The Effect of Sampling Temperature on Problem Solving in Large Language Models

Matthew Renze and Erhan Guven

Johns Hopkins University

What is the optimal sampling temperature for an LLM on problem-solving tasks?

Why is 0.7, in general, the default value of temperature for LLMs

https://datascience.stackexchange.com/questions/124584/why-is-0-7-in-general-the-default-value-of-temperature-for-llms

Low temperature (e.g. 0.2):
Ideal for tasks requiring precise
and predictable results,
such as technical writing
or formal documentation.

https://www.bighummingbird.com/blogs/llm-hyperparameter

How to choose the perfect LLM Temperature?

https://www.hopsworks.ai/dictionary/llm-temperature

[What is the] best Temperature for Gpt-4 api to get quality coding advice and samples?

https://www.reddit.com/r/ChatGPTCoding/comments/12i6k06/best_temperature_for_gpt4_api_to_get_quality/

Choosing the right temperature for your LLM

https://www.promptfoo.dev/docs/guides/evaluate-llm-temperature/

The magic lies in finding the right temperature for the job.

https://generative-ai.io/blog/understanding-temperature-for-llms-guide/

Hot or Cold? Adaptive Temperature Sampling for Code Generation with Large Language Models

https://ojs.aaai.org/index.php/AAAI/article/view/27798

What is the optimal temperature setting for Large Language Models (LLMs) to achieve maximum accuracy in information extraction?

https://typeset.io/questions/what-is-the-optimal-temperature-setting-for-large-language-4pv05kkwve

What is the optimal sampling temperature for an LLM on problem-solving tasks?

The Effect of Sampling Temperature on Problem Solving in Large Language Models

Matthew Renze
Johns Hopkins University
mrenze1@jhu.edu

Erhan Guven
Johns Hopkins University
eguven2@jhu.edu

Abstract

In this research study, we empirically investigate the effect of sampling temperature on the performance of Large Language Models (LLMs) on various problem-solving tasks. We created a multiple-choice question-and-answer (MCQA) exam by randomly sampling problems from standard LLM benchmarks. Then, we used nine popular LLMs with five prompt-engineering techniques to solve the MCQA problems while increasing the sampling temperature from 0.0 to 1.6. Despite anecdotal reports to the contrary, our empirical results indicate that changes in temperature from 0.0 to 1.0 do not have a statistically significant impact on LLM performance for problem-solving tasks. In addition, these results appear to generalize across LLMs, prompt-engineering techniques, and problem domains. All code, data, and supplemental materials are available on GitHub at: https://github.com/matthewrenze/jhu-llm-temperature.

Background

Sampling Temperature

LLM inference hyperparameter

$$\Pr(v_k) = \frac{e^{l_k/\tau}}{\sum_i e^{l_i/\tau}}$$

Sampling Temperature

LLM inference hyperparameter

Controls randomness of output

$$\Pr(v_k) = \frac{e^{l_k/\tau}}{\sum_i e^{l_i/\tau}}$$

Sampling Temperature

LLM inference hyperparameter

Controls randomness of output

Modification of softmax function

$$\Pr(v_k) = \frac{e^{l_k/\tau}}{\sum_i e^{l_i/\tau}}$$

Sampling Temperature

LLM inference hyperparameter

Controls randomness of output

Modification of softmax function

Adjusts probability mass functions

$$\Pr(v_k) = \frac{e^{l_k/\tau}}{\sum_i e^{l_i/\tau}}$$

Sampling Temperature

Lower τ → more deterministic

“Force equals mass times ___”

Token	Logits	Softmax
acceleration	-0.05	96.2%
velocity	-4.35	1.3%
gravity	-5.41	0.8%
distance	-6.67	0.4%
change	-6.71	0.1%

Sampling Temperature

Lower τ → more deterministic

Higher τ → more random

“Force equals mass times ___”

Token	Logits	Softmax
acceleration	-0.05	96.2%
velocity	-4.35	1.3%
gravity	-5.41	0.8%
distance	-6.67	0.4%
change	-6.71	0.1%

Sampling Temperature

Lower τ → more deterministic

Higher τ → more random

More “creative” but hallucinates

“Force equals mass times ___”

Token	Logits	Softmax
acceleration	-0.05	96.2%
velocity	-4.35	1.3%
gravity	-5.41	0.8%
distance	-6.67	0.4%
change	-6.71	0.1%

Sampling Temperature

Lower τ → more deterministic

Higher τ → more random

More “creative” but hallucinates

Exploring vs. exploiting solutions

“Force equals mass times ___”

Token	Logits	Softmax
acceleration	-0.05	96.2%
velocity	-4.35	1.3%
gravity	-5.41	0.8%
distance	-6.67	0.4%
change	-6.71	0.1%

Methods

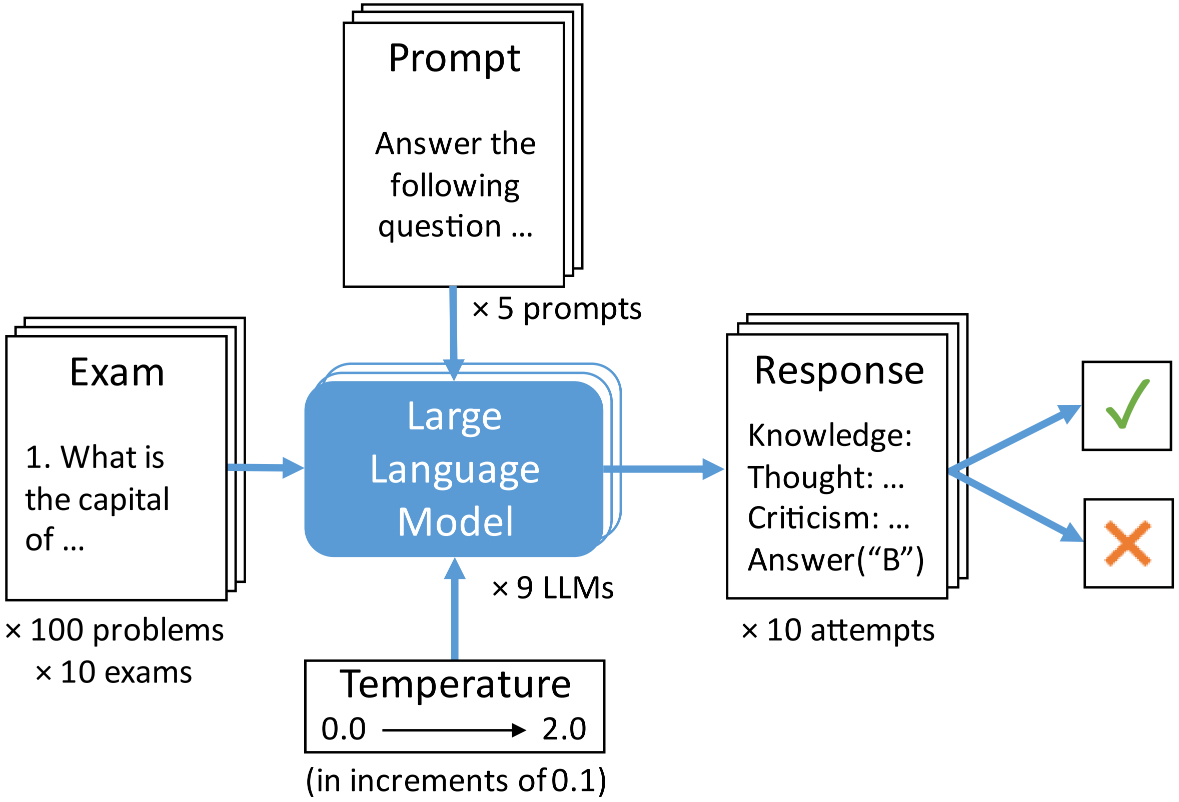
Experiment

9 models

5 prompts

10 exams

100 questions



Experiment

9 models

5 prompts

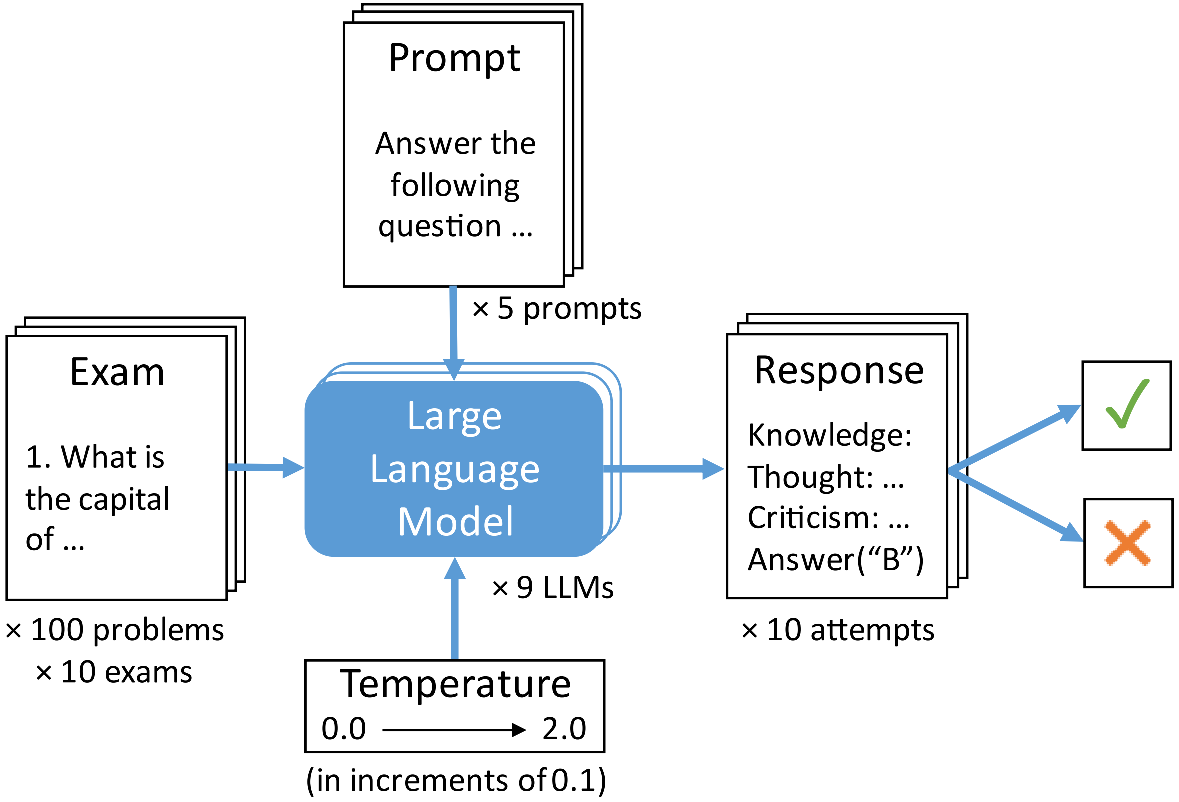
10 exams

100 questions

10 attempts

From 0.0 to 2.0

Increments of 0.1



Models

Name	Vendor	Released	License	Source
Claude 3 Opus	Anthropic	2024-03-04	Closed	Anthropic (2024)
Command R+	Cohere	2024-04-04	Open	Cohere (2024)
Gemini 1.0 Pro	Google	2023-12-06	Closed	Gemini Team (2023)
Gemini 1.5 Pro (Preview)	Google	2024-02-15	Closed	Gemini Team (2024)
GPT-3.5 Turbo	OpenAI	2022-11-30	Closed	OpenAI (2022)
GPT-4	OpenAI	2023-03-14	Closed	OpenAI (2023)
Llama 2 7B Chat	Meta	2023-07-18	Open	Meta (2023)
Llama 2 70B Chat	Meta	2023-07-18	Open	Meta (2023)
Mistral Large	Mistral AI	2024-02-26	Open	Mistral AI (2024)

Prompts

Baseline – no prompt engineering

Domain Expertise – "you are an expert in …"

Self-recitation – recite knowledge first

Chain of Thought – "think step-by-step"

Composite – all three + self-criticism

Exams

Problem Set	Benchmark	Domain	Questions	License	Source
ARC Challenge Test	ARC	Science	1,173	CC BY-SA	Clark (2018)
AQUA-RAT	AGI Eval	Math	254	Apache v2.0	Zhong (2023)
Hellaswag Val	Hellaswag	Common Sense Reasoning	10,042	MIT	Zellers (2019)
LogiQA (English)	AGI Eval	Logic	651	GitHub	Liu (2020)
LSAT-AR	AGI Eval	Law (Analytic Reasoning)	230	MIT	Wang (2021)
LSAT-LR	AGI Eval	Law (Logical Reasoning)	510	MIT	Wang (2021)
LSAT-RC	AGI Eval	Law (Reading Comprehension)	260	MIT	Wang (2021)
MedMCQA Valid	MedMCQA	Medicine	6,150	MIT	Pal (2022)
SAT-English	AGI Eval	English	206	MIT	Zhong (2023)
SAT-Math	AGI Eval	Math	220	MIT	Zhong (2023)

[System Prompt]

You are an expert in {{expertise}}.

Your task is to answer the following multiple-choice questions.

First, you should recite all of the relevant knowledge you have about the question and each option.

Next, you should think step-by-step through the problem to ensure you have the correct answer.

Then, you should critically evaluate your thoughts to identify any flaws in your facts, logic, and reasoning.

Finally, you MUST answer the question using the following format 'Action: Answer("[choice]")’

…

[Example Problem]

Question: What is the capital of the state where Johns Hopkins University is located?

Choices:

A: Baltimore
B: Annapolis
C: Des Moines
D: Las Vegas

[Example Solution]

Thought:

Johns Hopkins University is located in Baltimore.
Baltimore is a city located in the state of Maryland.
The capital of Maryland is Annapolis.
Therefore, the capital of the state where Johns Hopkins University is located is Annapolis.
The answer is B: Annapolis.

Action: Answer("B")

Topic: Science

Question: A certain atom has 20 electrons, 21 neutrons, and 20 protons. What is the atomic mass of the atom?

Choices:

A: 20

B: 21

C: 41

D: 61

Knowledge:

The atomic mass of an atom is the sum of the number of protons and neutrons in the nucleus of the atom.

The number of electrons in an atom is equal to the number of protons in the nucleus.

Knowledge:

The atomic mass of an atom is the sum of the number of protons and neutrons in the nucleus of the atom.

The number of electrons in an atom is equal to the number of protons in the nucleus.

Thought:

The number of protons in the nucleus is 20.

The number of neutrons in the nucleus is 21.

The atomic mass of the atom is the sum of the number of protons and neutrons, which is 20 + 21 = 41.

Therefore, the answer is C: 41.

Knowledge:

The atomic mass of an atom is the sum of the number of protons and neutrons in the nucleus of the atom.

The number of electrons in an atom is equal to the number of protons in the nucleus.

Thought:

The number of protons in the nucleus is 20.

The number of neutrons in the nucleus is 21.

The atomic mass of the atom is the sum of the number of protons and neutrons, which is 20 + 21 = 41.

Therefore, the answer is C: 41.

Criticism:

The thought process is correct and the answer is correct.

Knowledge:

The atomic mass of an atom is the sum of the number of protons and neutrons in the nucleus of the atom.

The number of electrons in an atom is equal to the number of protons in the nucleus.

Thought:

The number of protons in the nucleus is 20.

The number of neutrons in the nucleus is 21.

The atomic mass of the atom is the sum of the number of protons and neutrons, which is 20 + 21 = 41.

Therefore, the answer is C: 41.

Criticism:

The thought process is correct and the answer is correct.

Action: Answer("C")

Votes: C, B, C, C, C, C, C, A, C, C

Agent Answer: C

Correct Answer: C

Score: 0.8

Tokens: 1922

### Results ###

Start Time: 2024-06-07 19:40:36

End Time: 2024-06-07 21:21:14

Model Name: gpt-4

Agent Name: chain_of_thought

Exam Name: comprehensive-100

Temperature: 0.0

Questions: 100

Accuracy: 0.8100

Analysis

Kruskal-Wallis Test

Analysis

Kruskal-Wallis Test

Accuracy vs. temperature

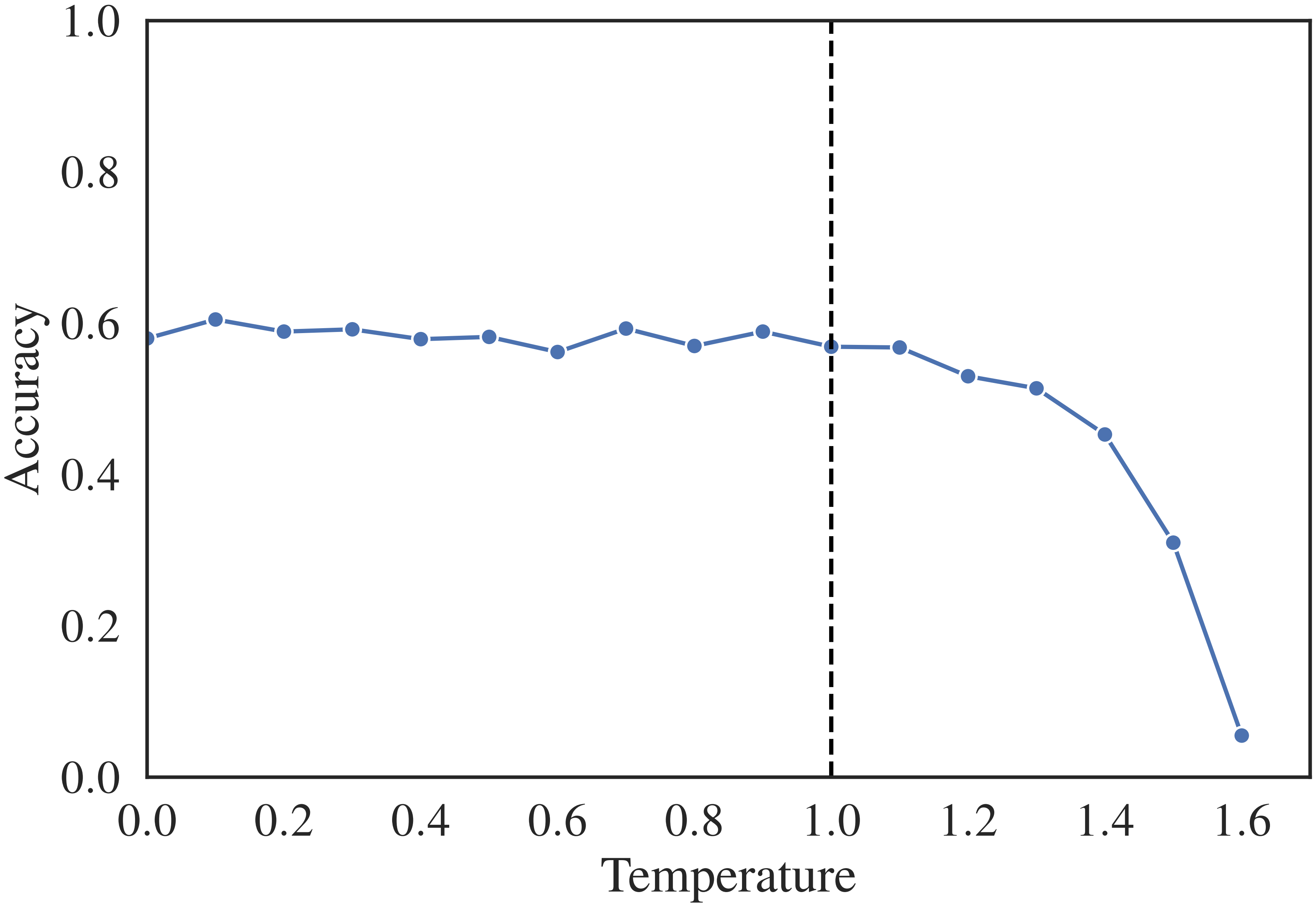
Analysis

Kruskal-Wallis Test

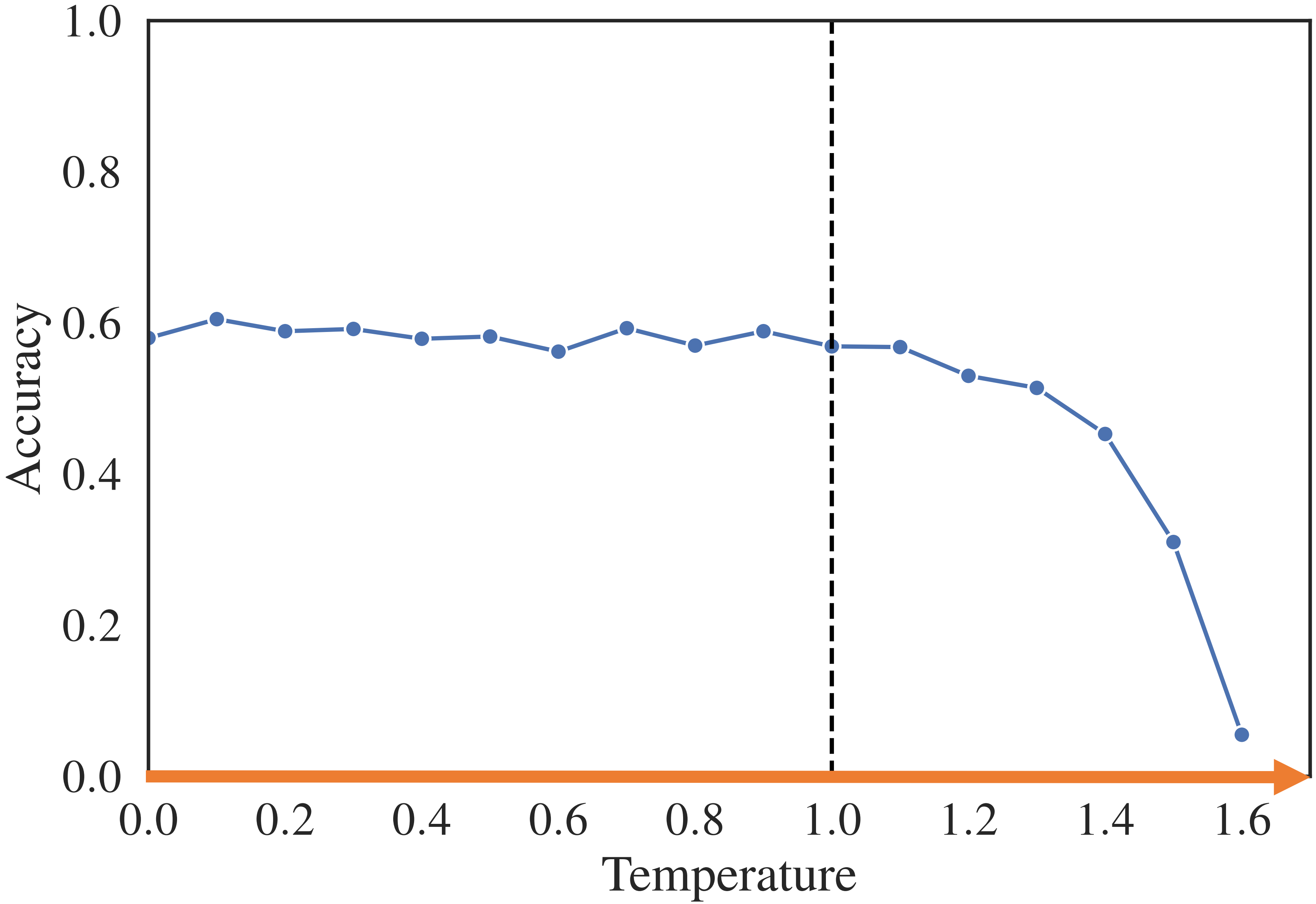
Accuracy vs. temperature

Data were bimodally distributed

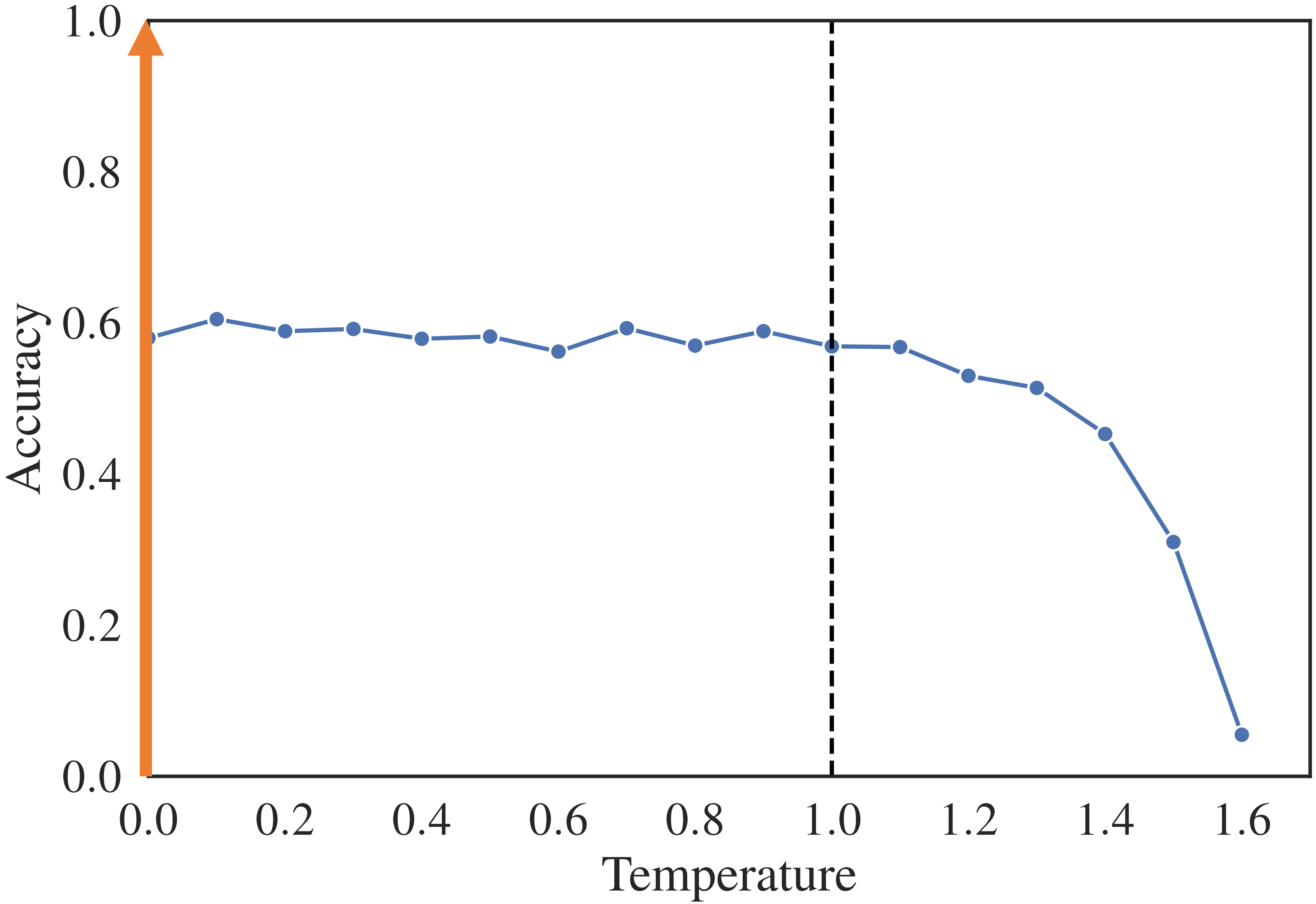
Results



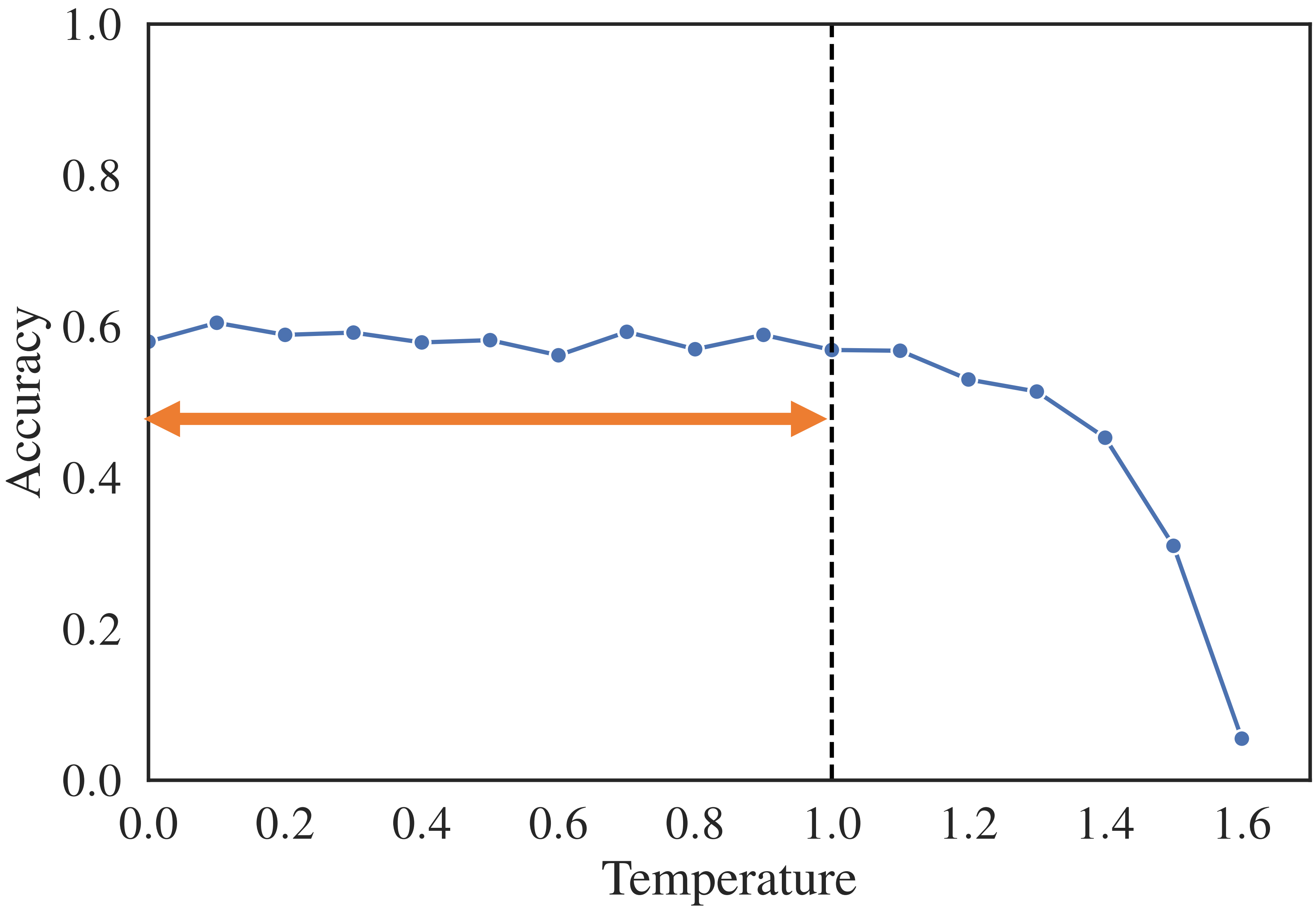
Accuracy by temperature from 0.0 to 1.6 for GPT-3.5 using CoT prompt on the 100-question exam.



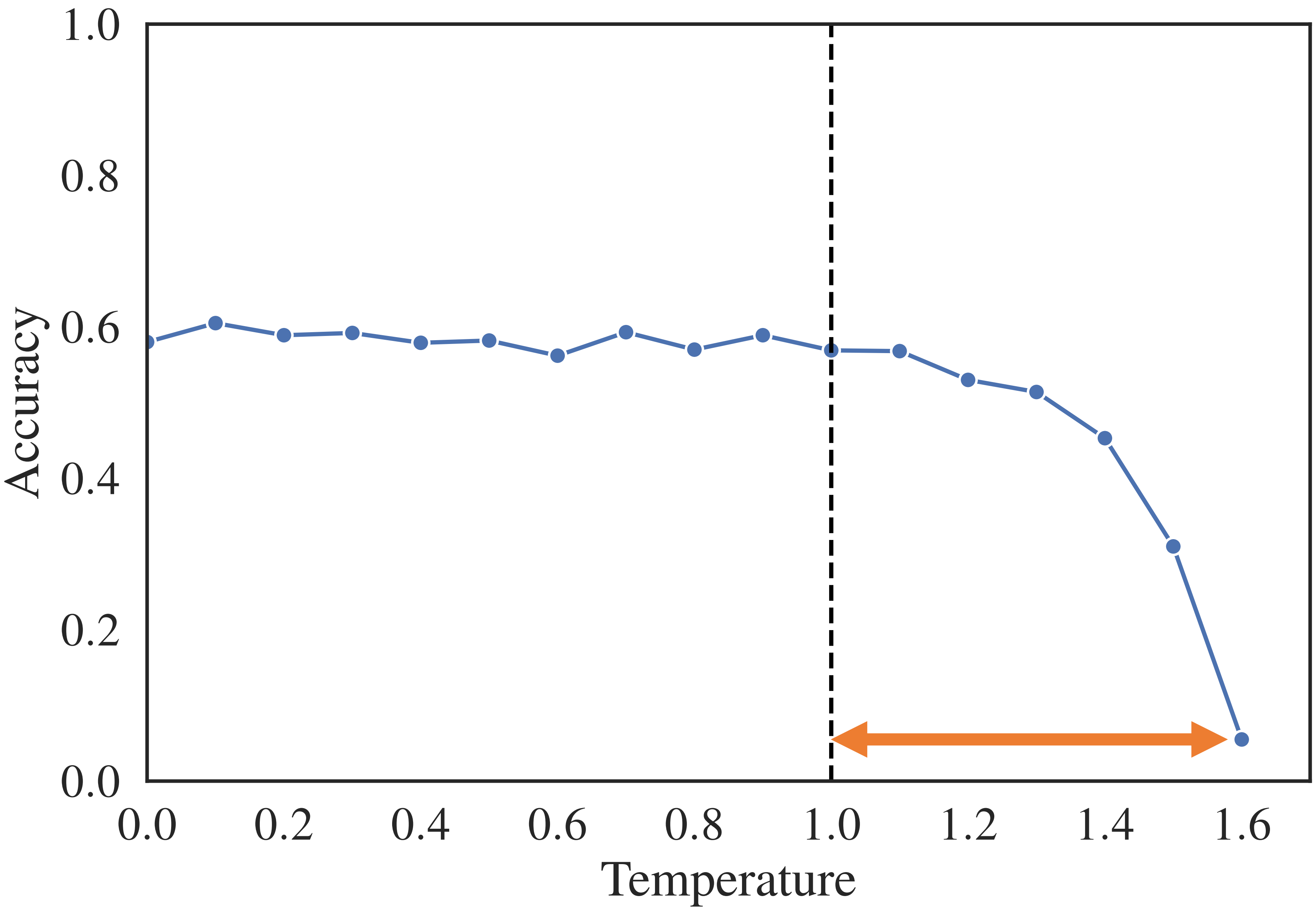
Accuracy by temperature from 0.0 to 1.6 for GPT-3.5 using CoT prompt on the 100-question exam.



Accuracy by temperature from 0.0 to 1.6 for GPT-3.5 using CoT prompt on the 100-question exam.



Accuracy by temperature from 0.0 to 1.6 for GPT-3.5 using CoT prompt on the 100-question exam.



Accuracy by temperature from 0.0 to 1.6 for GPT-3.5 using CoT prompt on the 100-question exam.

Quantitative Results

GPT-3.5,

CoT prompt,

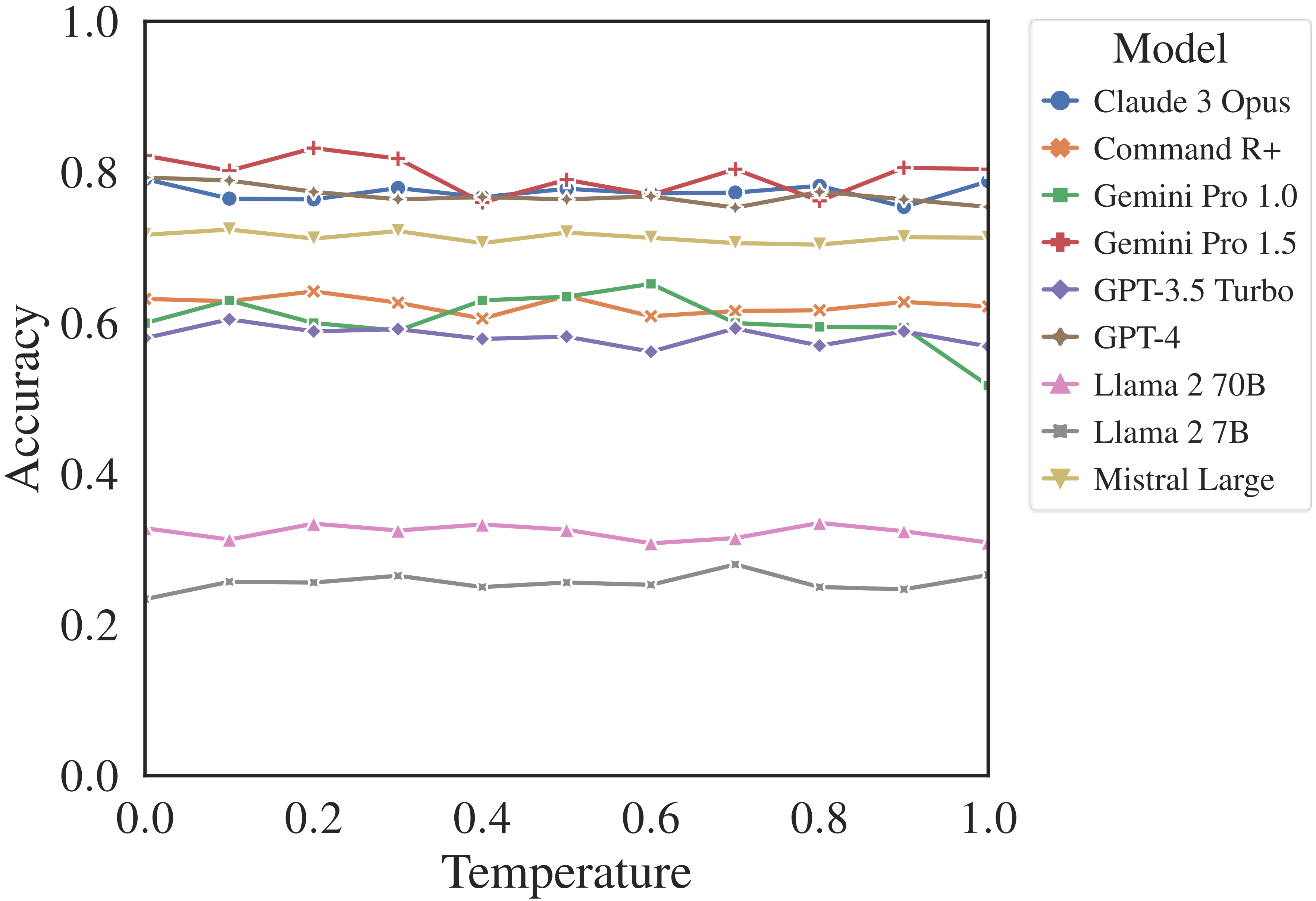
1,000 questions

Range of 0.0 to 1.0

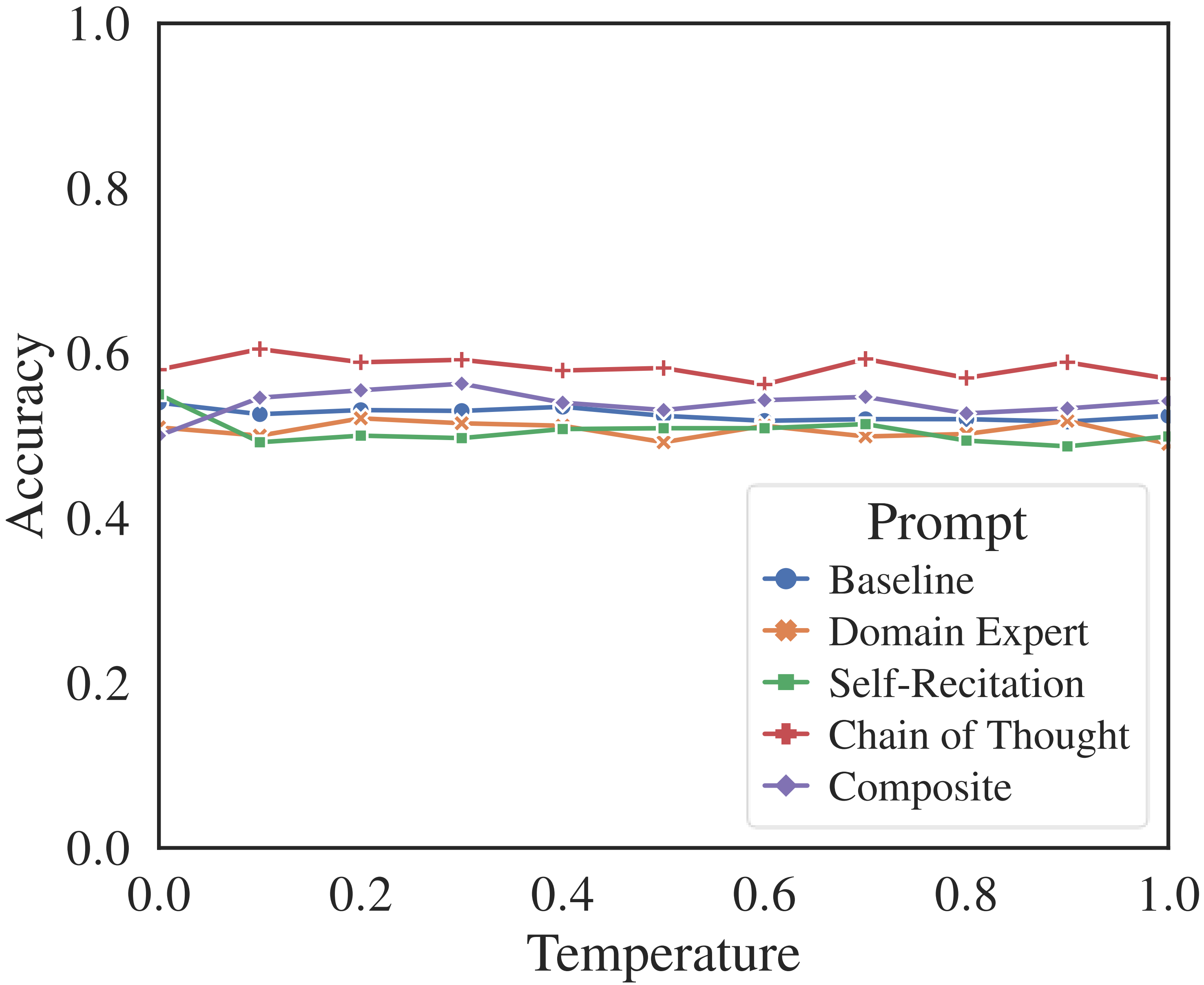
Quantitative Results

GPT-3.5,
CoT prompt,
1,000 questions
Range of 0.0 to 1.0

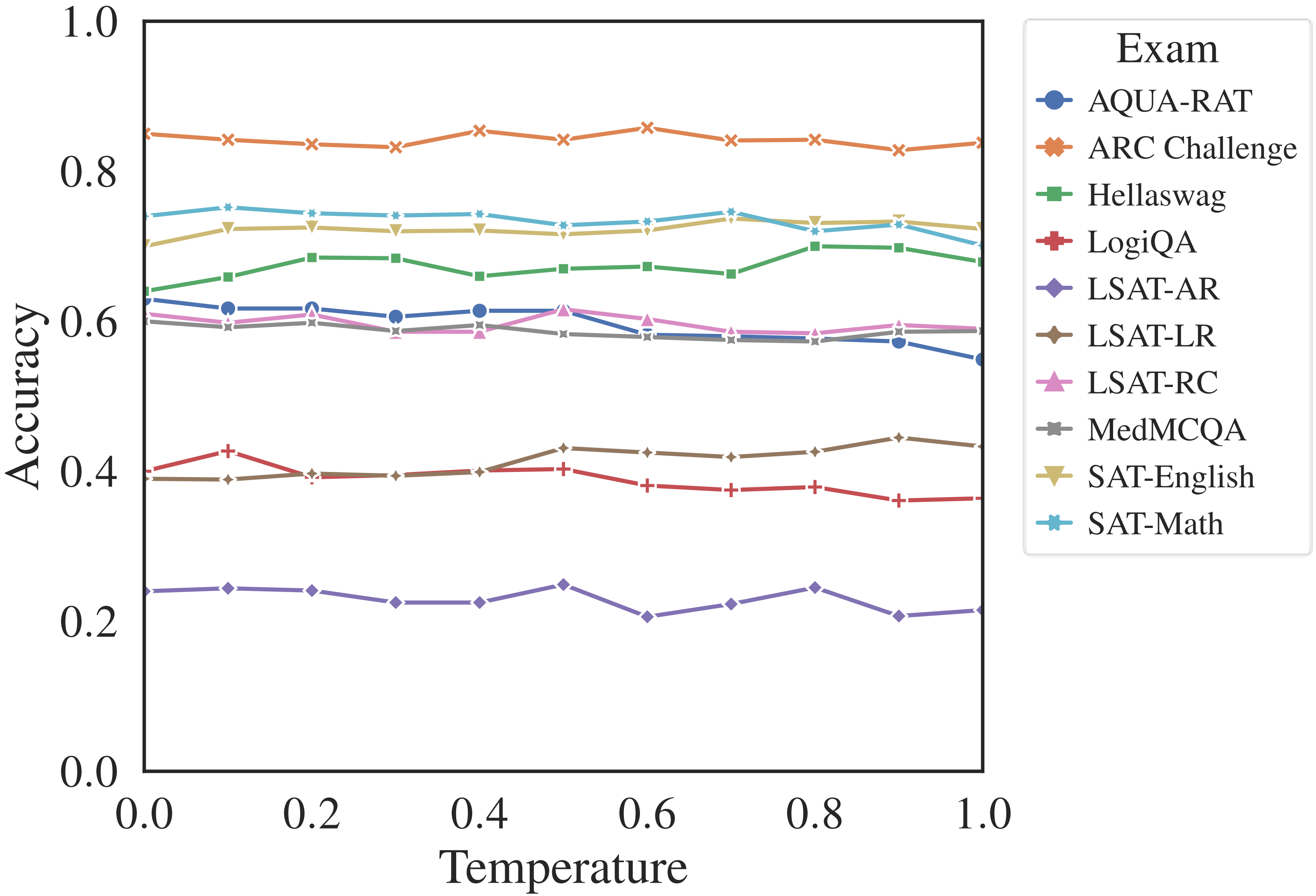
Kruskal-Wallis Test
$H(10) = 10.439$
$p = 0.403$
$p > 0.05$



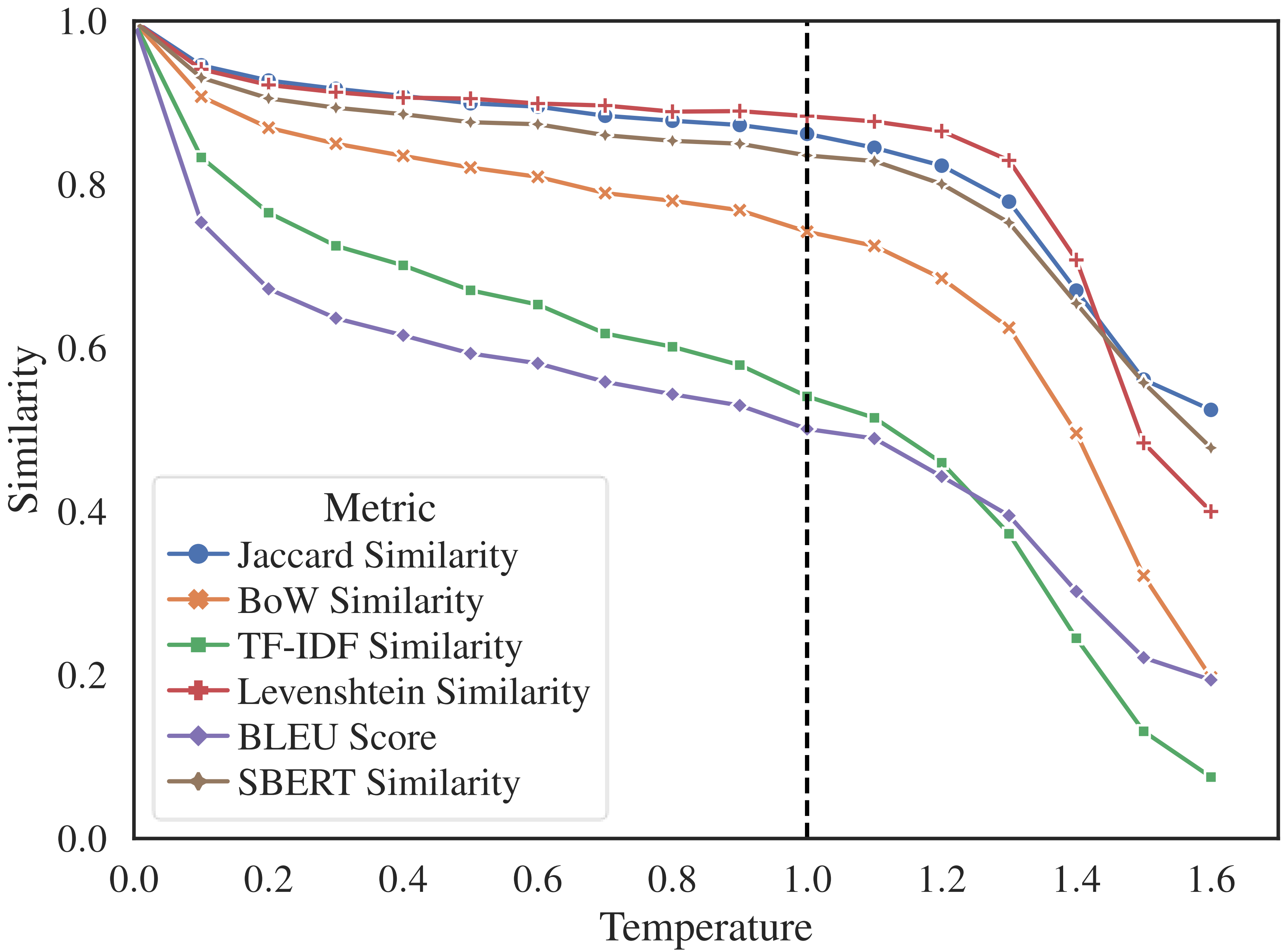
Accuracy by temperature and model using the CoT prompt on the 100-question exam.



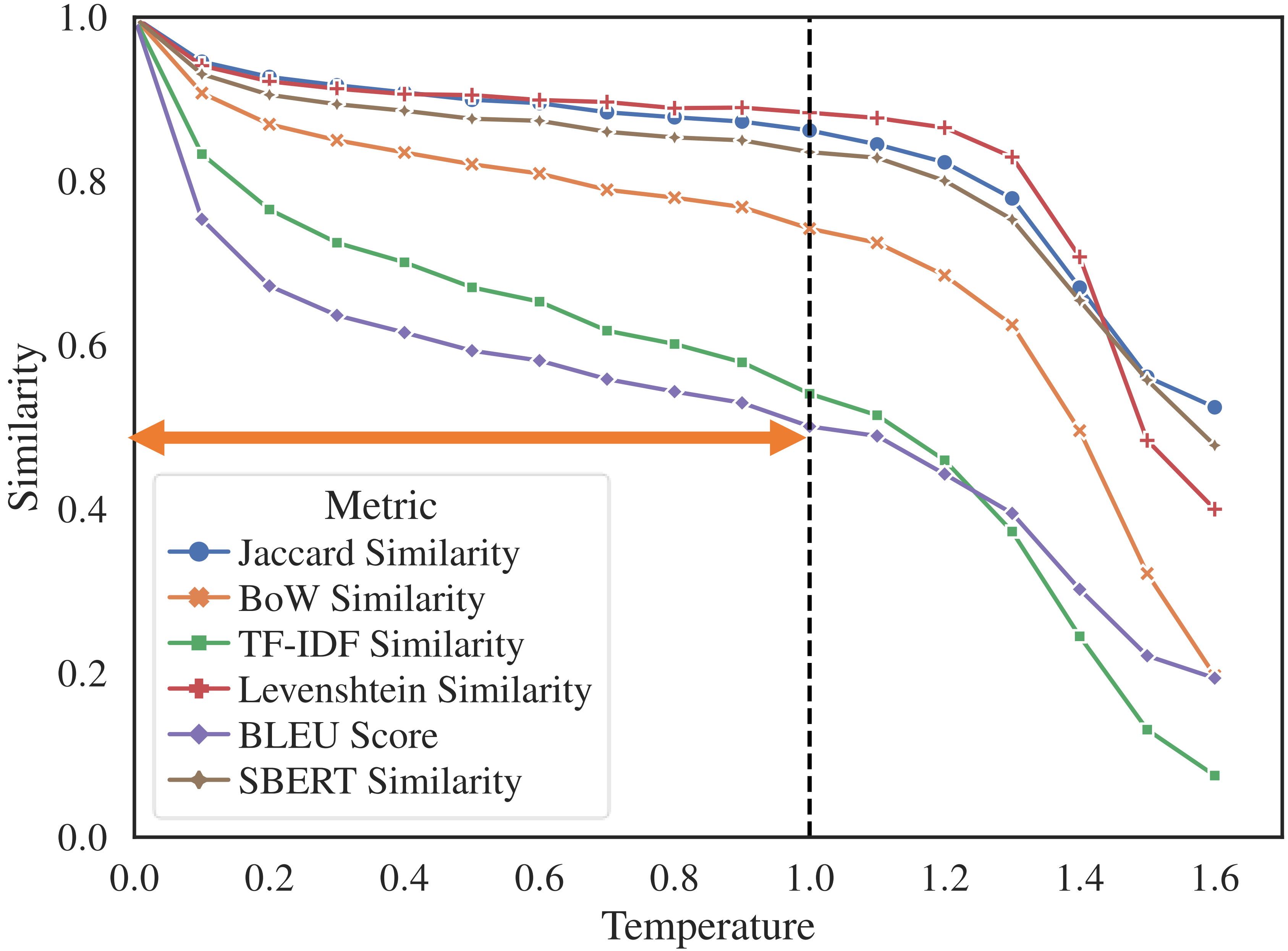
Accuracy by temperature and prompt for GPT-3.5 using the CoT prompt on the 100-question exam.



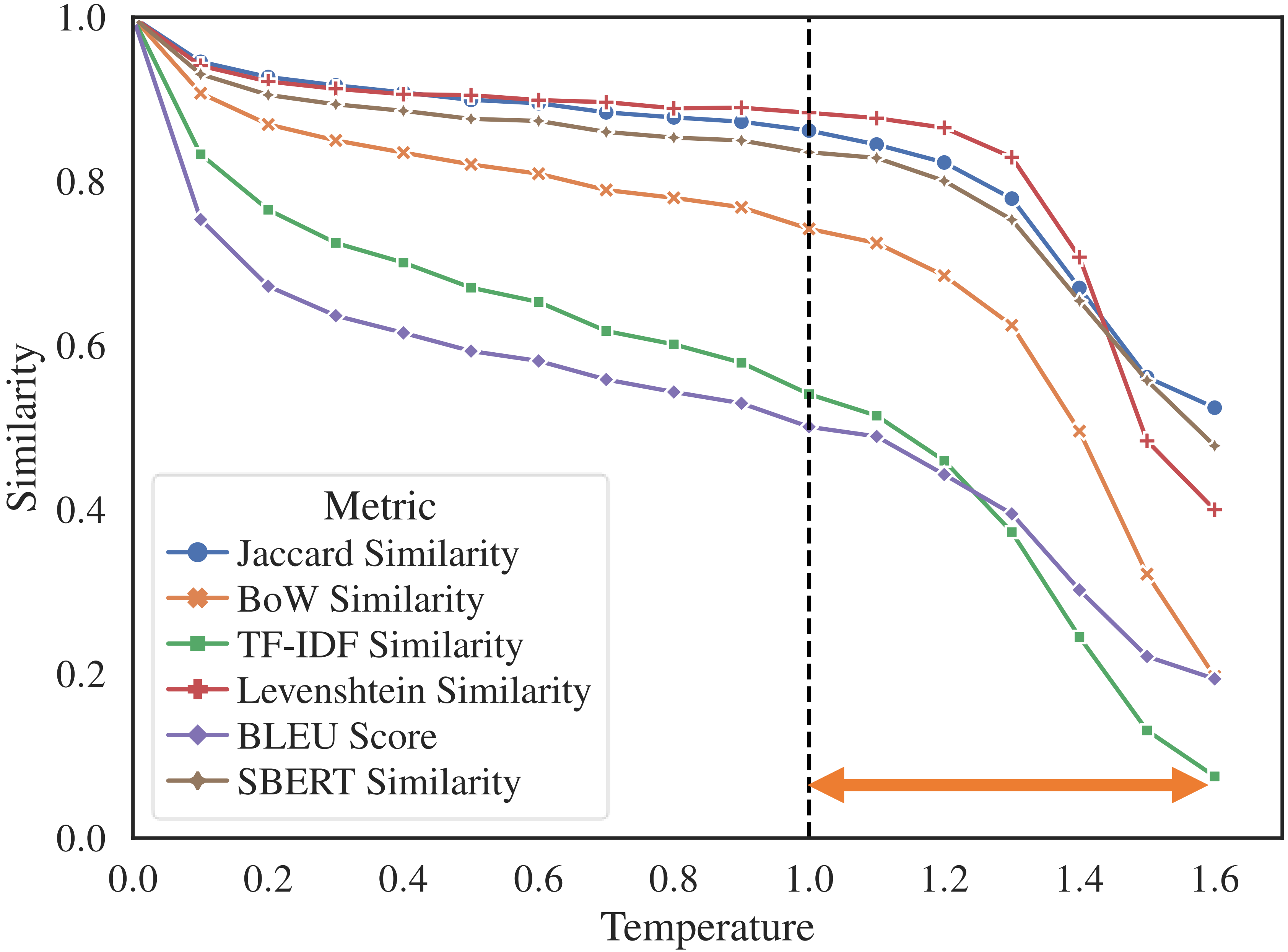
Accuracy by temperature and exam for GPT-3.5 using the CoT prompt.



Text similarity by temperature for GPT-3.5 using the CoT prompt on the 100-question exam.



Text similarity by temperature for GPT-3.5 using the CoT prompt on the 100-question exam.



Text similarity by temperature for GPT-3.5 using the CoT prompt on the 100-question exam.

Discussion

Interpretation

Temperature does not impact performance on MCQA problems.

Interpretation

Temperature does not impact performance on MCQA problems.

Always set temperature to 0.0 for problem-solving tasks.

Interpretation

Temperature does not impact performance on MCQA problems.

Always set temperature to 0.0 for problem-solving tasks.

Maximizes reproducibility

Interpretation

Temperature does not impact performance on MCQA problems.

Always set temperature to 0.0 for problem-solving tasks.

Maximizes reproducibility

Avoids the drop-off

Interpretation

Temperature does not impact performance on MCQA problems.

Always set temperature to 0.0 for problem-solving tasks.

Maximizes reproducibility

Avoids the drop-off

Minimizes tokens

Limitations

Limitations

Only 9 LLMs

Only 5 prompts

Only 10 domains

Limitations

Only 9 LLMs

Only 5 prompts

Only 10 domains

Only 1,000 problems

Limitations

Only 9 LLMs

Only 5 prompts

Only 10 domains

Only 1,000 problems

Only 1 hyperparameter

Implications

Implications

Saves AI engineers time/effort

Implications

Saves AI engineers time/effort

Reduces unproductive debate

Implications

Saves AI engineers time/effort

Reduces unproductive debate

Insight into model hallucination

Implications

Saves AI engineers time/effort

Reduces unproductive debate

Insight into model hallucination

Insight into solution-space search

Future Research

Future Research

Other LLMs

Other domains

Other problems

Future Research

Other LLMs

Other domains

Other problems

All temperatures

Future Research

Other LLMs

Other domains

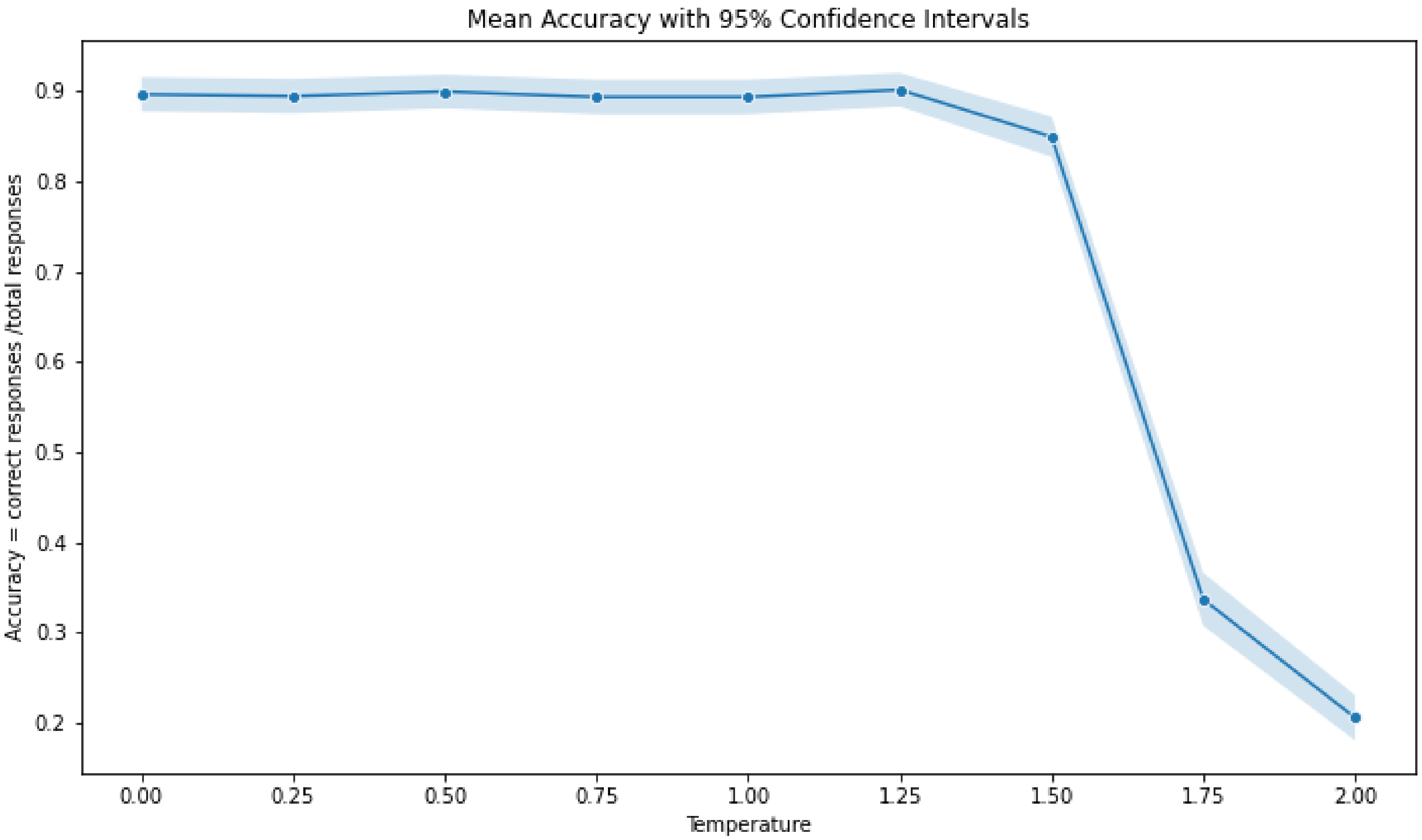
Other problems

All temperatures

In-depth error analysis

The Impact of Temperature on the Performance of Large Language Model Systems and Business Applications

Michael Gou



https://www.tickr.com/blog/posts/impact-of-temperature-on-llms/

Conclusion

Conclusion

What is the effect of temperature for LLMs on problem-solving?

Conclusion

What is the effect of temperature for LLMs on problem-solving?

Temperature does not impact performance on MCQA problems.

Conclusion

What is the effect of temperature for LLMs on problem-solving?

Temperature does not impact performance on MCQA problems.

Always set temperature to 0.0 for problem-solving tasks.

Learn more

https://matthewrenze.com/research/the-effect-of-sampling-temperature-on-llms/