Promises and Pitfalls of LLM for Measurement from Text

HSS 510: NLP for HSS

Taegyoon Kim

Jun 5, 2024

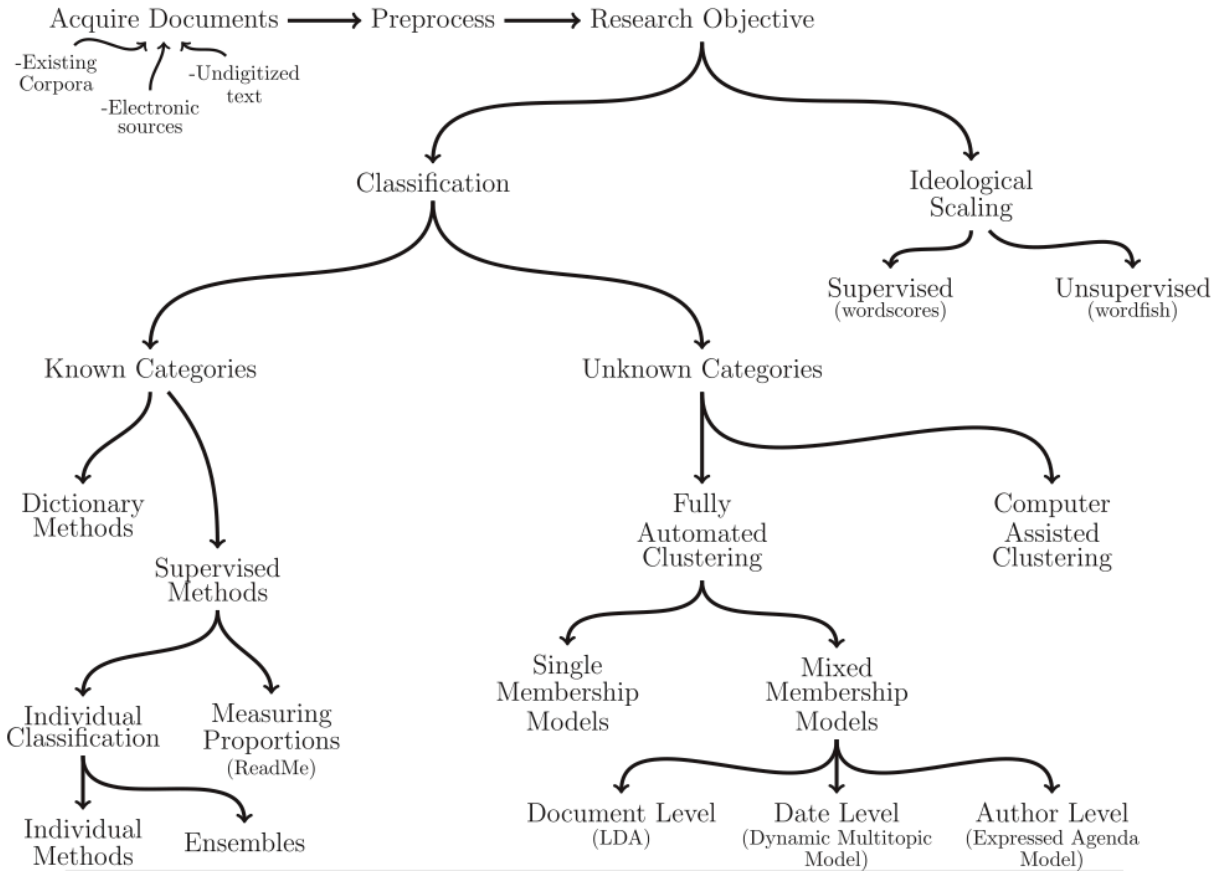
Agenda

Things to be covered

- LLM as a tool for measurement from text
  - Quick review of measurement methods with texts
  - Rise of LLM and prompt-based measurement
  - Best practices
  - Discussion of issues related to current practices
- Prompt-based relation inference
  - Presentation by Prof. Woo Seokkyun

Measurement from Texts in HSS

Many methods in text-as-data are for measurement



Measurement from Texts in HSS

Evolution of text classification/scaling

- Dictionary methods
- Supervised machine learning
- Fine-tuning pre-trained language models

Large Language Model as a Measurement Tool

Prompt-based classification/scaling of texts

- LLMs are queried with an instruction (with/without examples)
- They are asked to provide an answer, designed to capture the concept of interest
  - No example: zero-shot
  - One example: one-shot
  - Multiple examples: few-shot
- A stylized illustration of zero-shot prompting

```
"Classify the text as 'Relevant' or 'Not Relevant' to immigration"
"Classification:"
```

Large Language Model as a Measurement Tool

Prompt-based classification/scaling of texts

- Dispenses with the need of building a training set
- Sophisticated architecture + massive training data → various tasks/domains/language
- *Might* perform as well as cutting-edge fine-tuned models in terms of performance (rarely excel them at least for now)
- Much research suggests LLMs can be used for measurement
  - Chae and Davidson (2023); Gilardi et al. (2023); Ornstein et al. (2024); Rathje et al. (2024); Rytting et al. (2023); Tornberg (2023); Yang and Mencer (2023); Ziems et al. (2023)

Large Language Model as a Measurement Tool

A wide range of tasks

- Relevance
- Sentiment
- Stance
- Discrete emotions
- Moral expressions
- Topic

Large Language Model as a Measurement Tool

Performance on various tasks (Ziems et al. 2023)

Model / Data	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Utterance Level Tasks																
Dialect	3.3	3.0	0.2	4.5	23.4	24.8	30.3	32.9	0.5	0.5	1.2	9.1	17.1	14.7	11.7	23.2
Emotion	16.7	71.6	19.8	63.8	69.7	65.7	66.2	70.8	6.4	4.9	6.6	19.7	36.8	44.0	47.1	50.6
Figurative	25.0	99.2	16.6	23.2	18.0	32.2	53.2	62.3	10.0	15.2	10.0	19.4	45.6	57.8	48.6	17.5
Humor	49.5	73.1	51.8	37.1	54.9	56.9	29.9	56.8	38.7	33.3	34.7	29.2	29.7	33.0	43.3	61.3
Ideology	33.3	64.8	18.6	23.7	43.0	47.6	53.1	46.4	39.7	25.1	25.2	23.1	46.0	46.8	43.1	60.0
Impl. Hate	16.7	62.5	7.4	14.4	7.2	32.3	29.6	32.0	7.1	7.8	4.9	9.2	18.4	19.2	16.3	3.7
Misinfo	50.0	81.6	33.3	53.2	64.8	68.7	69.6	77.4	45.8	36.2	41.5	42.3	70.2	73.7	55.0	26.9
Persuasion	14.3	52.0	3.6	10.4	37.5	32.1	45.7	43.5	3.6	5.3	4.7	11.3	21.6	17.5	23.3	56.4
Sem. Chng.	50.0	62.3	33.5	41.0	56.9	52.0	36.3	41.6	32.8	38.9	41.3	35.7	41.9	37.4	44.2	21.2
Stance	33.3	36.1	25.2	36.6	42.2	43.2	49.1	48.1	18.1	17.7	17.2	35.6	46.4	41.3	48.0	76.0
Conversation Level Tasks																
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4
Document Level Tasks																
Event Arg.	22.3	65.1	–	–	–	–	–	–	–	–	8.6	8.6	21.6	22.9	22.3	23.0
Event Det.	0.4	75.8	9.8	7.0	1.0	10.9	41.8	50.6	29.8	47.3	47.4	44.4	48.8	52.4	51.3	14.8
Ideology	33.3	85.1	24.0	19.2	28.3	29.0	42.4	38.8	22.1	26.8	18.9	21.5	42.8	43.4	44.7	51.5
Tropes	36.9	-	1.7	8.4	13.7	14.6	19.0	28.6	7.7	12.8	16.7	15.2	16.3	26.6	36.9	44.9

Table 3
Zero-shot Classification Results across our selected CSS benchmark tasks. All tasks are evaluated with macro F-1, which is averaged across 5 prompt permutations for zero-shot models. Supervised baseline results are averaged over 3 random seeds. Best zero-shot models are in green. A dash indicates a model did not follow instructions.

Large Language Model as a Measurement Tool

Performance on various languages (Rathje et al. 2023)

Table 4. GPT-4 vs. Top-Performing Machine Learning Models

Language	Construct	Top-performing GPT model F1	Top-performing GPT model	Top-performing alternate model F1	Model type	Year of study
English	Sentiment	0.685	3.5 Turbo	0.677	LSTM-CNN	2017
Arabic	Sentiment	0.746	4 Turbo	0.610	Naive Bayes	2017
English	Discrete emotions	0.782	4 Turbo	0.785	BERT	2020
Indonesian	Discrete emotions	0.785	4 Turbo	0.795	BERT	2020
English	Offensiveness	0.746	4	0.829	BERT	2019
Turkish	Offensiveness	0.762	4 Turbo	0.826	XLM-BERT	2020
Swahili	Sentiment	0.560	3.5 Turbo	0.657	Fine-tuned XLM-R	2023
Hausa	Sentiment	0.682	4 Turbo	0.826	Fine-tuned XLM-R	2023
Amharic	Sentiment	0.646	4 Turbo	0.640	Fine-tuned XLM-R	2023
Yoruba	Sentiment	0.681	4 Turbo	0.800	Fine-tuned XLM-R	2023
Igbo	Sentiment	0.622	4	0.830	Fine-tuned XLM-R	2023
Twi	Sentiment	0.505	4	0.675	Fine-tuned XLM-R	2023
Kinyarwanda	Sentiment	0.661	4 Turbo	0.726	Fine-tuned XLM-R	2023
Tsonga	Sentiment	0.448	4 Turbo	0.607	Fine-tuned XLM-R	2023
Average	-	**0.665**	-	**0.735**	-	-

We compare the performance of GPT-3.5 and GPT-4 to the performance of the top machine learning models reported in the papers from which we retrieved the tested datasets. All top-performing model statistics (besides the GPT statistics) are taken from the papers from which the datasets originated. GPT sometimes out-performed the top-performing fine-tuned models, or at least came close to the performance of these top-performing models. The abbreviations are as follows: LSTM, Long Short Term Memory; CNN, Convolutional Neural Network; BERT, Bidirectional Encoder Representations from Transformers; XLM, Cross-Lingual Model; XLM-R, XLM combined with RoBERTa (a variant of BERT with more extensive pretraining).

Choosing a Model

In addition to high performance, we should consider

- Replicability
  - Much attention is focused on proprietary models (e.g., GPT family)
  - Possible model change over time within the same version
  - Deprecation of earlier versions (no longer accessible through API)
- Scalability
  - Models run locally depend on available computing power
  - API-based models have issues of rate limits and costs

Coding procedure

- Define concept (could reference the literature)
- Build a set of human labels (S)
  - Ideally with multiple coders (check ICR)
- Write a prompt
- Check the LLM performance on a subset of the training set (S_t)
- Iterate over the previous two steps
  - Update the prompt in a way that improves
  - Make sure that you do not get overly influenced by model results
- Once the performance reaches a satisfactory level, validate it on another subset of the training set (S_v)

Refining Prompts

Writing effective prompts can be a challenge

- Often requires multiple iterations of updating/testing
- There is no single answer, which necessitates *calibration* (Zhao et al. 2021)
- Useful materials
  - From DeepLearning.AI
  - From OpenAI

Hyper-parameter Tuning

Controlling stochastic behavior

- Temperature: adjusts the probability distribution over words
  - Low (high) temperature: increases (decreases) the probabilities of higher-probability words → more deterministic outcome
  - GPT: the default value is 1
- Top-K: limits token selection to the top K tokens
- Top-P: limits token selection to the tokens until the cumulative probability exceeds P%

Discussion: Inconsistent Comparison

What does it mean, “LLMs perform better than humans”?

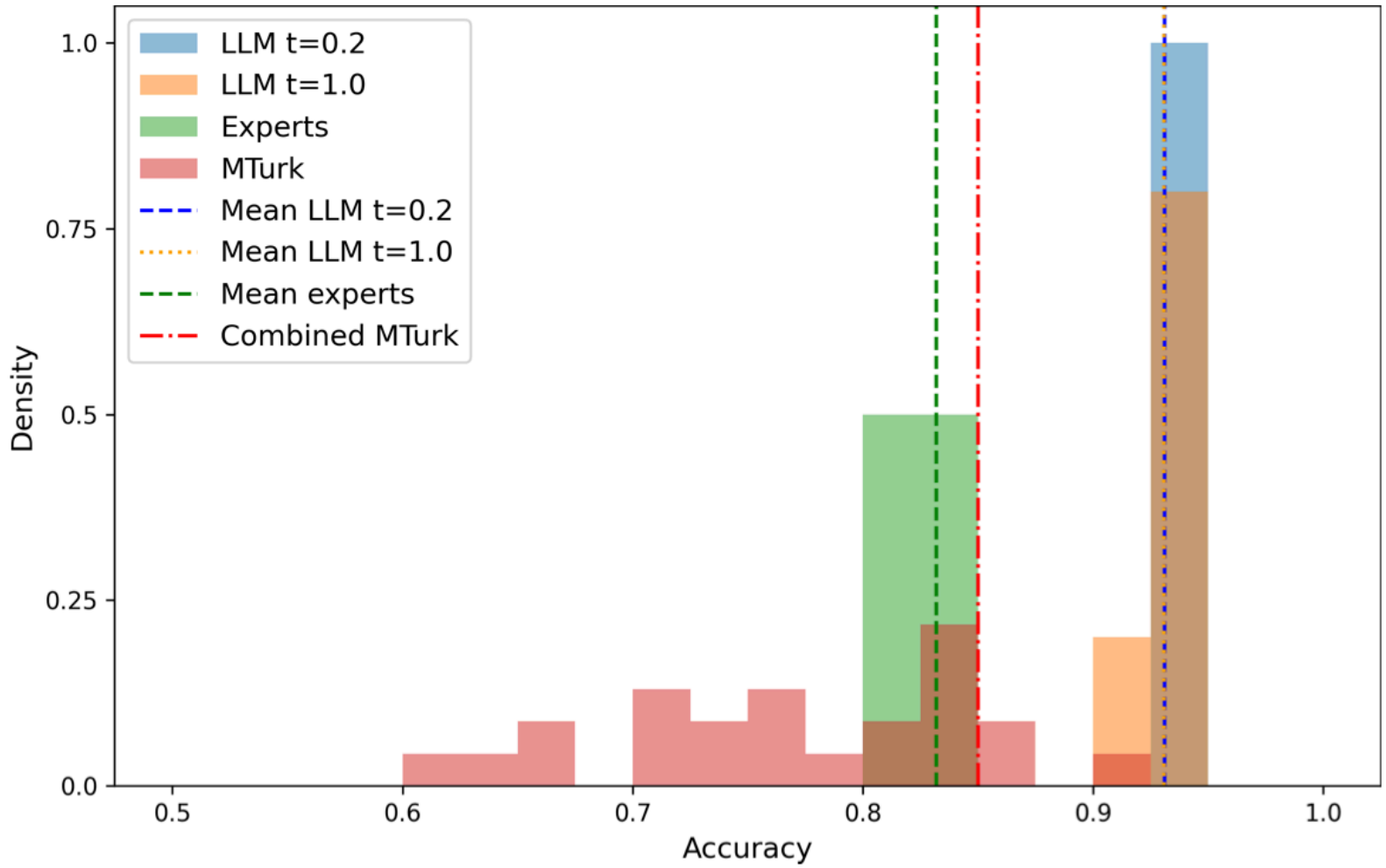
- To say how “good” an LLM model is, we need *true values* against which to evaluate it
- In addition, we need a baseline approach against which to compare the LLM model
- However, studies on the use of LLMs as a measurement tool are inconsistent in these regards
- This causes difficulty evaluating *validity*

Discussion: Inconsistent Comparison

What does it mean, “LLMs perform better than humans”?

- Case 1: there is objective truth, and **LLMs perform better than human coders**
- E.g., Tornberg (2023)
  - Predicting legislators’ political party using their public statements
  - LLMs do a better job than expert coders and crowd workers

Discussion: Inconsistent Comparison



Discussion: Inconsistent Comparison

What does it mean, “LLMs perform better than humans”?

- Case 2: there is no objective truth, it is proxied by **expert coders**, and **LLMs** outperform **crowd workers** in predicting **expert coders**’ decisions
- E.g., Gilardi et al. (2023)
  - Expert coders’ decisions on six concepts related to content moderation policy (expressed in tweets & news articles)
  - Compares **GPT** vs. **MTurk workers** in the task of predicting **expert coders**’ decisions

Discussion: Inconsistent Comparison

“ChatGPT outperforms MTurk for most tasks across the four datasets. On average, ChatGPT’s accuracy exceeds that of MTurk by about 25 percentage points”

→ However, expert coding and crowdsourcing are considered two alternative frameworks for generating “ground truth” data

Discussion: Inconsistent Comparison

What does it mean, “LLMs perform better than humans”?

- Case 3: there is no objective truth, it is proxied by one group of **expert coders**, and **LLMs** are compared with **another group of expert coders**
- E.g., Rytting et al. (2023)
  - Identifying topics in Congressional documents (U.S.) and NYT articles
  - Ground truth data generated by expert coders in a previous study
  - The ground truth is predicted by LLMs and new set of coders
  - “overall, these results again demonstrate that GPT-3 generally achieves on-par performance with humans”

Discussion: Inconsistent Evaluation Metric

Intercoder reliability (ICR)

- Metric for how much coders agree when coding the same data set
- In the context of machine learning, it can be used as a metric for evaluating a model's performance (here the model is seen as a coder)
- Higher ICR is *generally* preferred

Discussion: Inconsistent Evaluation Metric

How is ICR used in recent works?

- Case 1: ICR between multiple runs of an LLM model is used as a performance metric
- E.g., Gilardi et al. (2023)
  - “intercoder agreement is computed as the percentage of tweets that were assigned the same label by two different annotators (research assistant, crowd workers, or ChatGPT runs)”

Discussion: Inconsistent Evaluation Metric

How is ICR used in recent works?

- Case 2: ICR between expert coders' decisions can improve when an LLM's decisions are added
- E.g., Rytting et al. (2023)

Discussion: Inconsistent Evaluation Metric

- “adding GPT-3 as a coder adds a great deal to reliability for two measures (positivity, groups), slightly increases reliability of the coding for two others (extremity, issues), and reduces reliability in one (traits) … since adding GPT-3’s outputs to the human outputs generally either increases or maintains ICC across each attribute, we conclude that GPT-3 achieves human or better performance at this task”

Discussion: Inconsistent Evaluation Metric

Despite possible stochastic behavior, LLMs are likely to generate similar answers given the same instruction, leading to higher inter*coder* reliability

→ Equating this with ICR among human coders and even with performance makes little sense

Summary

- With continuous improvement, LLM performance might become comparable to (relatively small) carefully fine-tuned pre-trained models
- To serve as a valid, reliable, replicable measurement tool, much more work is necessary
  - Framework for selecting ground-truth data & baseline models
  - Careful use of evaluation metric(s)
  - Systematized optimization
  - Improved replicability