prompting helps when:
  - The task is challenging and requires multi-step reasoning.
  - A large (say +100B parameters) model is used.

Prompt chaining

- Use a sequence of prompts to hide the model’s reasoning process.
- Summarize or filter previous dialogue.
- Ask the model if it missed anything on previous passes.
- Not all tasks require in-depth thinking. Use prompt chaining judiciously to ensure the right balance of performance and latency.
- Summarize documents piecewise and construct a full summary recursively.

In favor of chaining prompts

- Easier to test.
- Same reasons as for splitting a computer program in code chunks.
- Less tokens per prompt. This may matter, because of cost and context window issues.
- Allows skipping part of the workflow when not needed for the task.
- For complex tasks, allows to keep track of what happens out of the model.
- Allows the use of external tools.

Reasoning models

- The model builds the chain of thought by itself.
- This part is presented separately in some chat apps. For instance, DeepSeek encloses it between the tags `<think>` and `</think>`.
- In chat apps, most of the prompts do not require reasoning, and reasoning delays the response, so it is presented as an option. Under the hood, the choice involves choosing between two different models (possibly with ther same architecture but different parameter values).