Text Encoders Lack Knowledge: Leveraging Generative LLMs for Domain-Specific Semantic Textual Similarity

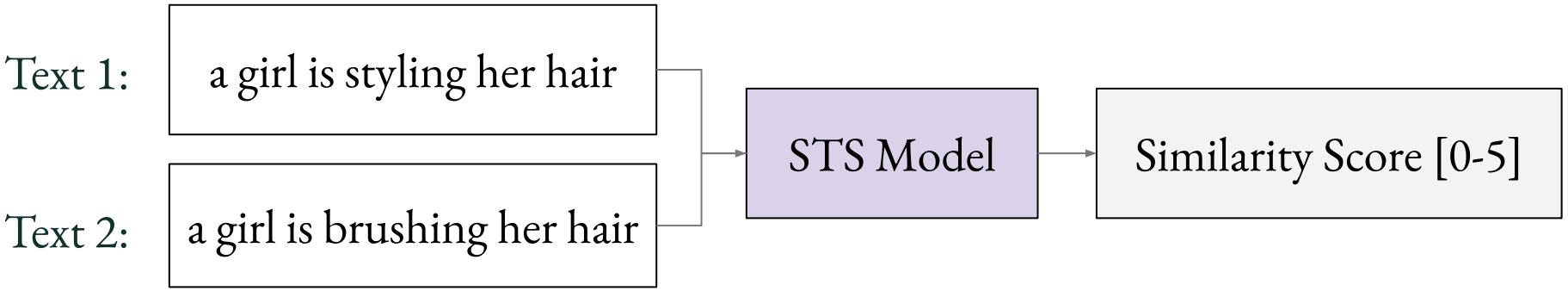
Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman, Sarah M. Preum

Dartmouth College

joseph.m.gatto.gr@dartmouth.edu

Semantic Textual Similarity (STS)

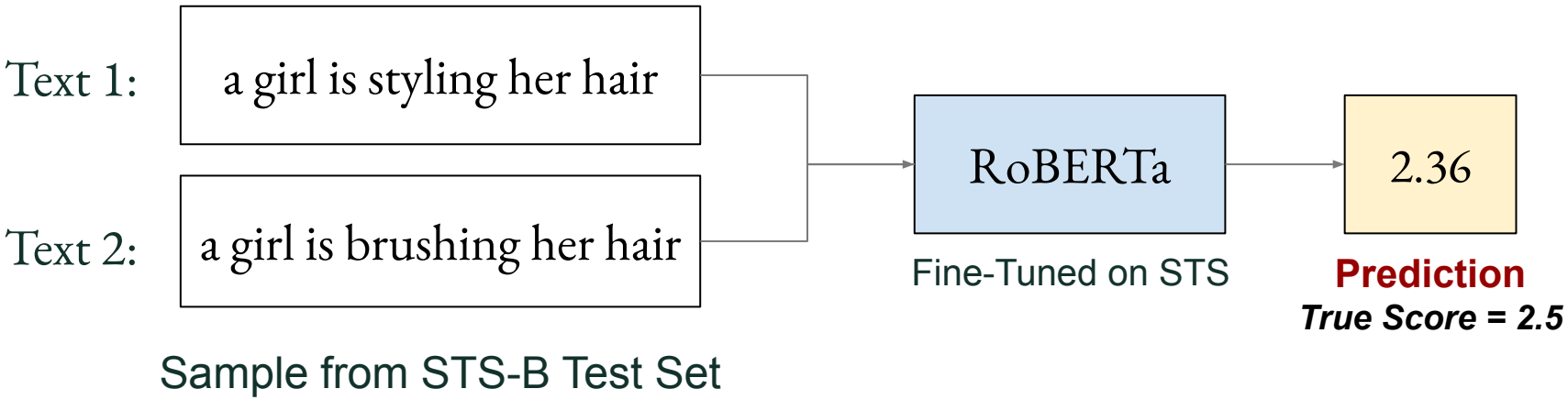
STS Objective: Predict a continuous value between 0-5 that represents the semantic similarity between two texts



Sample from STS-B Test Set

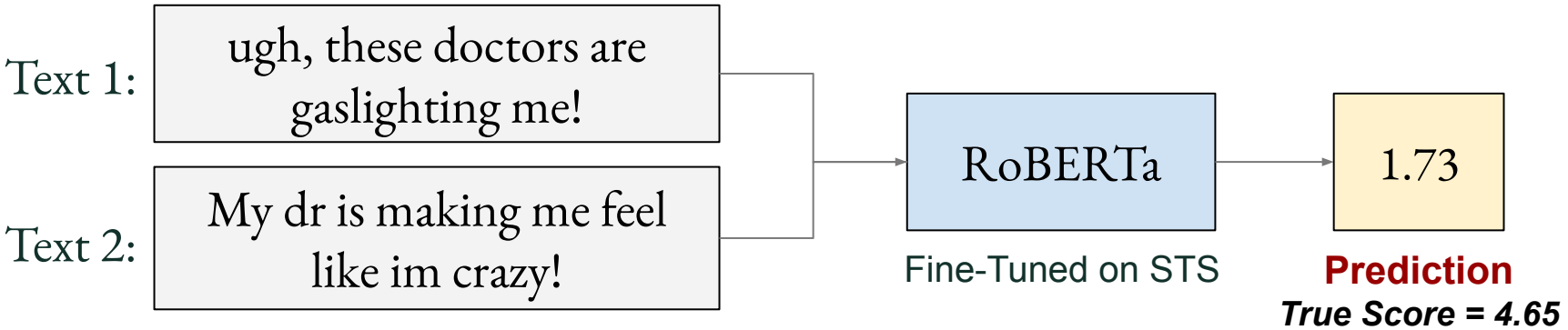
Semantic Textual Similarity (STS)

👍 ***Existing STS models work well* on simple texts**



Semantic Textual Similarity (STS)

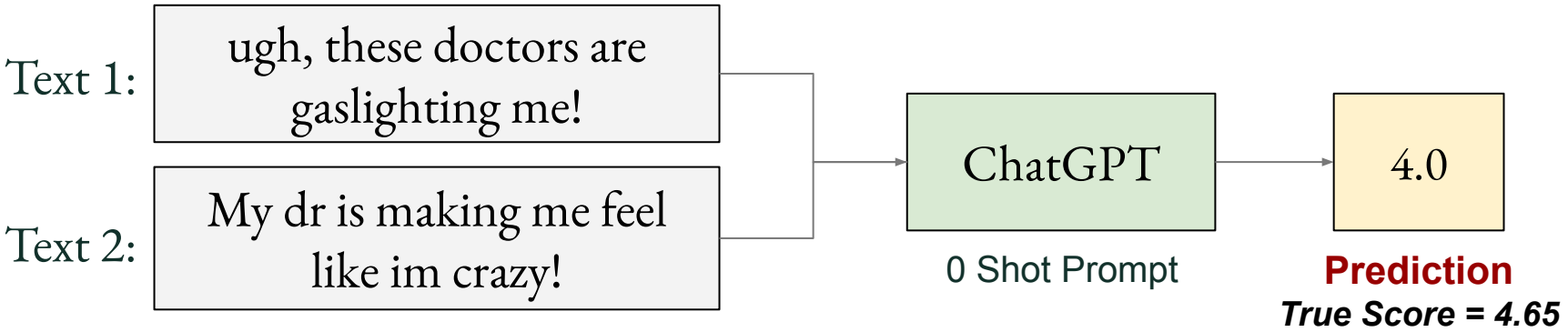
❌ ***However, they struggle to model the relationship between complex domain-specific texts***



Sample from our STS-Health Dataset

Semantic Textual Similarity (STS)

We show that **LLMs significantly outperform text encoders** when predicting STS between **complex text pairs** which require significant **world knowledge** to understand!



Sample from our STS-Health Dataset

Why Use Text Generation for a Regression Task?

Recent works have claimed that LLMs have **“no use case”** on regression tasks such as STS-B*

* Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. arXiv preprint arXiv:2304.13712.

Why Use Text Generation for a Regression Task?

Intuition: **3 Reasons Why LLMs are Well-Suited for STS**

Why Use Text Generation for a Regression Task?

🍐 LLMs are great at pairwise textual comparisons → ***Making them well-suited for STS***

Why Use Text Generation for a Regression Task?

🍐 LLMs are great at pairwise textual comparisons → ***Making them well-suited for STS***

🧠 LLMs contain significant *world knowledge* → ***Remove need for domain-specific STS annotation***

Why Use Text Generation for a Regression Task?

🍐 LLMs are great at pairwise textual comparisons → ***Making them well-suited for STS***

🧠 LLMs contain significant *world knowledge* → ***Remove need for domain-specific STS annotation***

🔢 LLMs are familiar with percentages → ***STS is a percentage prediction task!***

Research Questions

1. What is the optimal LLM prompting strategy for STS?

2. How do LLMs compare to text encoders on popular STS benchmarks?
   Eval Sets: STS12-16, STS-B, SICK-R

3. How do LLMs compare to text encoders on domain-specific STS challenge sets?
   Domains: Health 🚑, News 📰, Sports 🏀

Methods: Optimal STS Prompting Strategies

💡 **Key Idea:** Reformulate STS as a percentage prediction task

Prompt: Output a number between **0 and 5** describing the semantic similarity between the following two sentences:

Sentence 1: <text>
Sentence 2: <text>

Prior works only explore LLMs for STS in the 0-5 scale

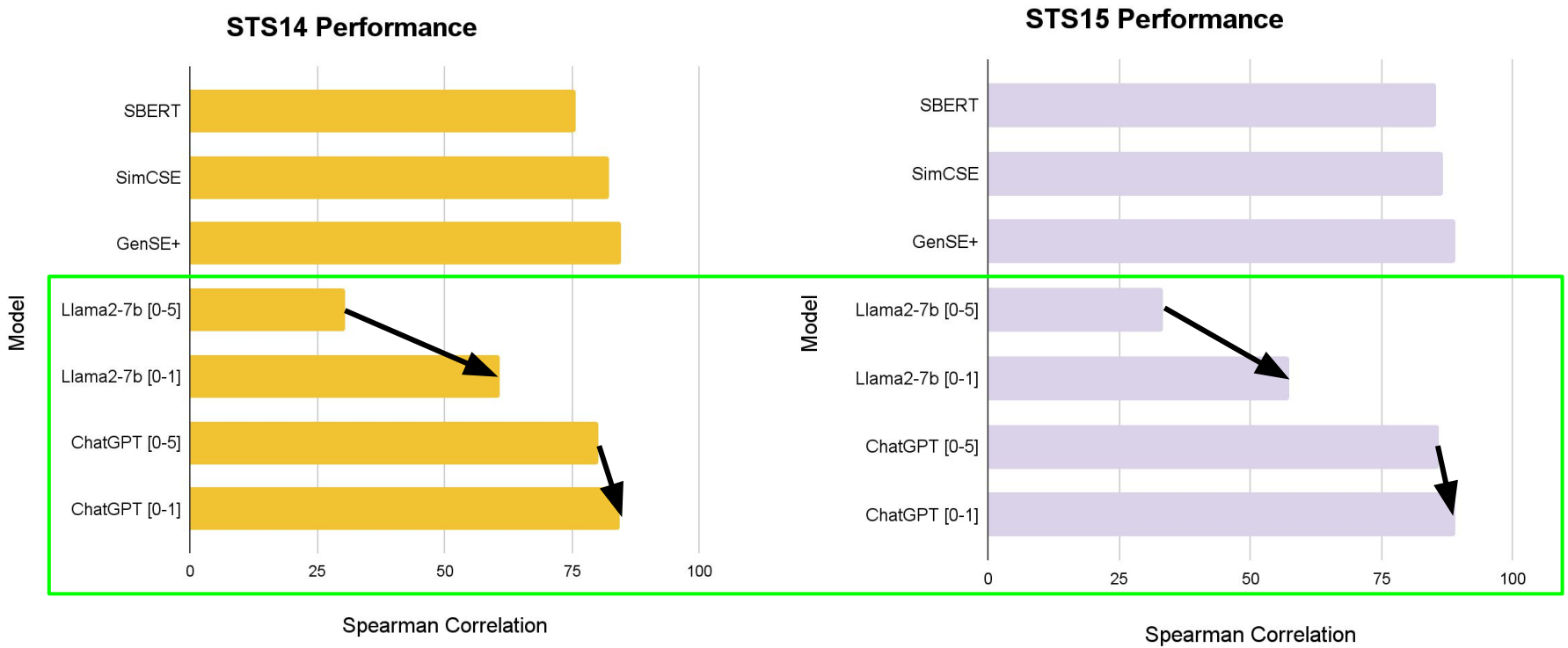
Prompt: Output a number between **0 and 1** describing the semantic similarity between the following two sentences:

Sentence 1: <text>
Sentence 2: <text>

We map the labels between 0-1 so that LLMs can leverage learned knowledge of percentages

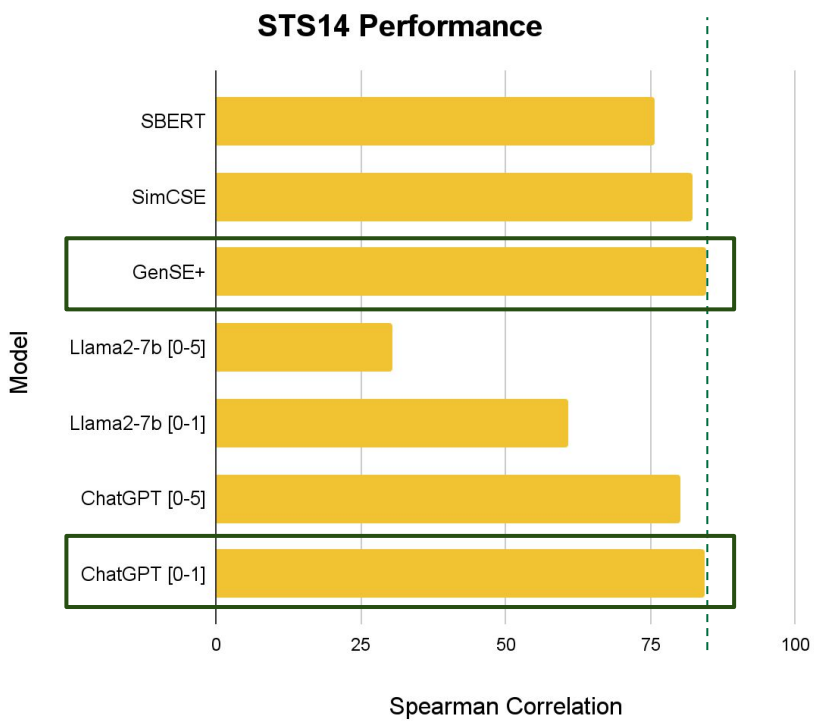
RQ1: What is the optimal prompting strategy for STS?

Result: 💡Prompting for percentages improves performance on 6/7 datasets!



RQ2: How do LLMs compare to Text Encoders on STS Benchmarks?

Result: 💡On 6/7 benchmark datasets, our LLM approach performs on-par or better than SOTA text encoders!



RQ3: How do LLMs compare to text encoders on domain-specific STS challenge sets?

Three STS Challenge Sets (n=100) collected *after* May 2023

1. **STS-Sports** 🏀
   Data Source: r/NBA, r/NBATalk, r/NFL

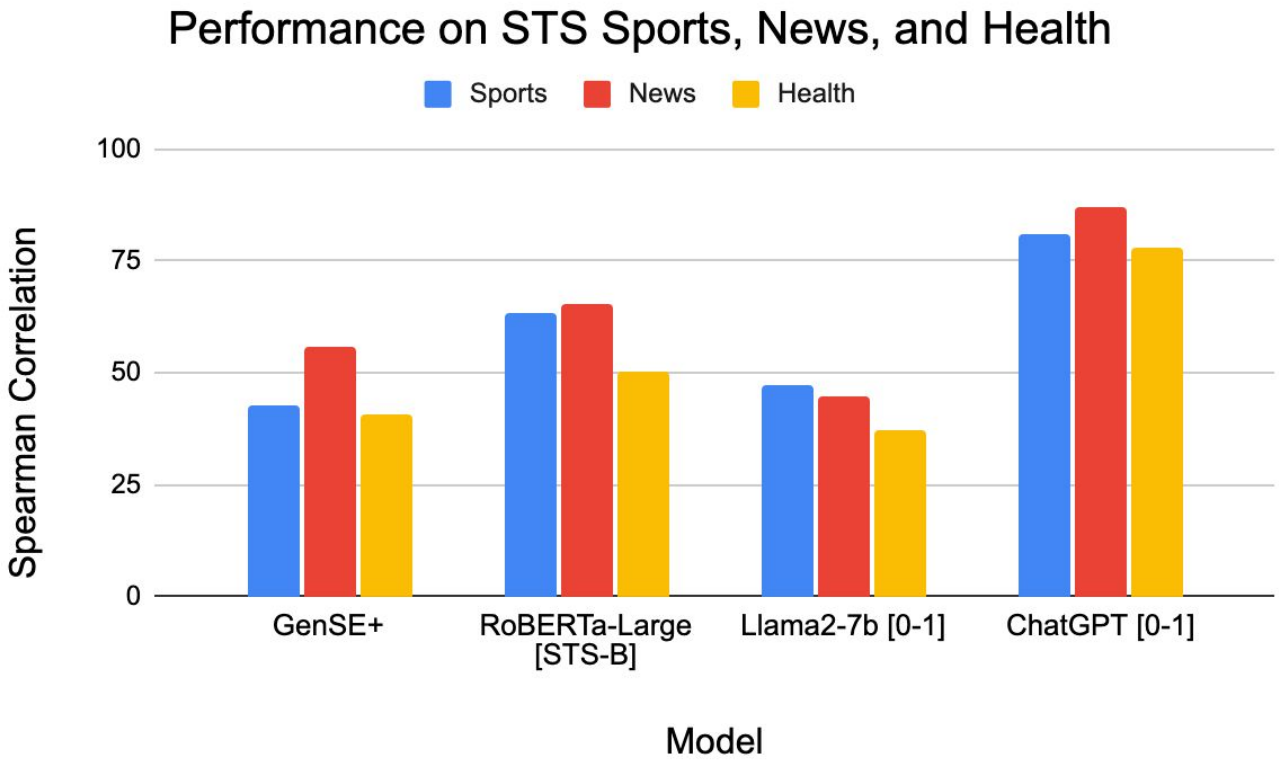
2. **STS-Health** 🚑
   Data Source: r/covidlonghaulers, r/LongCovid

3. **STS-News** 📰
   Data Source: r/politics

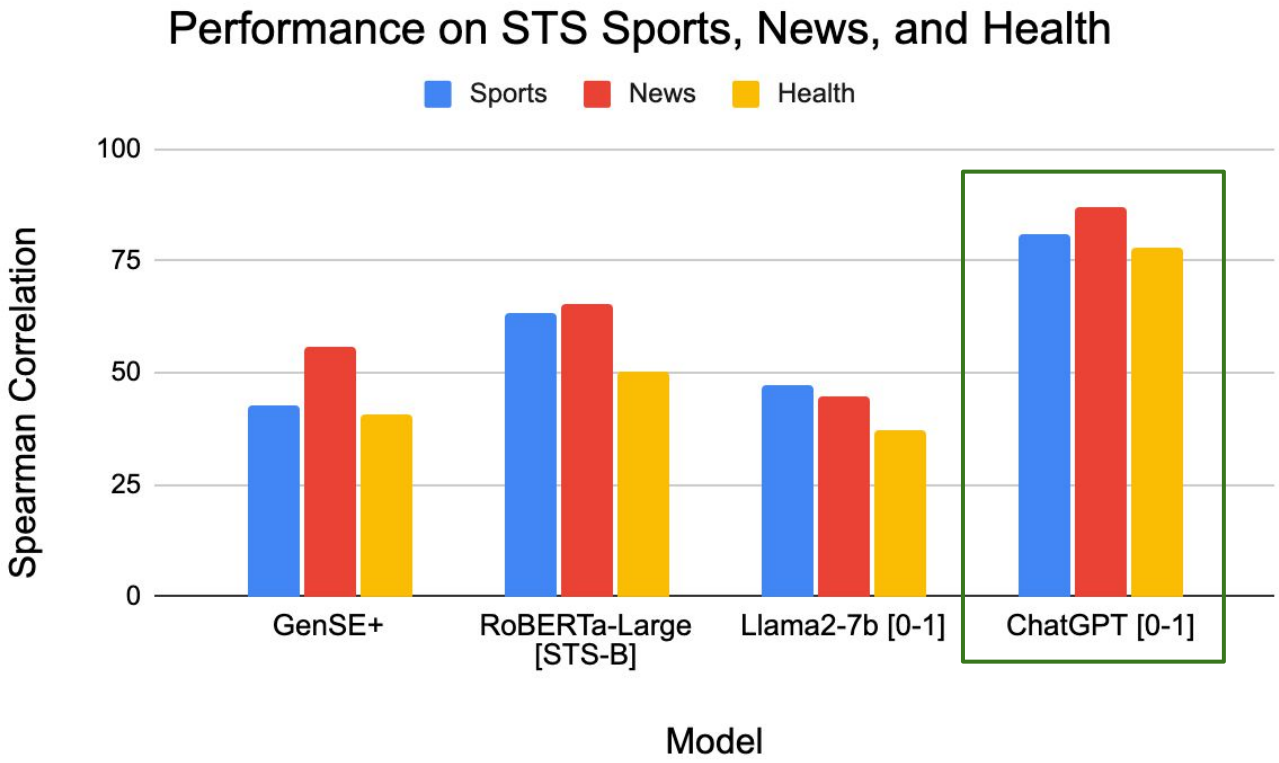
📰 ***STS-News Example***

Text 1: Governors Kathy Hochul & Phil Murphy join 8 other governors in opposing school textbook censorship
Text 2: The New Jersey and New York governors oppose censorship of school textbooks.
Label: 0.86/1.0

RQ3: How do LLMs compare to text encoders on domain-specific STS challenge sets?



RQ3: How do LLMs compare to text encoders on domain-specific STS challenge sets?



Limitation: The ChatGPT API is not free! 💰 💰 💰

This method may only be well-suited to ***small-scale tasks*** requiring STS

Key Takeaways

1. Reformulating STS in the context of percentage prediction improves LLMs' performance on STS

2. LLMs can perform on-par with SOTA unsupervised text encoders on popular STS benchmarks

3. LLMs significantly outperform existing STS models on complex, domain-specific STS samples

***Text Encoders Lack Knowledge*: Leveraging Generative LLMs for Domain-Specific Semantic Textual Similarity**

Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman, Sarah M. Preum
Department of Computer Science, Dartmouth College
{joseph.m.gatto.gr, sarah.masud.preum} @ dartmouth.edu

Thank You!

Data

Full Paper