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  ✦ Georgia Tech  ‡ Dartmouth
Research Questions

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

- Psychology
- Political Science
- Literature
- History
- Sociology
- Linguistics
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
RQ: Are LLMs useful tools in the Computational Social Scientist’s toolkit?

- politeness recognition
- humor recognition
- emotion recognition
- empathy classification
- stance detection
- ideology detection
- agent framing
- relationship dynamics
- event extraction
- power relations identification
- social role detection
- dialect feature identification

**Psychology**

**Political Science**

**Literature**

**History**

**Sociology**

**Linguistics**
Research Questions

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

<table>
<thead>
<tr>
<th>Psychology</th>
<th>Political Science</th>
<th>Literature</th>
<th>History</th>
<th>Sociology</th>
<th>Linguistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politeness recognition</td>
<td>Stance detection</td>
<td>Agent framing</td>
<td>Event extraction</td>
<td>Power relations identification</td>
<td>Dialect feature identification</td>
</tr>
<tr>
<td>Humor recognition</td>
<td>Ideology detection</td>
<td>Relationship dynamics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotion recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathy classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Stanford Politeness Corpus** (Danescu-Niculescu-Mizil et al., 2013)

**r/Jokes + Pun of the Day** (Weller and Seppi 2019)

**CARER** (Saravia et al. 2018)

**EPITOME** (Sharma et al., 2020)

**SemEval-2016 Stance Dataset** (Mohammad et al., 2016)

**Ideological Books Corpus** (Gross et al., 2013)

**Article Bias Corpus** (Baly et al., 2013)

**WikiEvents** (Li et al., 2021)

**Hippocorpus** (Sap et al., 2020)

**Wikipedia Talk Pages** (Danescu-Niculescu-Mizil et al. 2012)

**CMU Movie Corpus** (Bamman et al. 2013)

**Indian English Minimal Pairs** (Demszky et al. 2019)

**Psychology**

**Political Science**

**Literature**

**History**

**Sociology**

**Linguistics**

**Psychology**

**Political Science**

**Literature**

**History**

**Sociology**

**Linguistics**

**Psychology**

**Political Science**

**Literature**

**History**

**Sociology**

**Linguistics**
RQ1: Zero-Shot Classification Performance

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

- **Flan-T5** (instruction-tuned)
  - small: 80M
  - base: 250M
  - large: 780M
  - XL: 3B
  - XXL: 11B
  - UL2: 20B
RQ2: Model-Selection.

GPT-3
- text-ada 350M
- text-babbage 1.3B
- text-curie 6.7B
- text-davinci 175B
  - 001
  - 002 (more / better data)
  - 003 (+ RLHF)

3.5 turbo ChatGPT
(+ dialog tuning / RLHF)

GPT-4
- 1.8T

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

<> **Findings:** Performance scales with model size
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:**

What LLM to use? → Do you have labeled data already?
RQ2: Scaling Laws – Benefits of Open Source

Recommendation:

- What LLM to use?
- Do you have labeled data already?
- YES
- OPEN-SOURCE
Recommendation:

RQ2: Scaling Laws – Benefits of Open Source

What LLM to use?

Do you have labeled data already?

Do you have your own GPUs?

NO

YES

OPEN-SOURCE
RQ2: Scaling Laws – Benefits of Open Source

Recommendation:

What LLM to use?

Do you have labeled data already?

Do you have your own GPUs?

YES

OPEN-SOURCE

YES

OPEN-SOURCE

NO
RQ2: Scaling Laws – **Benefits of Open Source**

**Recommendation:**

- **What LLM to use?**
  - Do you have labeled data already?
    - NO
      - Do you have your own GPUs?
        - NO
          - Closed, Proprietary
        - YES
          - OPEN-SOURCE
    - YES
      - OPEN-SOURCE
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.
Findings: Performance is not tied to academic discipline but rather by the complexity of the input.

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

RQ3: Performance Depends on Task-Complexity
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: **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

<table>
<thead>
<tr>
<th>Task</th>
<th>COVID Aspect Summarization</th>
<th>Misinformation Explanation</th>
<th>Figurative Language</th>
<th>Hate Speech Explanation</th>
<th>Positive Reframing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>CDC Comm. Specialist</td>
<td>Public Policy Grad Student</td>
<td>Grammarly Writing Expert</td>
<td>Journalism Degree</td>
<td>Psychology Degree</td>
</tr>
</tbody>
</table>
RQ4: **High-Quality Generation Results**

RQ4: **Functionality.**

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

- Faithfulness

<table>
<thead>
<tr>
<th>Task</th>
<th>COVID Aspect Summarization</th>
<th>Misinformation Explanation</th>
<th>Figurative Language</th>
<th>Hate Speech Explanation</th>
<th>Positive Reframing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>CDC Comm. Specialist</td>
<td>Public Policy Grad Student</td>
<td>Grammarly Writing Expert</td>
<td>Journalism Degree</td>
<td>Psychology Degree</td>
</tr>
</tbody>
</table>
RQ4: **High-Quality Generation Results**

RQ4: **Functionality.**

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

- Faithfulness
- Relevance

<table>
<thead>
<tr>
<th>Task</th>
<th>COVID Aspect Summarization</th>
<th>Misinformation Explanation</th>
<th>Figurative Language</th>
<th>Hate Speech Explanation</th>
<th>Positive Reframing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>CDC Comm. Specialist</td>
<td>Public Policy Grad Student</td>
<td>Grammarly Writing Expert</td>
<td>Journalism Degree</td>
<td>Psychology Degree</td>
</tr>
</tbody>
</table>
RQ4: High-Quality Generation Results

RQ4: Functionality.

→ Findings: zero-shot GPT-4 beats reference levels of:

- Faithfulness
- Relevance
- Coherence

Diagram:

- Positive Reframing
- Figurative Language

Table:

<table>
<thead>
<tr>
<th>Task</th>
<th>COVID Aspect Summarization</th>
<th>Misinformation Explanation</th>
<th>Figurative Language</th>
<th>Hate Speech Explanation</th>
<th>Positive Reframing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>CDC Comm. Specialist</td>
<td>Public Policy Grad Student</td>
<td>Grammarly Writing Expert</td>
<td>Journalism Degree</td>
<td>Psychology Degree</td>
</tr>
</tbody>
</table>
RQ4: **High-Quality Generation Results**

**RQ4: Functionality.**

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

- Faithfulness
- Relevance
- Coherence
- Fluency

<table>
<thead>
<tr>
<th>Task</th>
<th>COVID Aspect Summarization</th>
<th>Misinformation Explanation</th>
<th>Figurative Language</th>
<th>Hate Speech Explanation</th>
<th>Positive Reframing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>CDC Comm. Specialist</td>
<td>Public Policy Grad Student</td>
<td>Grammarly Writing Expert</td>
<td>Journalism Degree</td>
<td>Psychology Degree</td>
</tr>
</tbody>
</table>
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
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: **Viability.** Can LLMs augment the human annotation pipeline?

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

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

**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.