Lecture 4: Large language models (LLMs)/decoder-based models in Social Science

Miriam Hurtado Bodell

LINKÖPING UNIVERSITY

Recap decoders/LLMs

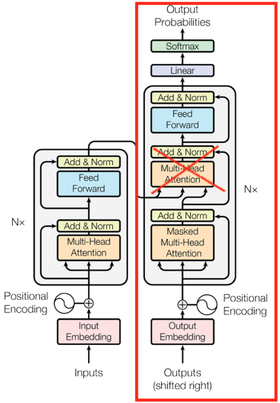

Figure 1: The Transformer - model architecture.

LLMs/Decoder-based models in the social sciences

This is the latest research frontier = still many unknowns. How *can* and *should* LLMs be used for social science?

But the consensus seems to be that there are both opportunities and potential "dangers" with using LLMs for social scientific research

Opportunities: text annotation, Törnberg. (2024)

No consensus on best practices for how to use LLMs for text annotation:

1. Choose an appropriate model
2. Follow a systematic coding procedure
3. Develop a prompt codebook
4. Validate your model
5. Engineer your prompts
6. Specify your LLM parameters
7. Discuss ethical and legal implications
8. Examine model stochasticity
9. Consider that your data may be in the training data

Opportunities: **measurement**/text annotation, Törnberg. (2024), choose an appropriate model

1. Choose an appropriate model
   - Reproducibility: use a fixed version of the LLM and ensure that the model will be available in future
   - Ethics and legality: e.g., not storing research data compliance with relevant data privacy regulations
   - Transparency: should exists clear documentations of; methods, data sources, and assumptions etc.
   - Culture and language: The LLM should adequately support the language(s) and cultures of your textual data (many English-centric)
   - Scalability: need to handle scale of you data
   - Complexity: need to handle the complexity of the task

Opportunities: **measurement**/text annotation, Törnberg. (2024), choose an appropriate model

2. Follow a systematic coding procedure
   - Iterative process: annotate small samle of data & compare with model output, where do the model get it wrong? Ask LLM to give reason for label. Revise prompt or codebook.
3. Develop a prompt codebook
4. Validate your model
   - Validate after your developed your *final* prompt
   - Go beyond accuracy: e.g., F1-score, Cohen's κ, MSE, errors in different subset

Opportunities: **measurement**/text annotation, Törnberg. (2024)

No consensus on best practices for how to use LLMs for text annotation:

5. Engineer your prompts
   - Contain the following elements: context ("who answers"), question (what), and constraints (format).
   - Give "I don't know" option
   - Chain of though/"Let's think step by step" can help
6. Specify your LLM parameters: max length (of text), temperature (high $\rightarrow$ more randomness, for classification set low/0), top-p (range of considered token, choose low), etc.

Opportunities: **measurement**/text annotation, Törnberg. (2024), example prompt

As an expert annotator with a focus on social media content analysis, your role involves scrutinizing Twitter messages related to the US 2020 election. Your expertise is crucial in identifying misinformation that can sway public opinion or distort public discourse.

Does the message contain misinformation regarding the US 2020 election?

Provide your response in JSON format, as follows:
{ "contains_misinformation": "Yes/No/Uncertain", "justification": "Provide a brief justification for your choice." }

Options:
- Yes
- No
- Uncertain

Remember to prioritize accuracy and clarity in your analysis, using the provided context and your expertise to guide your evaluation. If you are uncertain about the classification, choose 'Uncertain' and provide a rationale for this uncertainty.

Twitter message: [MESSAGE]

Answer:

Opportunities: **measurement**/text annotation, Törnberg. (2024)

7. Discuss ethical and legal implications
   - Input data can be used for future training (don't feed sensitive or copyright-protected texts)
8. Examine model stochasticity
   - Does the same prompt return the same result if run several times? Do small variations in the prompt result in different results?
9. Consider that your data may be in the training data
   - Avoid using publicly available databases (e.g. Manifesto Data) as validation data, they can be in training data

Opportunities: **silicon samples**, Argyle et al. (2023)

Main idea: Use LLMs to generate survey answers for "respondents" with different social-demographics to correct biases in GPT, works if high algorithmic fidelity (AF)

AF = the degree to which patterns of relationships between ideas, attitudes, and sociocultural contexts accurately mirror those within human subpopulations

Opportunities: **silicon samples**, Argyle et al. (2023)

Task 1: Use real survey respondents to create a "backstory," then ask to generate words/text describing partisans. Compare with humans in the survey.

	Describing Democrats	Describing Republicans
Strong Republicans	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. Financially, I am upper-class. In terms of my age, I am young. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist**.	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Conservative** 2. **Male** 3. **White (or Caucasian)** 4. **Christian**.
Strong Democrats	Ideologically, I describe myself as liberal. Politically, I am a strong Democrat. Racially, I am white. I am female. Financially, I am poor. In terms of my age, I am old. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal**. 2. **Young**. 3. **Female**. 4. **Poor**.	Ideologically, I describe myself as extremely liberal. Politically, I am a strong Democrat. Racially, I am hispanic. I am male. Financially, I am upper-class. In terms of my age, I am middle-aged. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Ignorant** 2. **Racist** 3. **Misogynist** 4. **Homophobic**.

Finding: GPT3 and humans are (almost) indistinguishable

Opportunities: **silicon samples**, Argyle et al. (2023)

Task 2: predict voting in 2012, 2016, 2020 given the socio-demographics of survey respondents (training data until 2019)

Variable	2012 Tetra.	2012 Prop. Agree	2016 Tetra.	2016 Prop. Agree	2020 Tetra.	2020 Prop. Agree
Whole sample	0.90	0.85	0.92	0.87	0.94	0.89
Men	0.90	0.85	0.93	0.88	0.95	0.88
Women	0.91	0.86	0.92	0.86	0.94	0.90
Strong partisans	0.99	0.97	1.00	0.97	1.00	0.97
Weak partisans	0.73	0.74	0.71	0.74	0.84	0.82
Leaners	0.90	0.85	0.93	0.87	0.95	0.89
Independents	0.31	0.59	0.41	0.62	0.02	0.53
Conservatives	0.84	0.84	0.88	0.86	0.91	0.89
Moderates	0.65	0.77	0.76	0.78	0.71	0.77
Liberals	0.81	0.95	0.73	0.95	0.86	0.97
Whites	0.87	0.82	0.91	0.85	0.94	0.89
Blacks	0.71	0.97	0.87	0.96	0.81	0.94
Hispanics	0.86	0.86	0.93	0.90	0.88	0.83
Attends church	0.91	0.86	0.93	0.88	0.94	0.88
Does not attend church	0.88	0.85	0.90	0.85	0.93	0.90
High interest in politics	0.95	0.90	0.97	0.93	0.97	0.92
Low interest in politics	0.71	0.74	0.75	0.75	0.83	0.81
Discusses politics	0.92	0.87	0.94	0.88	0.95	0.90
Does not discuss politics	0.83	0.82	0.81	0.79	0.80	0.79
18–30 years old	0.90	0.87	0.90	0.86	0.90	0.87
31–45 years old	0.90	0.85	0.92	0.87	0.94	0.90
46–60 years old	0.90	0.86	0.92	0.86	0.92	0.87
Over 60	0.90	0.85	0.93	0.87	0.96	0.91

Finding: GPT-3 underestimate Romney 2012, Trump 2016 & Biden 2020

Opportunities: **silicon samples**, Argyle et al. (2023)

Task 3: predict answers to survey given other survey responses relating to socio-demographic variables, attitudes, and behaviors

Finding: GPT-3 captures human-like patterns between different survey items

Opportunities: **silicon samples**, Argyle et al. (2023)

Conclusion: AF of GPT3 is high enough to use for social science

So what?

- Help to design survey questions, experimental treatments, and codebooks to guide human research (low cost)
- Give insights into which variables researchers should include in their studies of public opinion if they want to accurately understand (Americans') voting behavior
- Generate synthetic data for analyses when there is no real data

But other studies finds issues with e.g., positive bias, worse for some subgroups, etc. (to be continued!)

Opportunities: **inference**/simulation studies

Agent-based modeling a long standing simulation framework used to map how micro-level assumptions to macro-level outcomes, but agents' interactions are typically naive (interaction = tie or in the same place on map)

LLMs has been used to simulate conversations between agents with different "personalities" and help us study emergent behaviors of individuals and groups

Opportunities: inference/simulation studies, e.g. Park et al. (2023)

Give 25 agents:

- A **personality** (prompt with a description of a person; personality traits, jobs, relations), **memory** (store list of personality, relations, previous experience), **reflection** and **planning** ability
- Each iteration they output a text that describes what they do & interact with other agents in local area with a prob.

Emergent social behaviors

- Information diffusion (gossip)
- Relationship memory (ask questions about convos. in past)
- Coordination (one throws party, invite others, that show up)

Opportunities: inference/simulation studies, e.g. Törnberg et al. (2023)

Aim: Test how the level of toxicity in online conversations depend on platform news feed algorithm

- Generate 500 agents with personalities based on survey respondents in the US; the agents "read" newspaper articles and are asked to write a social media comment and/or like the article
- Compare 3 news feed algorithms: (1) interact with posts most liked by friends (30 homophilic ties) (2) interact with post most liked by all users, and (3) interact with posts most liked by people with opposing political view

Opportunities: inference/simulation studies, e.g. Törnberg et al. (2023)

	Toxicity	E-I interpartisan comments	E-I interpartisan likes
Platform 1	0.09	-0.89	-0.97
Platform 2	0.13	-0.70	-0.78
Platform 3	0.07	0.33	-0.18

Table 2. The resulting toxicity and interpartisan interaction.

<u>Results:</u> Platform algorithm impacts toxicity and cross-partisian interaction

Limitations and potential dangers:

Biases:

- More liberal than conservative, more young than old, more male than female, more extrovert than introvert etc.
- Higher recall than precision (more false positives than false negatives)
- Some LLMs have filters to avoid generating certain content (this changes over time) $\rightarrow$ not true representations of culture (?)

Errors:

- Hallucinations/make up content, e.g. literature reviews with non-existing papers
- Not better than encoder-only models in many classification tasks

Limitations and potential dangers:

Ethnics:

- Informed consent when exposing real humans to LLMs or LLM-created content
- Working conditions of people working with training LLMs (sweatshops)
- Environmental impact for training LLMs

Replicability:

- Many LLMs updated continuously and will produce new answers over time

Summary

Using encoder-style transformers in the social science is new → still difficult to know how and when then should be used

Using decoders for text annotation show great potential, but encoders are likely better for many social scientific concepts (Ollion et al. 2023)

Decoders may be used to perform better/less costly surveys

Can be used in simulation/experimental settings to simulate human values, conversations, and behaviors

But still associated with much uncertainty, and potential dangers in letting loose LLMs for policy relevant research

Thank you!

www.liu.se

LINKÖPING UNIVERSITY