How to get the model to do what you want

Rudolf Rosa

RAG, CoT

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

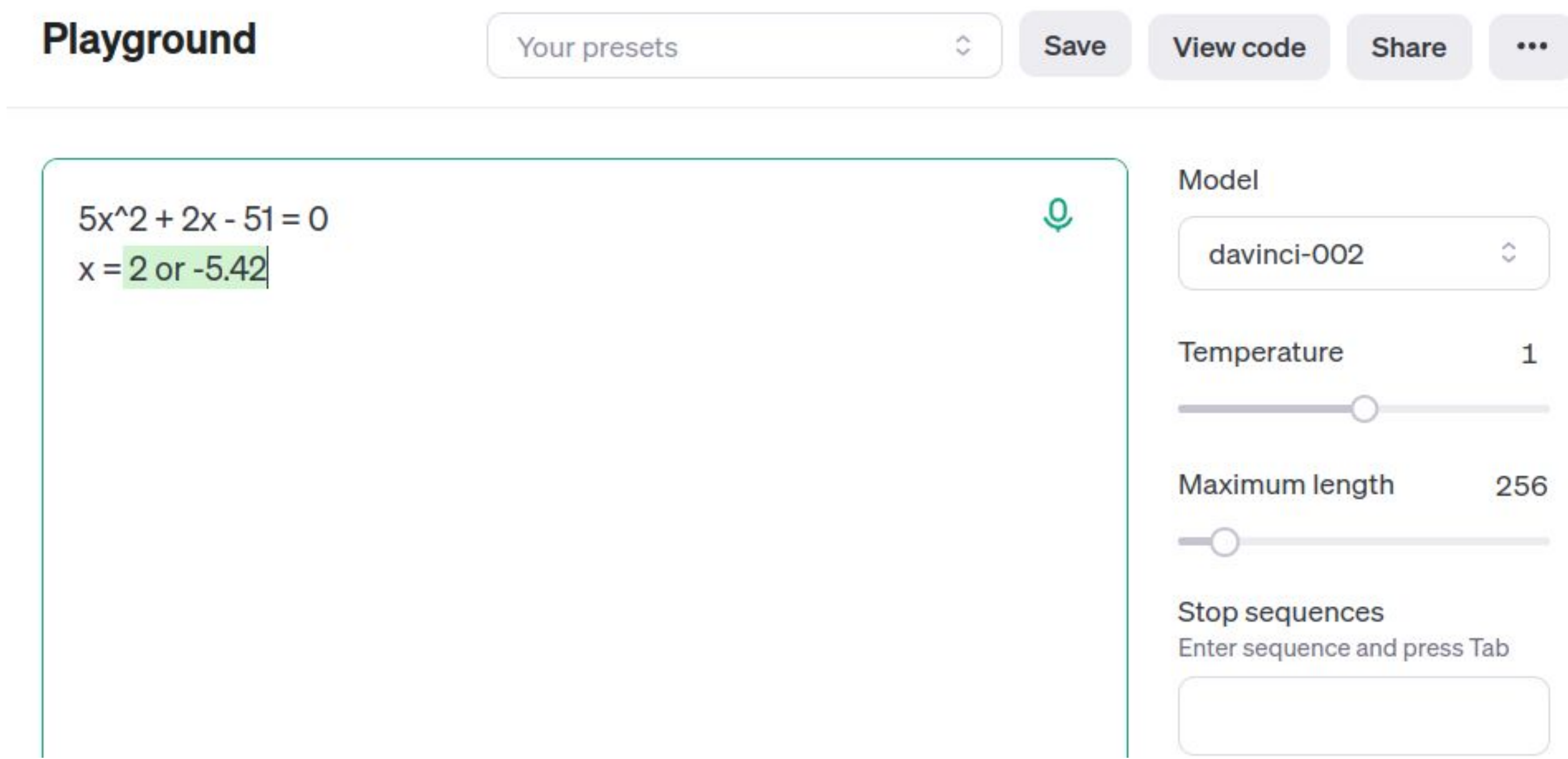
BY SA
unless otherwise stated

Chain of Thought

Chain of Thought

- classical via prompting: “think step by step”…
  - difficult:
    - 5x^2 + 2x - 51 = 0
    - x = …
  - easier:
    - 5x^2 + 2x - 51 = 0
    - D = 4 + 4 * 5 * 51
    - D = 1024
    - sqrt(D) = 32
    - x = (-2 +- 32) / (2 * 5)
    - x = {-3.4; 3}
- generalized
  - add annotations of the step-by-step thinking & fine-tune on that
  - model learns to do the step-by-step thinking itself, without prompting it to do so

Chain of Thought



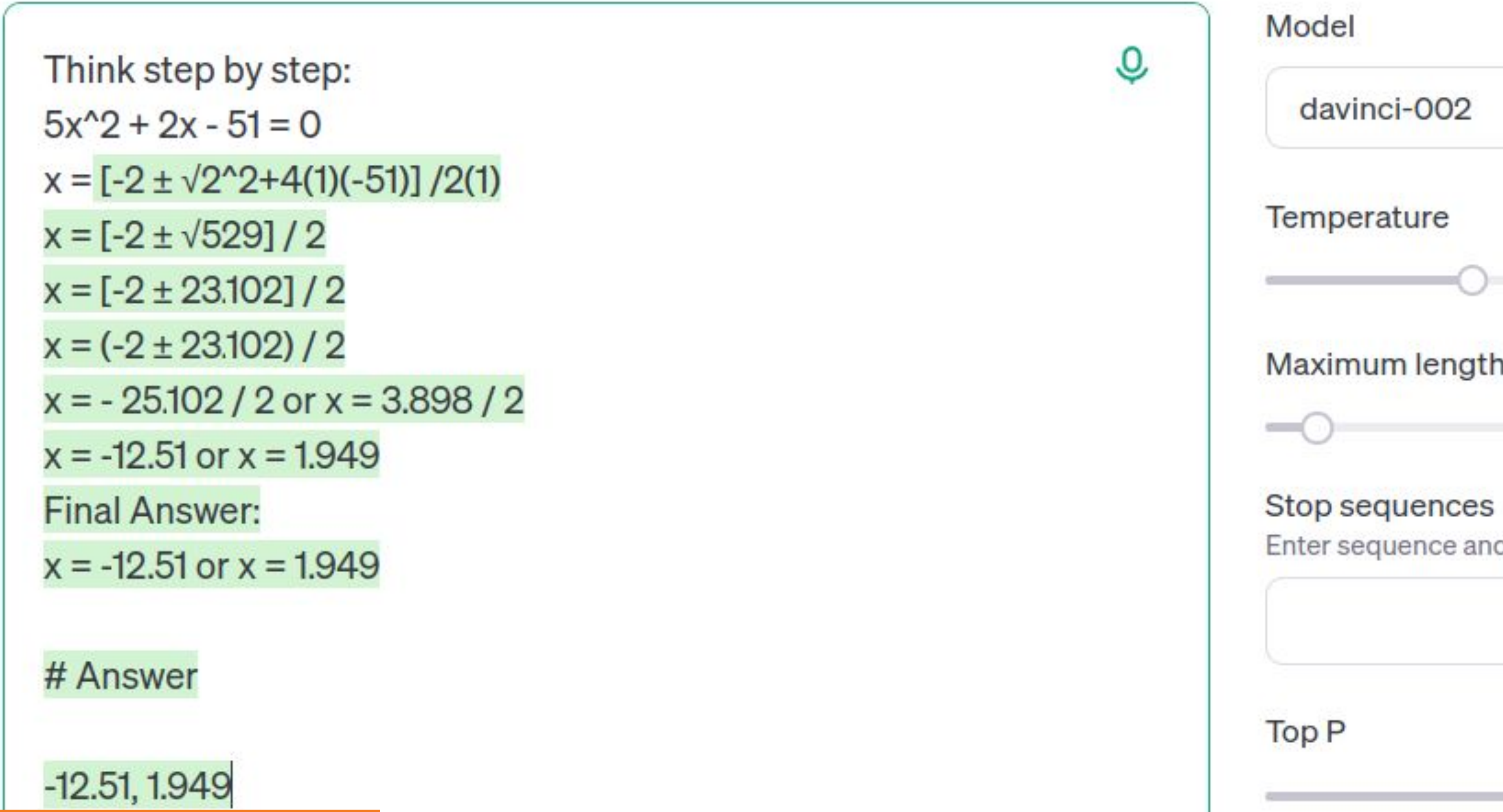
5x^2 + 2x - 51 = 0
x = 2 or -5.42

Model: davinci-002

Temperature: 1

Maximum length: 256

Stop sequences



```
Think step by step:
5x^2 + 2x - 51 = 0
x = [-2 ± √2^2+4(1)(-51)] /2(1)
x = [-2 ± √529] / 2
x = [-2 ± 23.102] / 2
x = (-2 ± 23.102) / 2
x = - 25.102 / 2 or x = 3.898 / 2
x = -12.51 or x = 1.949
Final Answer:
x = -12.51 or x = 1.949

# Answer

-12.51, 1.949
```

Model: davinci-002

CoT example (EduPo project): generating poems

jak zřel amor, malé dítě,
tiše dřímat těžký sen,
na prahu již čekal na tě
ne, on z jizby vyšel ven.

# ABAB
T # 8 # ítě # jak zřel amor, malé dítě,
T # 7 # sen # tiše dřímat těžký sen,
T # 8 # tě # na prahu již čekal na tě
T # 7 # ven # ne, on z jizby vyšel ven.

Chain of Thought: Why does it work?

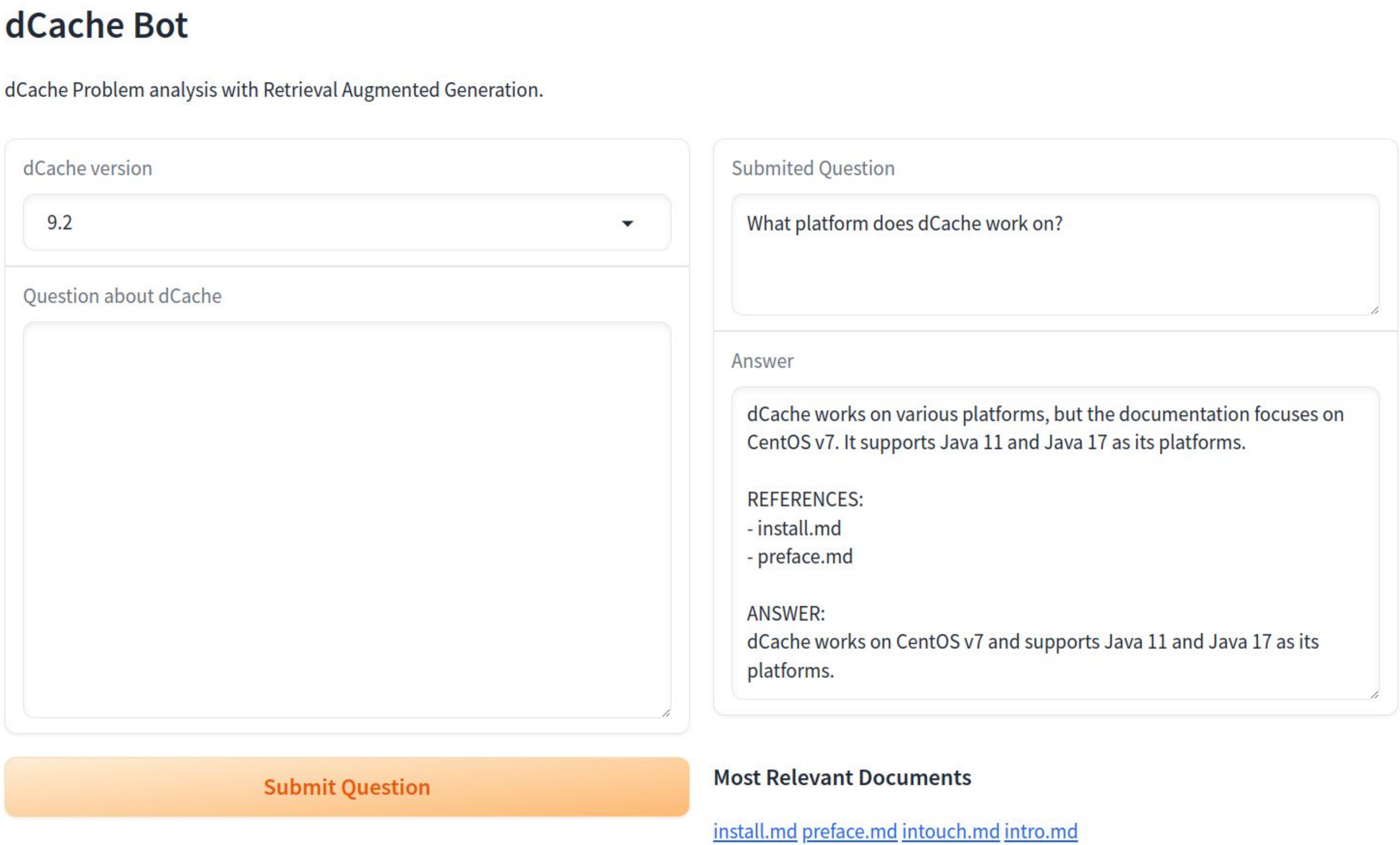
- learning small steps from data
- local reasoning
- more capacity to compute the answer
- …

Retrieval Augmented Generation

Retrieval Augmented Generation

- documents → embeddings
  - various options for document embeddings, e.g. using BERT
  - long documents → split into parts
- query (→ reformulation) → embedding
  - reformulation: e.g. “Generate three possible documents containing an answer to the question.”
- cosine similarity → candidate documents → context

Retrieval Augmented Generation



dCache Bot

dCache Problem analysis with Retrieval Augmented Generation.

dCache version: 9.2

Question about dCache:

Submited Question: What platform does dCache work on?

Answer:

dCache works on various platforms, but the documentation focuses on CentOS v7. It supports Java 11 and Java 17 as its platforms.

REFERENCES:
- install.md
- preface.md

ANSWER:
dCache works on CentOS v7 and supports Java 11 and Java 17 as its platforms.

Most Relevant Documents

install.md preface.md intouch.md intro.md

What are some positives and negatives?

- adds relevant context
    - generation easier
    - generation more trustworthy
- implicit or even explicit reference to sources
- can use on own documents
- …

- can easily leak data from documents
- retrieval is simpler than LLM
    - → not as strong
    - cosine similarity of documents might not be best criterion
- need good document embeddings
- …

How to get the model to do what you want

- CoT = chain of thought
- RAG = retrieval augmented generation