Panel Session 1 Comments

EPSA 2025 LLM Workshop

Frederik Hjorth
Associate Professor, Ph.D.
fghjorth.github.io

UNIVERSITY OF
COPENHAGEN

1 Muti et al.

2 Bailey et al.

3 Ceron

4 Bormann & Viganò

5 Elshehawy et al.

General comments

- Very compelling motivating case
- Great use case for LLMs
- Very well-chosen inclusion of qualitative data

General comments

- Very compelling motivating case
- Great use case for LLMs
- Very well-chosen inclusion of qualitative data

General comments

- Very compelling motivating case
- Great use case for LLMs
- Very well-chosen inclusion of qualitative data

General comments

- Very compelling motivating case
- Great use case for LLMs
- Very well-chosen inclusion of qualitative data

Potential improvements

- Motivating case (i.e., Giulia Cecchettin) could be explained better
- Link more explicitly to literature on **stance detection** (Bestvater and Monroe cited, but not applied)
- Carefully designed codebook → why not use it to guide the LLM?
- Use of Perspective's toxicity classifier is too credulous, see e.g. Pozzobon et al. 2023 (`https://aclanthology.org/2023.emnlp-main.472.pdf`)
- Ultimately LLM performance is not impressive ⇝ either refine prompt or use fine-tuned model

Potential improvements

- Motivating case (i.e., Giulia Cecchettin) could be explained better
- Link more explicitly to literature on **stance detection** (Bestvater and Monroe cited, but not applied)
- Carefully designed codebook → why not use it to guide the LLM?
- Use of Perspective's toxicity classifier is too credulous, see e.g. Pozzobon et al. 2023 (`https://aclanthology.org/2023.emnlp-main.472.pdf`)
- Ultimately LLM performance is not impressive ⇝ either refine prompt or use fine-tuned model

Potential improvements

- Motivating case (i.e., Giulia Cecchettin) could be explained better
- Link more explicitly to literature on **stance detection** (Bestvater and Monroe cited, but not applied)
- Carefully designed codebook → why not use it to guide the LLM?
- Use of Perspective's toxicity classifier is too credulous, see e.g. Pozzobon et al. 2023 (https://aclanthology.org/2023.emnlp-main.472.pdf)
- Ultimately LLM performance is not impressive ⇝ either refine prompt or use fine-tuned model

Potential improvements

- Motivating case (i.e., Giulia Cecchettin) could be explained better
- Link more explicitly to literature on **stance detection** (Bestvater and Monroe cited, but not applied)
- Carefully designed codebook → why not use it to guide the LLM?
- Use of Perspective's toxicity classifier is too credulous, see e.g. Pozzobon et al. 2023 (https://aclanthology.org/2023.emnlp-main.472.pdf)
- Ultimately LLM performance is not impressive ⇝ either refine prompt or use fine-tuned model

Potential improvements

- Motivating case (i.e., Giulia Cecchettin) could be explained better
- Link more explicitly to literature on **stance detection** (Bestvater and Monroe cited, but not applied)
- Carefully designed codebook → why not use it to guide the LLM?
- Use of Perspective's toxicity classifier is too credulous, see e.g. Pozzobon et al. 2023 (`https://aclanthology.org/2023.emnlp-main.472.pdf`)
- Ultimately LLM performance is not impressive ⇝ either refine prompt or use fine-tuned model

Potential improvements

- Motivating case (i.e., Giulia Cecchettin) could be explained better
- Link more explicitly to literature on **stance detection** (Bestvater and Monroe cited, but not applied)
- Carefully designed codebook → why not use it to guide the LLM?
- Use of Perspective's toxicity classifier is too credulous, see e.g. Pozzobon et al. 2023 (`https://aclanthology.org/2023.emnlp-main.472.pdf`)
- Ultimately LLM performance is not impressive ⇝ either refine prompt or use fine-tuned model

1 Muti et al.

2 Bailey et al.

3 Ceron

4 Bormann & Viganò

5 Elshehawy et al.

General comments

- Introduction is highly compelling, expertly structured
- Key concepts clearly defined
- Very careful measurement strategy

General comments

- Introduction is highly compelling, expertly structured
- Key concepts clearly defined
- Very careful measurement strategy

General comments

- Introduction is highly compelling, expertly structured
- Key concepts clearly defined
- Very careful measurement strategy

General comments

- Introduction is highly compelling, expertly structured
- Key concepts clearly defined
- Very careful measurement strategy

Potential improvements

- ‘Empirical Context’ section relatively sparse → include e.g. timeline, map
- ‘Coding Procedure’ is heavy on preprocessing, but light on actual coding
- ‘Mechanisms’ section a bit unclear—are LLMs also used for classification here?
- Be upfront about limits to staggered treatment adoption design—e.g., could the same event trigger mobilization and GBV?

Potential improvements

- ‘Empirical Context’ section relatively sparse → include e.g. timeline, map
- ‘Coding Procedure’ is heavy on preprocessing, but light on actual coding
- ‘Mechanisms’ section a bit unclear—are LLMs also used for classification here?
- Be upfront about limits to staggered treatment adoption design—e.g., could the same event trigger mobilization and GBV?

Potential improvements

- 'Empirical Context' section relatively sparse → include e.g. timeline, map
- 'Coding Procedure' is heavy on preprocessing, but light on actual coding
- 'Mechanisms' section a bit unclear—are LLMs also used for classification here?
- Be upfront about limits to staggered treatment adoption design—e.g., could the same event trigger mobilization and GBV?

Potential improvements

- ‘Empirical Context’ section relatively sparse → include e.g. timeline, map
- ‘Coding Procedure’ is heavy on preprocessing, but light on actual coding
- ‘Mechanisms’ section a bit unclear—are LLMs also used for classification here?
- Be upfront about limits to staggered treatment adoption design—e.g., could the same event trigger mobilization and GBV?

Potential improvements

- ‘Empirical Context’ section relatively sparse → include e.g. timeline, map
- ‘Coding Procedure’ is heavy on preprocessing, but light on actual coding
- ‘Mechanisms’ section a bit unclear—are LLMs also used for classification here?
- Be upfront about limits to staggered treatment adoption design—e.g., could the same event trigger mobilization and GBV?

1 Muti et al.

2 Bailey et al.

3 Ceron

4 Bormann & Viganò

5 Elshehawy et al.

Ceron

[Based on presentation]

1 Muti et al.

2 Bailey et al.

3 Ceron

4 Bormann & Viganò

5 Elshehawy et al.

General comments

- LLMs for event data is a very compelling use case
- Huge potential to augment current efforts such as UCDP GED or GDELT
- All methodological choices very lucidly explained

General comments

- LLMs for event data is a very compelling use case
- Huge potential to augment current efforts such as UCDP GED or GDELT
- All methodological choices very lucidly explained

General comments

- LLMs for event data is a very compelling use case
- Huge potential to augment current efforts such as UCDP GED or GDELT
- All methodological choices very lucidly explained

General comments

- LLMs for event data is a very compelling use case
- Huge potential to augment current efforts such as UCDP GED or GDELT
- All methodological choices very lucidly explained

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 'processed'
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 'processed'
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 'processed'
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 'processed'
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 ‘processed’
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 'processed'
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

Potential improvements

- Feels double-barreled: (i) introduction of CAIN and (ii) evaluation of LLM performance
- Which parts of (i) are really necessary for (ii)? Make those links explicit
- Sampling process: 20k news articles → 2k annotated → 348 'processed'
- ⇝ say more about the sampling process here
- OCR issues seem serious! Could a separate LLM OCR correction step be useful?
- Results look very promising—but would be useful to see consequences for a downstream estimate

1 Muti et al.

2 Bailey et al.

3 Ceron

4 Bormann & Viganò

5 Elshehawy et al.

General comments

- Clear contribution: looking beyond media bias
- Super impressive data collection (Appendix A1 a paper unto itself!)
- Very elegant visualization throughout

General comments

- Clear contribution: looking beyond media bias
- Super impressive data collection (Appendix A1 a paper unto itself!)
- Very elegant visualization throughout

General comments

- Clear contribution: looking beyond media bias
- Super impressive data collection (Appendix A1 a paper unto itself!)
- Very elegant visualization throughout

General comments

- Clear contribution: looking beyond media bias
- Super impressive data collection (Appendix A1 a paper unto itself!)
- Very elegant visualization throughout

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- ⇝ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- ⇝ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- ⇝ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- ⇝ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- ⇝ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- ⇝ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

Potential improvements

- Be more explicit about the implied mechanism: why are police stations (supposedly) doing this?
- $\rightsquigarrow$ strong arguments that police have (some) discretion, but little about *incentives*
- Why would these incentives be time-varying? Very important for the appropriateness of RDiT!
- What about police discretion wrt timing? Could officers frontload outgroup reporting before an election (like the Comey letter)?
- Take day-of-week effects more seriously (e.g., using FEs)—placebo test not fully persuasive
- Result appears limited to quite narrow bandwidths—strong justification for those would be needed

🙏