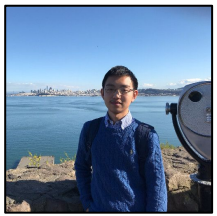
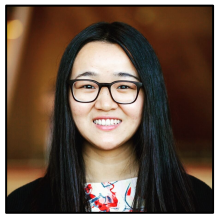
Can Large Language Models *Transform* Computational Social Science?

Caleb Ziems[†*], William Held[✦*], Omar Shaikh[†*], Jiaao Chen[✦*], Zhehao Zhang[‡*], Diyi Yang[†]





* All heavily contributed to the implementation of this work

Stanford NLP

[†]Stanford

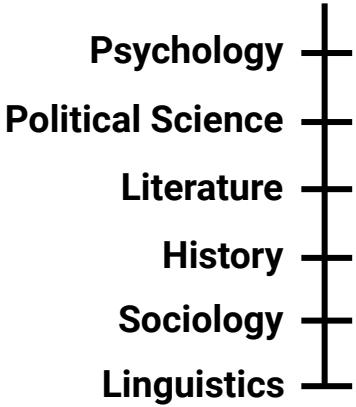
[✦]Georgia Tech

[‡]Dartmouth

Social and Language Technologies
SALT

Research **Questions**

RQ: Are LLMs useful tools in the **Computational Social Scientist's** toolkit?



Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's **toolkit?**

(Supervised) **Text Classification**

Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's **toolkit?**

(Supervised) — **Text Classification**

(Unsupervised) — **Text Clustering**

Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's **toolkit?**

(Supervised) — **Text Classification**

(Unsupervised) — **Text Clustering**

(Semi-supervised) — **Natural Language Generation**

Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's toolkit?

1. **Viability**
2. **Model-Selection**
3. **Domain-Utility**
4. **Functionality**

Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's toolkit?

1. **Viability**
2. **Model-Selection**
3. **Domain-Utility**
4. **Functionality**

Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's toolkit?

1. **Viability**
2. **Model-Selection**
3. **Domain-Utility**
4. **Functionality**

Research **Questions**

RQ: Are LLMs useful tools in the Computational Social Scientist's toolkit?

1. **Viability**
2. **Model-Selection**
3. **Domain-Utility**
4. **Functionality**

Research **Questions**

RQ: Are LLMs useful **tools** in the **Computational Social Scientist's** toolkit?

politeness recognition humor recognition emotion recognition empathy classification	**Psychology**
stance detection ideology detection	**Political Science**
agent framing relationship dynamics	**Literature**
event extraction	**History**
power relations identification social role detection	**Sociology**
dialect feature identification	**Linguistics**

Research Questions

RQ: Are LLMs useful **tools** in the **Computational Social Scientist's** toolkit?

politeness recognition	**Psychology**	**Stanford Politeness Corpus** (Danescu-Niculescu-Mizil et al., 2013)
humor recognition		**r/Jokes + Pun of the Day** (Weller and Seppi 2019)
emotion recognition		**CARER** (Saravia et al. 2018)
empathy classification		**EPITOME** (Sharma et al., 2020)
stance detection	**Political Science**	**SemEval-2016 Stance Dataset** (Mohammad et al., 2016)
ideology detection		**Ideological Books Corpus** (Gross et al., 2013)
agent framing	**Literature**	**Article Bias Corpus** (Baly et al. 2020)
relationship dynamics		**WikiEvents** (Li et al., 2021)
event extraction	**History**	**Hippocorpus** (Sap et al., 2020)
power relations identification	**Sociology**	**Wikipedia Talk Pages** (Danescu-Niculescu-Mizil et al. 2012)
social role detection		**CMU Movie Corpus** (Bamman et al. 2013)
dialect feature identification	**Linguistics**	**Indian English Minimal Pairs** (Demszky et al. 2019)

RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

RQ1: Zero-Shot Classification Performance

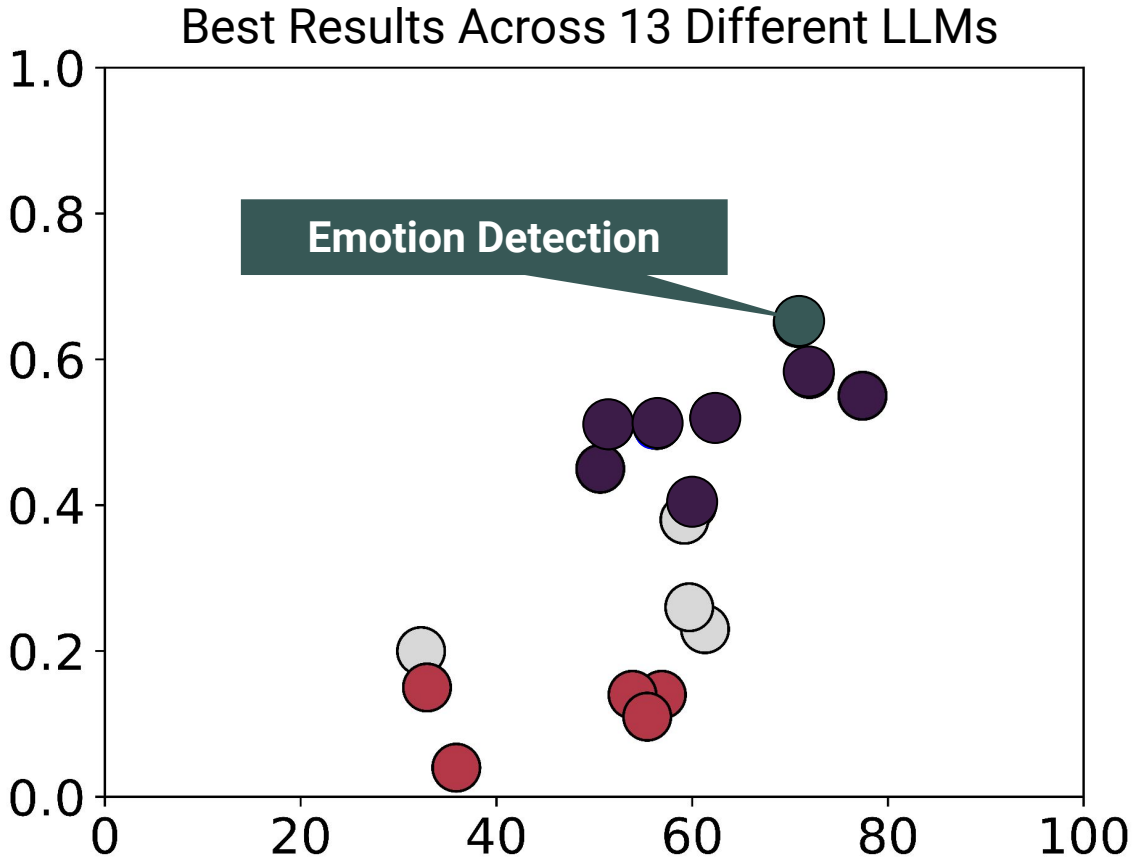
RQ1: Viability. Can LLMs augment the human annotation pipeline?

↪ **Finding: LLMs can make annotation more efficient** but we still need humans in the loop

RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

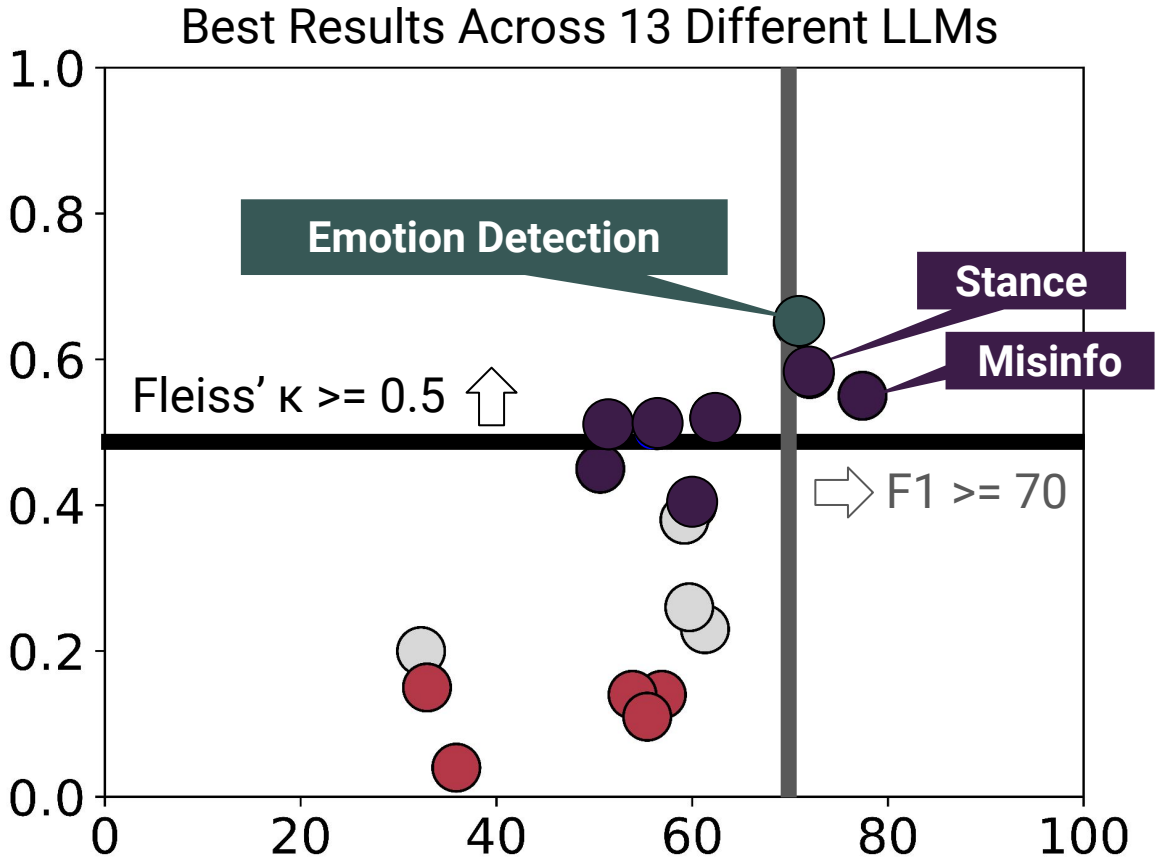
↪ **Finding: LLMs can make annotation more efficient** but we still need humans in the loop



RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

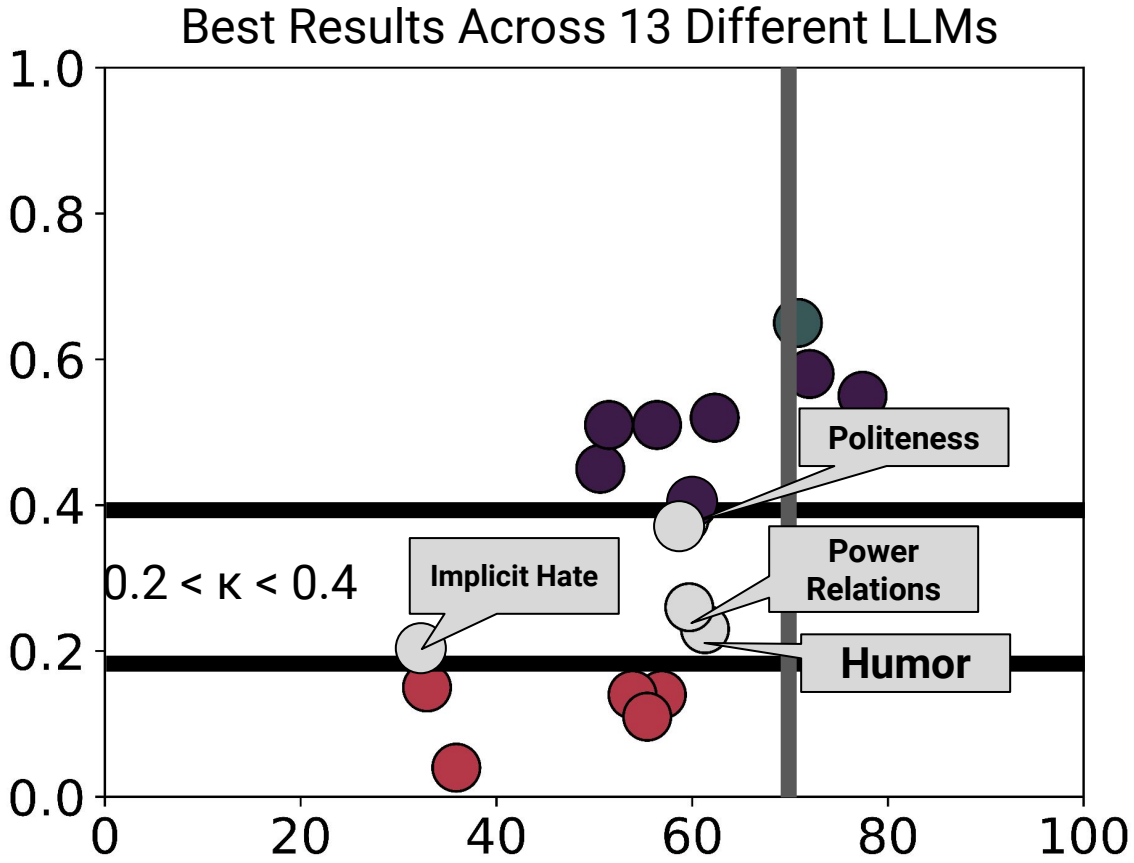
↪ **Finding: LLMs can make annotation more efficient** but we still need humans in the loop



RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

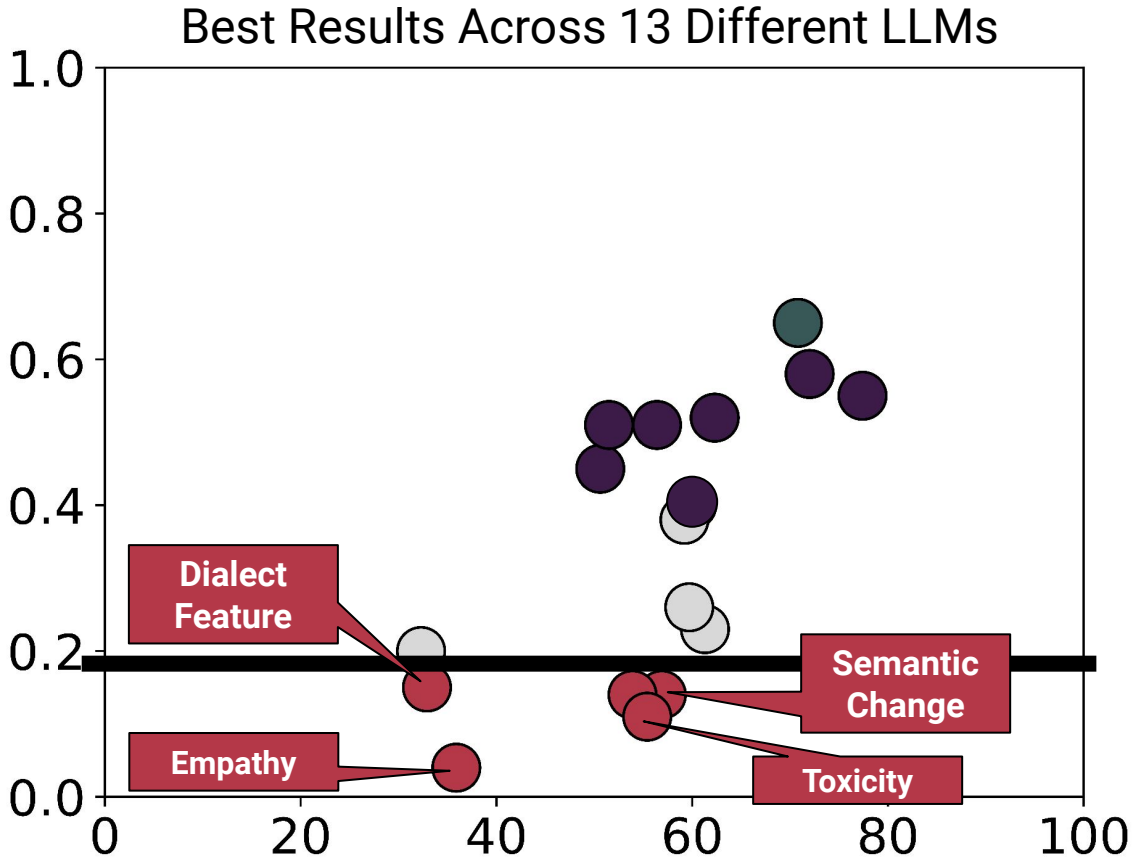
↪ **Finding:** LLMs can make annotation more efficient **but we still need humans in the loop**



RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

↪ **Finding:** LLMs can make annotation more efficient **but we still need humans in the loop**

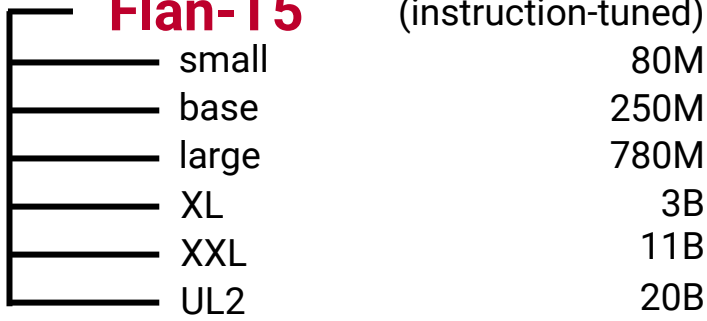


RQ2: CSS Performance Follows Scaling Laws

RQ2: Model-Selection.

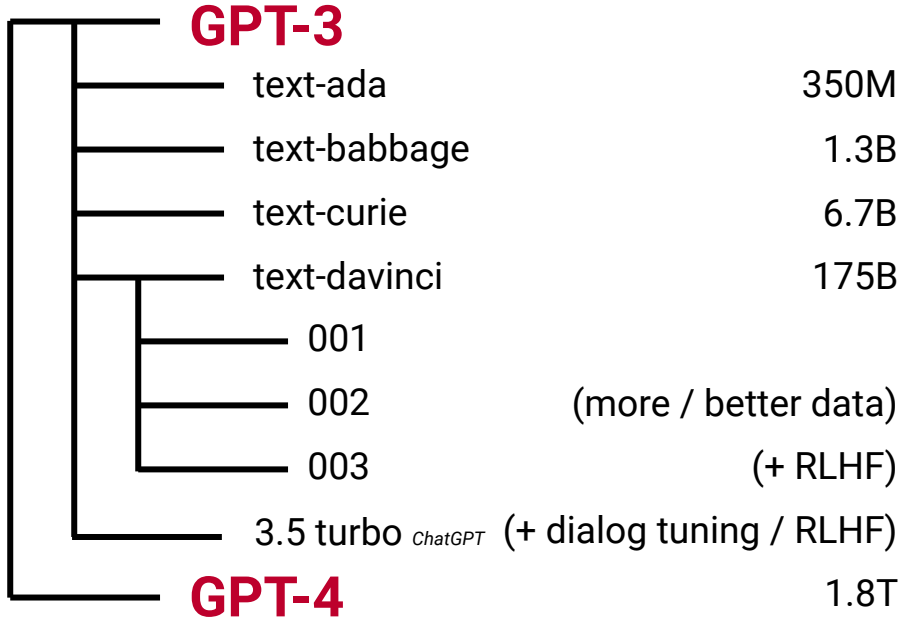
RQ2: CSS Performance Follows Scaling Laws

RQ2: Model-Selection.



RQ2: CSS Performance Follows Scaling Laws

RQ2: Model-Selection.



RQ2: CSS Performance Follows Scaling Laws

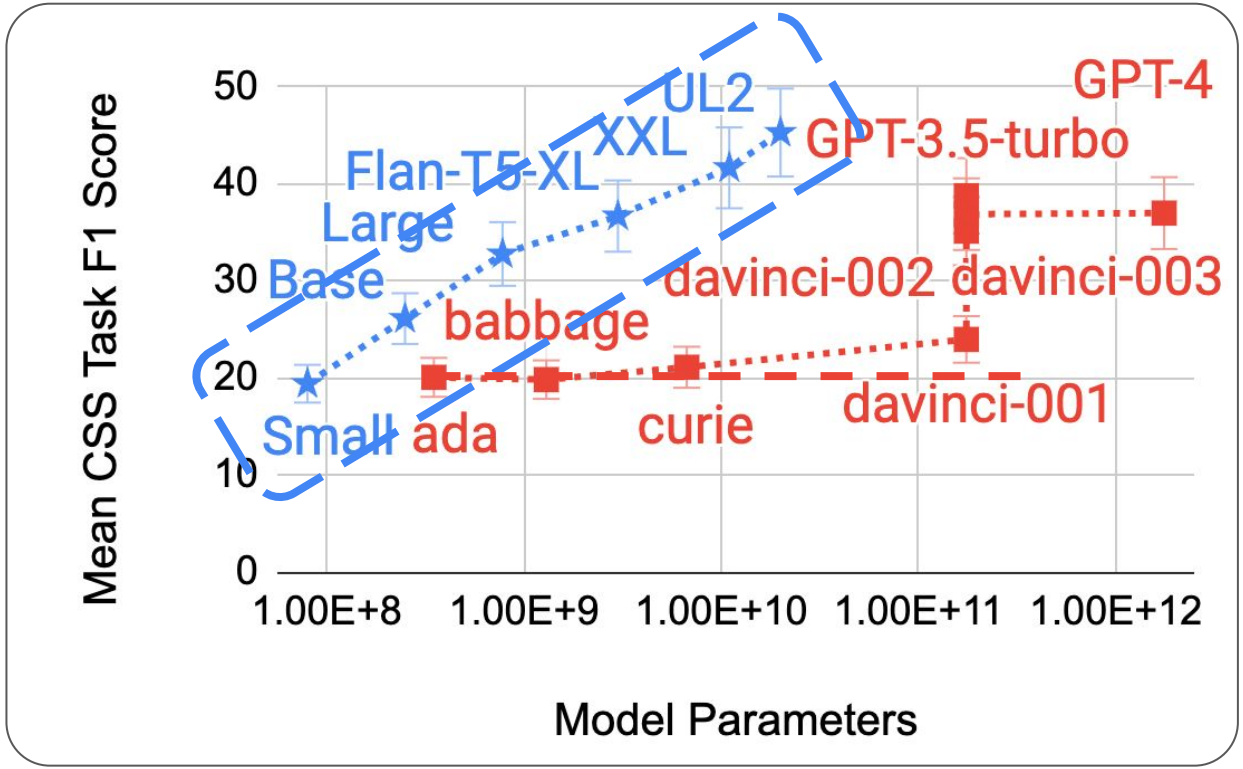
RQ2: Model-Selection.
How does model size, architecture and pretraining affect downstream performance on CSS tasks?

↪ **Findings:** Performance ***scales with*** model size

RQ2: CSS Performance Follows Scaling Laws

RQ2: Model-Selection. How does model size, architecture and pretraining affect downstream performance on CSS tasks?

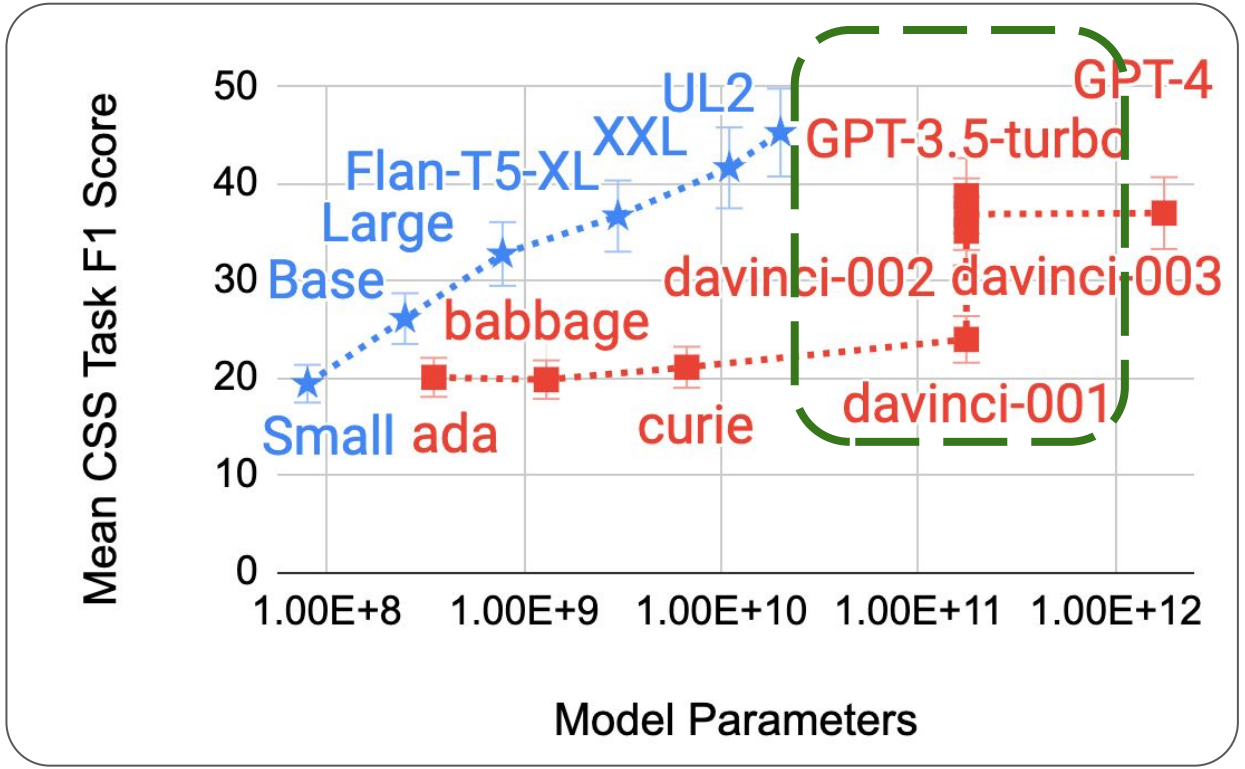
↪ **Findings:** Performance ***scales with*** model size



RQ2: CSS Performance Follows Scaling Laws

RQ2: Model-Selection. How does model size, architecture and pretraining affect downstream performance on CSS tasks?

↪ **Findings:** Performance ***scales with*** model size



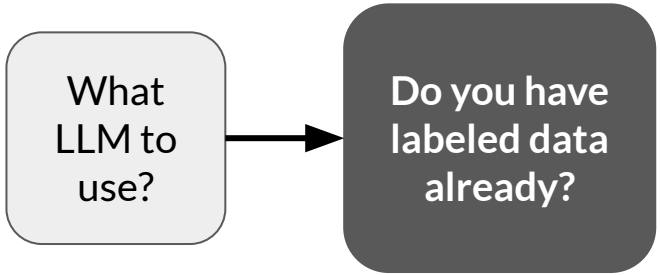
RQ2: Scaling Laws – Benefits of Open Source

Recommendation:

What LLM to use?

RQ2: Scaling Laws – Benefits of Open Source

Recommendation:

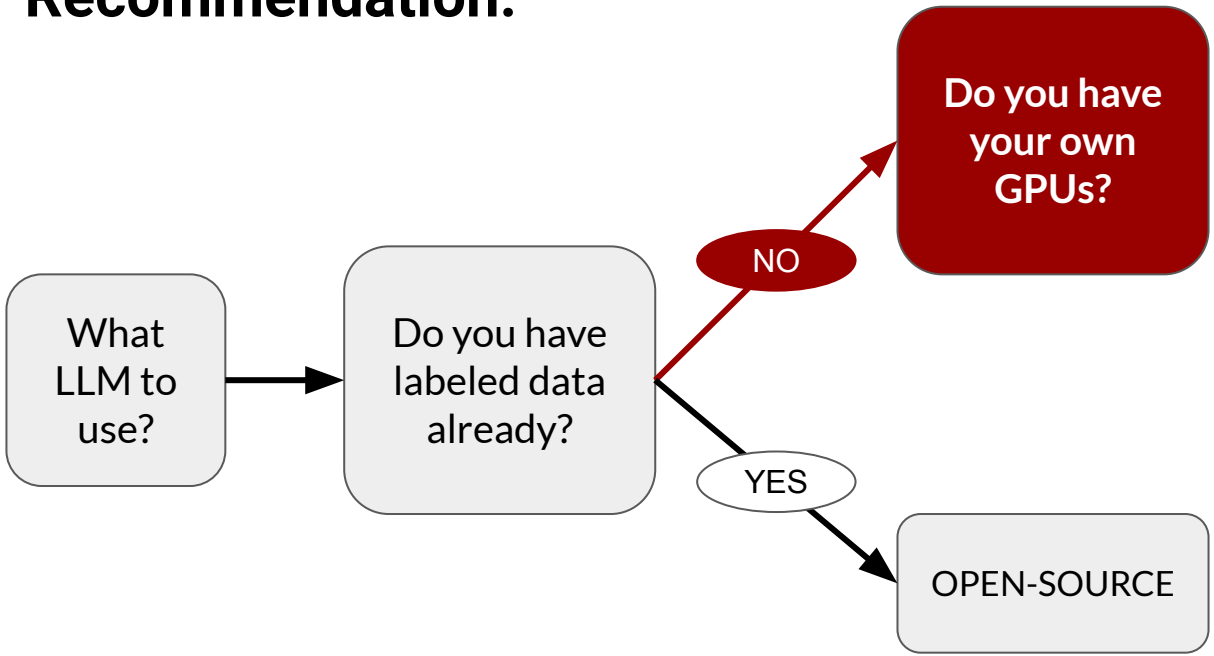


RQ2: Scaling Laws – Benefits of Open Source

Recommendation:

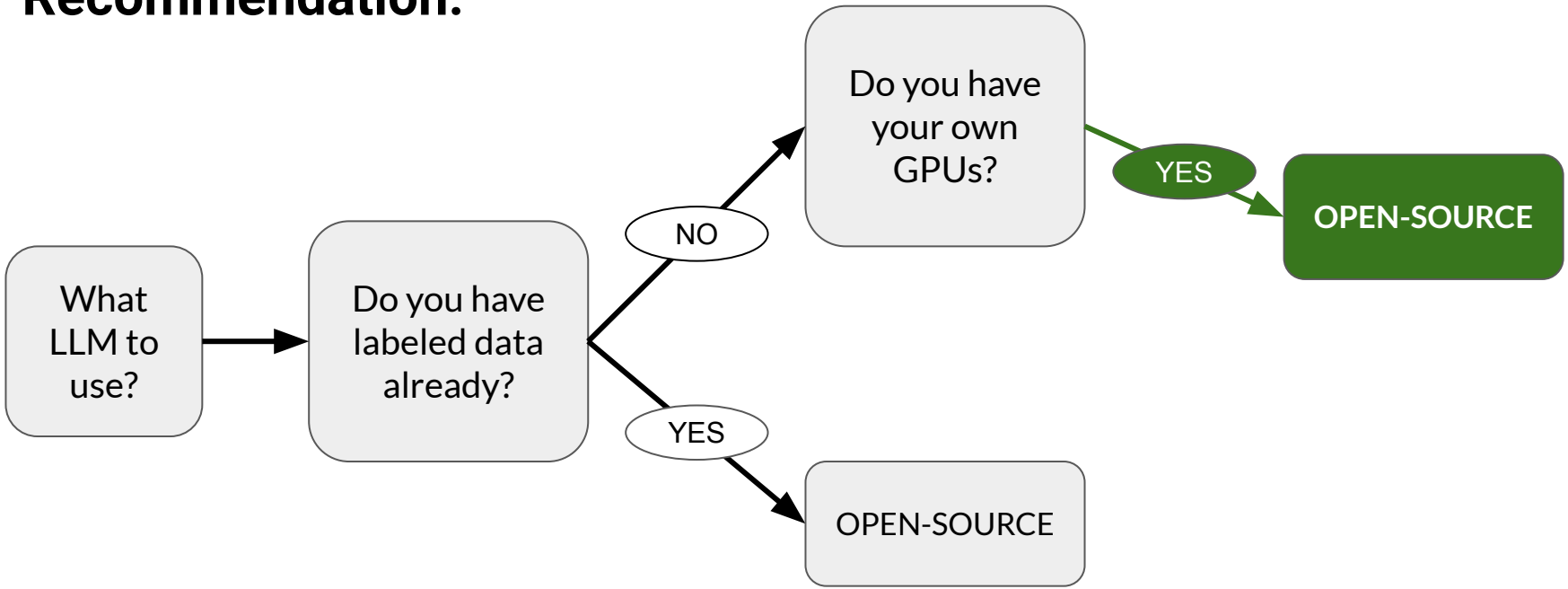
RQ2: Scaling Laws – Benefits of Open Source

Recommendation:



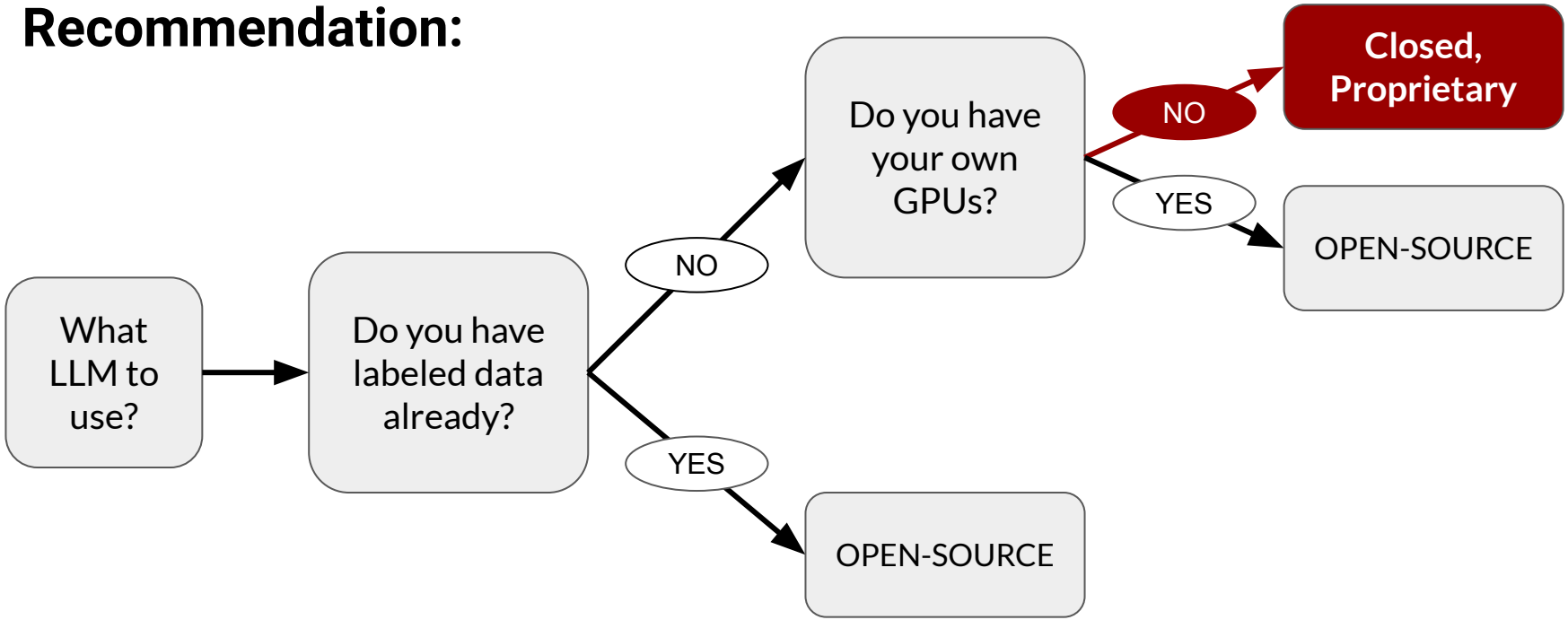
RQ2: Scaling Laws – Benefits of Open Source

Recommendation:



RQ2: Scaling Laws – Benefits of Open Source

Recommendation:



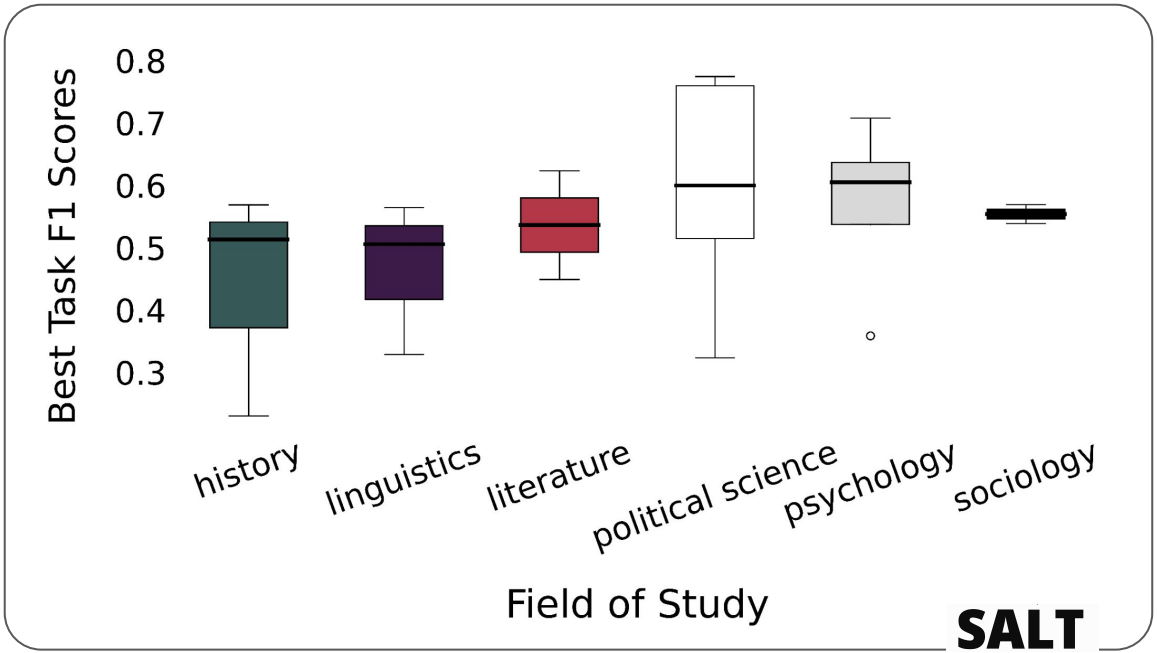
RQ3: Performance Depends on Task-Complexity

RQ3: Domain-Utility. Are LLMs better adapted for some subfields than others?

RQ3: Performance Depends on Task-Complexity

RQ3: Domain-Utility. Are LLMs better adapted for some subfields than others?

↪ **Findings:** Performance is ***not tied to academic discipline***

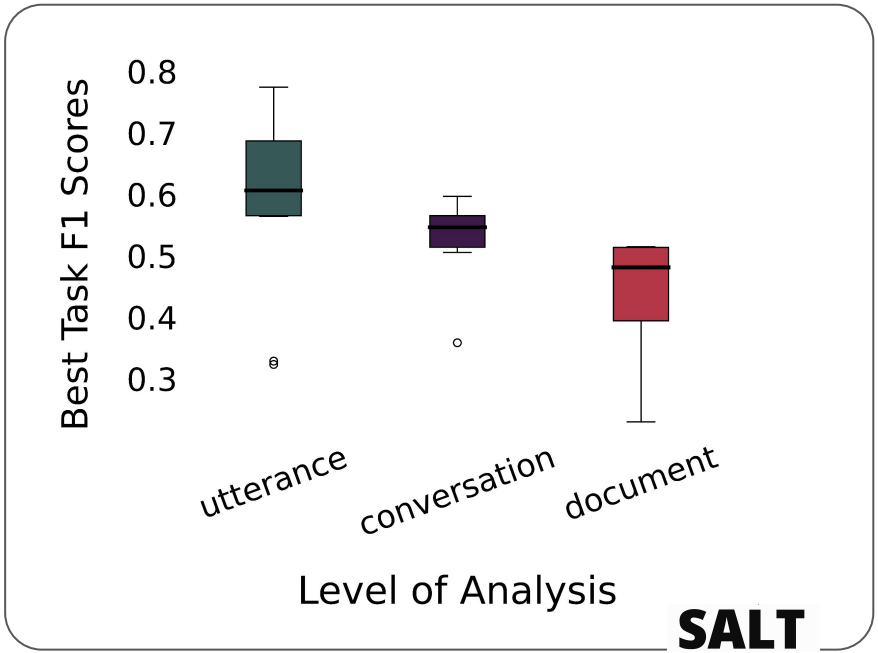


RQ3: Performance Depends on Task-Complexity

RQ3: Domain-Utility. Are LLMs better adapted for some subfields than others?

↪ **Findings:** Performance is ***not tied to academic discipline***

but rather by the ***complexity*** of the ***input***



RQ3: Performance Depends on Task-Complexity

Recommendations:

- Validate on a small sample
- Weigh benefits with risks
- Move beyond Western studies

RQ4: High-Quality Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

*emotion-specific **summarization***

CovidET (Zhan et al., 2022)

figurative language explanation

FLUTE (Chakrabarty et al., 2022)

implied misinformation explanation

Misinfo Reaction Frames (Gabriel et al., 2017)

hate speech explanation

Social Bias Inference Corpus (Sap et al. 2020)

positive reframing

Positive Psychology Frames (Ziems et al. 2022)

RQ4: High-Quality Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

emotion-specific summarization

CovidET (Zhan et al., 2022)

figurative language ***explanation***

FLUTE (Chakrabarty et al., 2022)

implied misinformation ***explanation***

Misinfo Reaction Frames (Gabriel et al., 2017)

hate speech ***explanation***

Social Bias Inference Corpus (Sap et al. 2020)

positive reframing

Positive Psychology Frames (Ziems et al. 2022)

RQ4: High-Quality Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

emotion-specific summarization

CovidET (Zhan et al., 2022)

figurative language explanation

FLUTE (Chakrabarty et al., 2022)

implied misinformation explanation

Misinfo Reaction Frames (Gabriel et al., 2017)

hate speech explanation

Social Bias Inference Corpus (Sap et al. 2020)

positive reframing

Positive Psychology Frames (Ziems et al. 2022)

RQ4: High-Quality Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

***emotion*-specific summarization**

CovidET (Zhan et al., 2022)

figurative language explanation

FLUTE (Chakrabarty et al., 2022)

implied misinformation explanation

Misinfo Reaction Frames (Gabriel et al., 2017)

hate speech explanation

Social Bias Inference Corpus (Sap et al. 2020)

positive reframing

Positive Psychology Frames (Ziems et al. 2022)

RQ4: High-Quality Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

emotion-specific summarization

CovidET (Zhan et al., 2022)

***figurative language** explanation*

FLUTE (Chakrabarty et al., 2022)

implied misinformation explanation

Misinfo Reaction Frames (Gabriel et al., 2017)

hate speech explanation

Social Bias Inference Corpus (Sap et al. 2020)

positive reframing

Positive Psychology Frames (Ziems et al. 2022)

RQ4: High-Quality Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

emotion-specific summarization

CovidET (Zhan et al., 2022)

figurative language explanation

FLUTE (Chakrabarty et al., 2022)

*implied **misinformation** explanation*

Misinfo Reaction Frames (Gabriel et al., 2017)

hate speech explanation

Social Bias Inference Corpus (Sap et al. 2020)

positive reframing

Positive Psychology Frames (Ziems et al. 2022)

RQ4: **High-Quality** Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

↪ **Findings:** zero-shot GPT-4 produces ***helpful and informative generations*** in all five evaluation tasks

RQ4: **High-Quality** Generation Results

RQ4: Functionality. Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?

↪ **Findings:** zero-shot GPT-4 produces ***helpful and informative generations*** in all five evaluation tasks

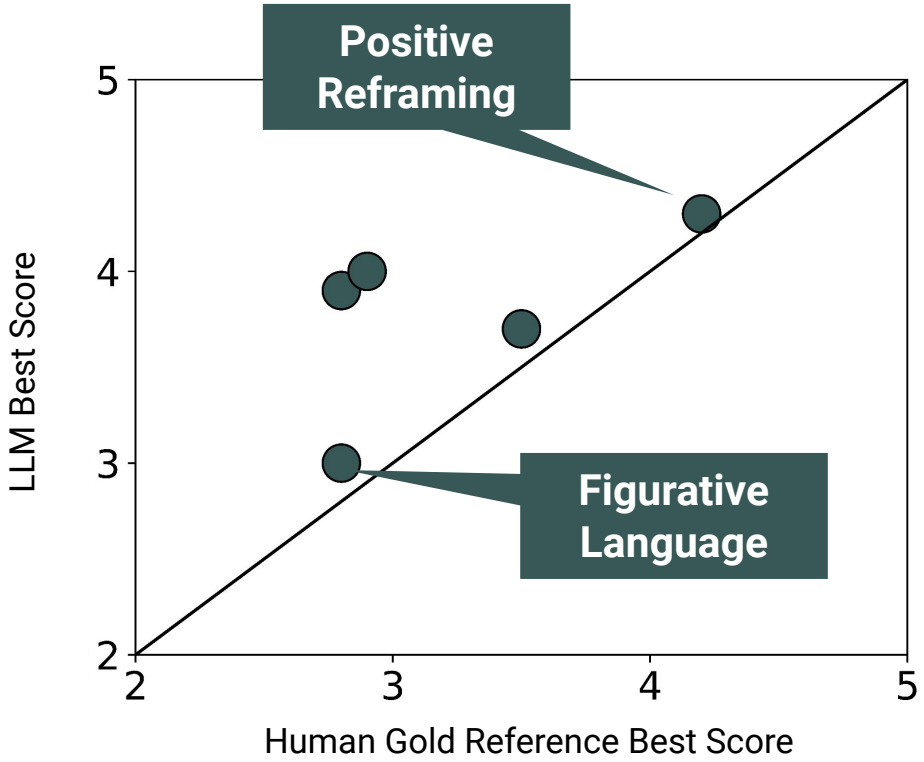
Task	COVID Aspect Summarization	Misinformation Explanation	Figurative Language	Hate Speech Explanation	Positive Reframing
Expert	**CDC Comm. Specialist**	**Public Policy Grad Student**	**Grammarly Writing Expert**	**Journalism Degree**	**Psychology Degree**

RQ4: High-Quality Generation Results

RQ4: Functionality.

↪ **Findings:** zero-shot GPT-4 ***beats reference*** levels of:

Faithfulness



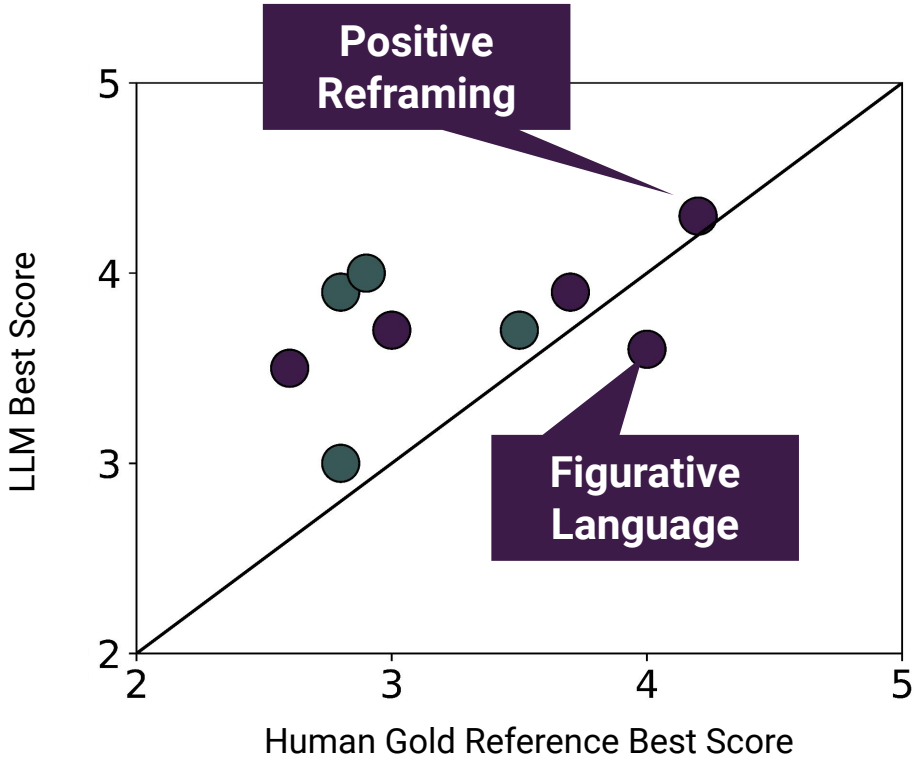
Task	COVID Aspect Summarization	Misinformation Explanation	Figurative Language	Hate Speech Explanation	Positive Reframing
Expert	**CDC Comm. Specialist**	**Public Policy Grad Student**	**Grammarly Writing Expert**	**Journalism Degree**	**Psychology Degree**

RQ4: High-Quality Generation Results

RQ4: Functionality.

↪ **Findings:** zero-shot GPT-4 ***beats reference*** levels of:

- ***Faithfulness***
- ***Relevance***



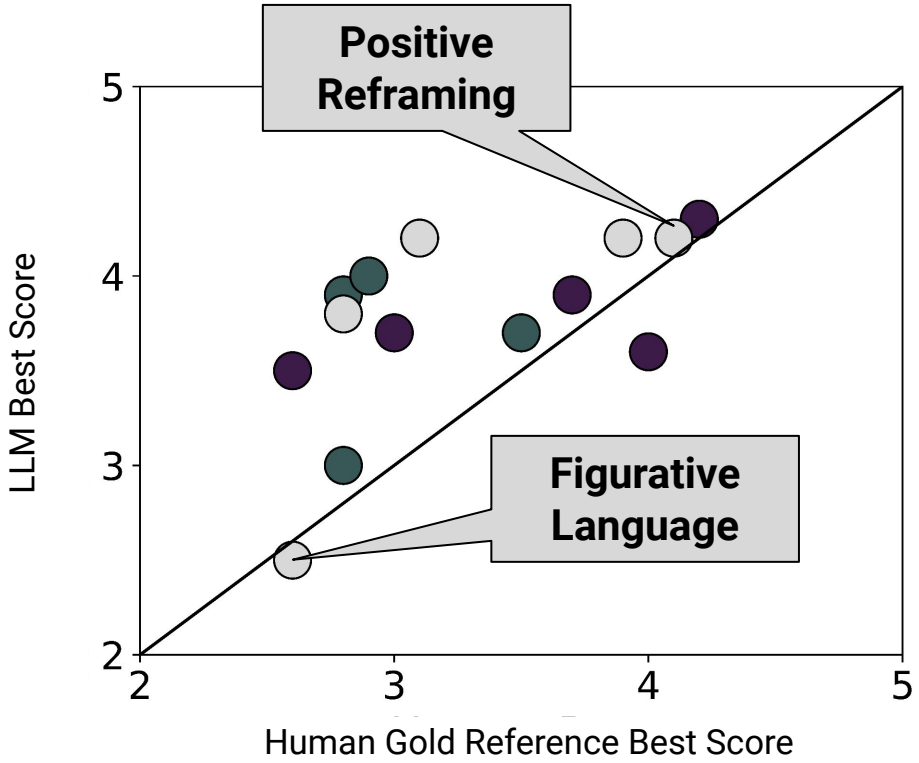
Task	COVID Aspect Summarization	Misinformation Explanation	Figurative Language	Hate Speech Explanation	Positive Reframing
Expert	**CDC Comm. Specialist**	**Public Policy Grad Student**	**Grammarly Writing Expert**	**Journalism Degree**	**Psychology Degree**

RQ4: High-Quality Generation Results

RQ4: Functionality.

↪ **Findings:** zero-shot GPT-4 ***beats reference*** levels of:

- ***Faithfulness***
- ***Relevance***
- ***Coherence***



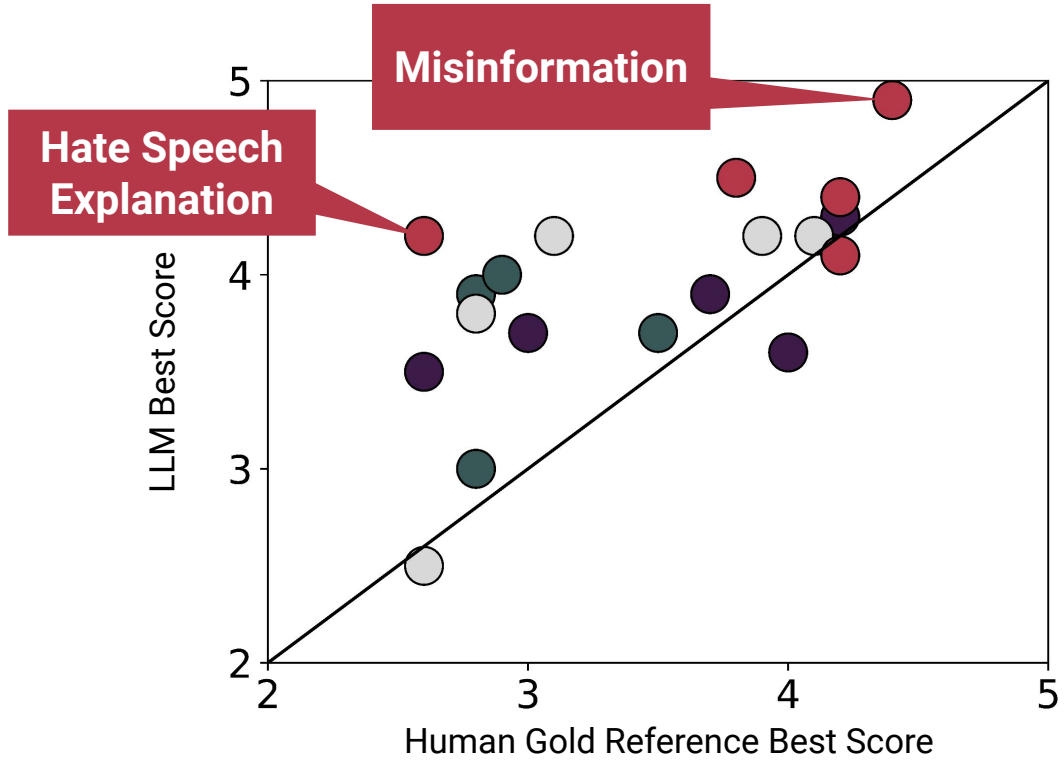
Task	COVID Aspect Summarization	Misinformation Explanation	Figurative Language	Hate Speech Explanation	Positive Reframing
Expert	**CDC Comm. Specialist**	**Public Policy Grad Student**	**Grammarly Writing Expert**	**Journalism Degree**	**Psychology Degree**

RQ4: High-Quality Generation Results

RQ4: Functionality.

↪ **Findings:** zero-shot GPT-4 ***beats reference*** levels of:

- ***Faithfulness***
- ***Relevance***
- ***Coherence***
- ***Fluency***



Task	COVID Aspect Summarization	Misinformation Explanation	Figurative Language	Hate Speech Explanation	Positive Reframing
Expert	**CDC Comm. Specialist**	**Public Policy Grad Student**	**Grammarly Writing Expert**	**Journalism Degree**	**Psychology Degree**

Discussion

CSS Challenges for LLMs:

1. Subtle expert taxonomies
2. Size of the target label space
3. Structural parsing
4. Temporal grounding

Discussion

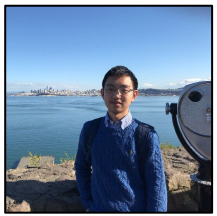
Recommendations:

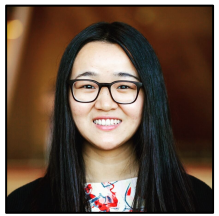
1. Integrate LLMs in the loop to transform large-scale data labeling
2. Consider open-source LLMs for classification
3. Reinvest in expert annotation

Can Large Language Models *Transform* Computational Social Science?

Caleb Ziems[†*], William Held[✦*], Omar Shaikh[†*], Jiaao Chen[✦*], Zhehao Zhang[‡*], Diyi Yang[†]





* All heavily contributed to the implementation of this work

Stanford NLP

[†]Stanford

[✦]Georgia Tech

[‡]Dartmouth

Social and Language Technologies
SALT

RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation
B: Trustworthy

Constraint: Answer with only the option above that is most accurate and nothing else.

RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation
B: Trustworthy

Constraint: Answer with only the option above that is most accurate and nothing else.

RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation ↩
B: Trustworthy ↩

Constraint: Answer with only the option above that is most accurate and nothing else.

RQ1: Zero-Shot Classification Performance

RQ1: Viability. Can LLMs augment the human annotation pipeline?

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation
B: Trustworthy

Constraint: Answer with only the option above that is most accurate and nothing else.