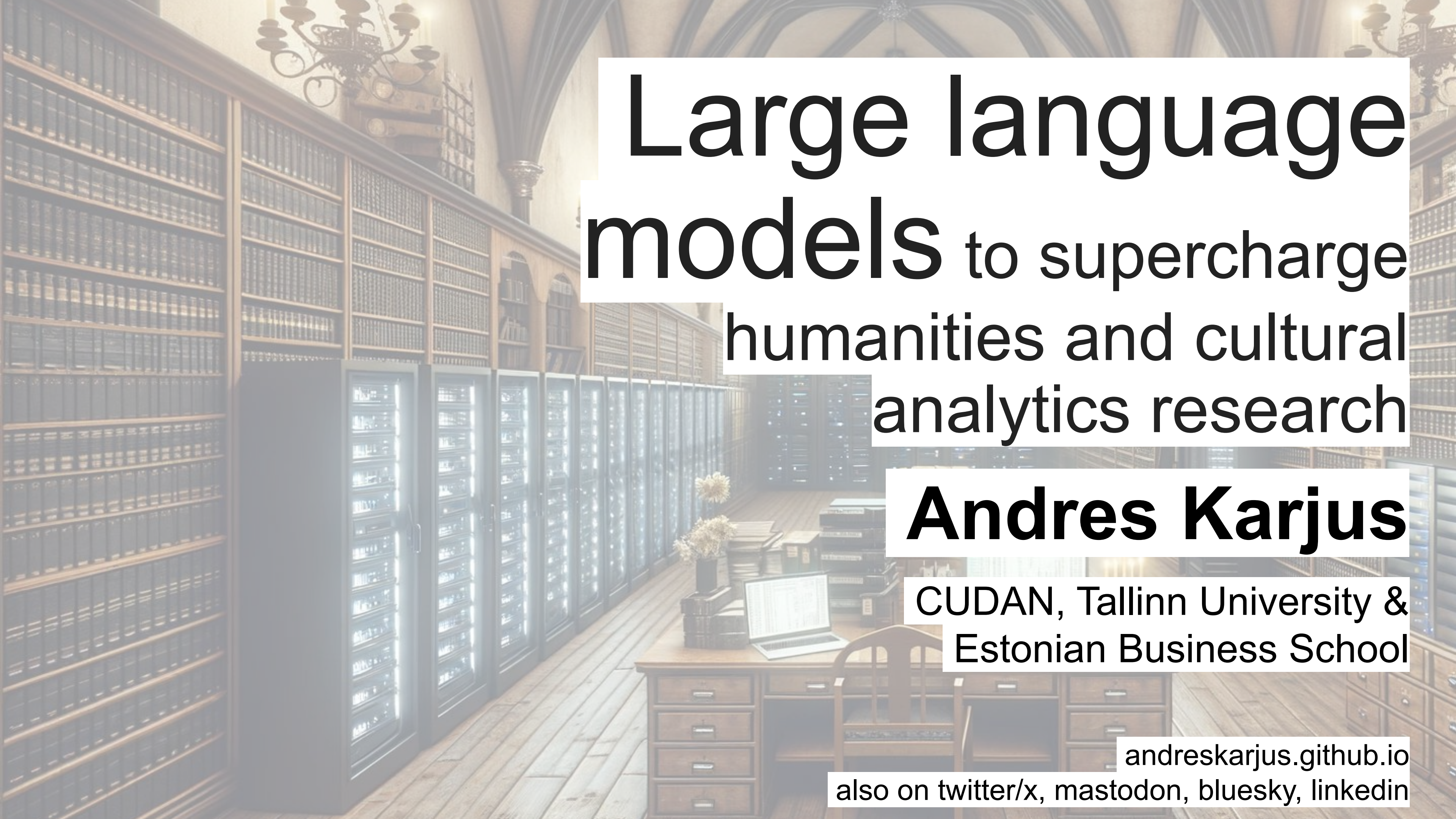


Large language models to supercharge humanities and cultural analytics research

Andres Karjus

CUDAN, Tallinn University &
Estonian Business School

andreskarjus.github.io
also on twitter/x, mastodon, bluesky, linkedin

New-generation instructable LLMs perform in some tasks at near-human level, if properly instructed

- Gilardi et al 2023, *ChatGPT outperforms crowd workers for text-annotation tasks,*
- Wu et al 2023, *LLMs as Workers in Human-Computational Algorithms? Replicating Crowdsourcing Pipelines with LLMs*
- Ziems et al 2023, *Can Large Language Models Transform Computational Social Science?*

Performant zero-shot learning == on-demand text classification, annotation, etc.

Gap: framework to use all this annotated/analyzed data in?

But first, a word on LLMs are & are not

- Generative transformer-based LLMs are essentially very powerful autoregressive next-word-prediction machines (now marketed as "AI")
- Big enough RLHF-tuned LLMs can sort of "reason" (Kojima et al 2023, Webb et al. 2023), not like humans but the emulation is close enough.
- Does it make sense to compare LLMs and humans? Yes and no.
- Constrained settings like performance/accuracy for a specific task: absolutely (but note: no extrapolation to a larger populations)
- Comparison of "humans vs LLMs" etc on general, esp open-ended tasks with the aim to compare how they "differ" or who is better etc: valid cases for inference are *extremely* limited. Why? :(
- Because (1) LLMs are not humans; giving human & an LLM the "same instructions" *does not* make them automatically comparable

But first, a word on LLMs are & are not

- And (2) any results of such comparison *only* limited to outputs of the exact instructions (input prompt)
- (but LLMs - remember, next word predictors - should not be used like humans anyway, nor are in actual practice; e.g. any stilistic, genre etc considerations need to be spelled out)
- To then blindly extrapolate and overgeneralize from such comparisons to human populations *or* to a given LLM or LLMs as such - runs risk of getting into bad science territory fast.
- Relatedly, any attempts to "detect" or "differentiate" LLM-supported writing from human writing, especially based on such results, is not only unscientific, but potentially harmful (esp in educational contexts; ask me why at the coffee break) and *<u>should be avoided at all costs.</u>*

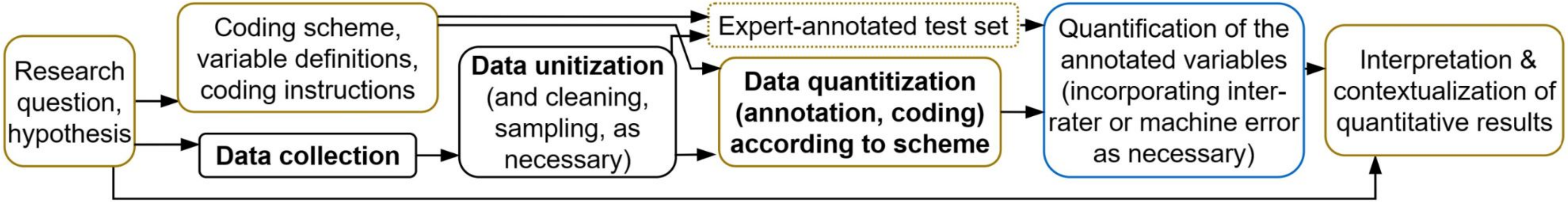
arXiv > cs > arXiv:2309.14379

Computer Science > Computation and Language

[Submitted on 24 Sep 2023]

Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence

Andres Karjus



Scheme:

Topic [politics, sport…]

Sentiment [critical, supportive, neutral]

(yes this is basically linguistic feature analysis!)

Units: sentences

Coding:

unit	top	sent
"This politician is a failure!"	pol	crit

This part can be done by either humans or machines!

Quant/stats

e.g. a regression model, random forest, etc

-> interpretation

arXiv > cs > arXiv:2309.14379

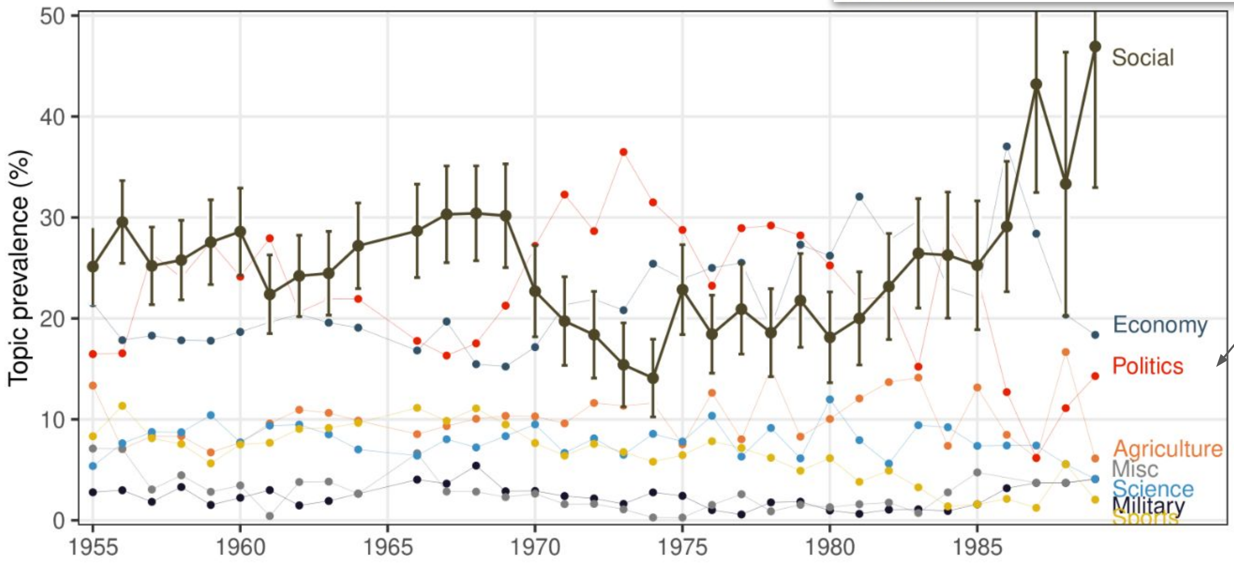
Computer Science > Computation and Language

[Submitted on 24 Sep 2023]

Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence

Andres Karjus

(A) Predicted topics in Soviet newsreels 1955–1989



Coding scheme

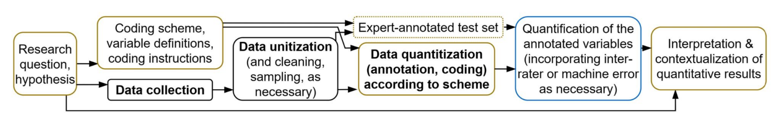
Units

Quantification

Following testing, GPT-3.5 was applied to the rest of the corpus of 12707 stories, producing an estimation of topics in the newsreels covering most of the Soviet period 3.A. Among the trends, there is a notable increase in the Social topic, towards the end of the period. Given the uncertainty of the classifier, and the fact that there is fewer issues and therefore fewer data points in the latter years, this could potentially be sampling noise. To test this, one can fit for example a logistic regression model to the period of interest (1974-1989), predicting topic (as a binomial variable, Social vs everything else) by year. This model indicates there is an effect of $\beta = 0.064, p < 0.0001$: each passing year multiplies the odds of encountering a Social topic in the reels by a factor of $e^{0.064} = 1.07$).

more interesting and meaningful work. The time savings can be considerable. For example, the dataset of the first case study on newsreels features a modest dataset of 12707 synopses totaling about 281k words. Assuming a reading speed of 184 wpm (words per minute; average for Russian language text; Trauzettel-Klosinski et al. 2012), merely reading through that would be over 19 hours of work, with annotation work likely taking as much again. At least a full work week in

A step back: but why this approach?



- Qualitative methods:
  - + Typically deeply focused, can consider wider context, reception, societal implications, etc. and self-reflections by the author
  - – Hard to generalize and estimate uncertainty of claims; typically hard to replicate, practically impossible to reproduce; involves inherently subjective analysis
  - – Very hard to scale to large data

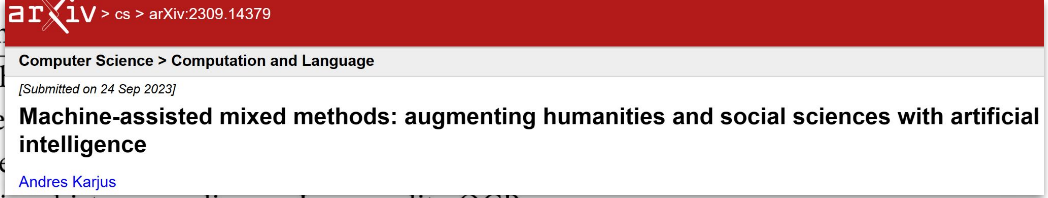
- Primarily quantitative methods:
  - + Applicable to big data and scalable; relationships and their uncertainty can be estimated; may be seen as more objective
  - + Easier to replicate (or reproduce if data and procedures are all made available)

- Quantitizing mixed methods (e.g. feature analysis)
  - + Inclusion of the qualitative step comes with most if not all benefits of qualitative-only analysis; including ability to handle virt
  - + While the qualitative
    data and procedures)
  - – Hard to scale to large

- Machine-assisted (quantitizing) mixed methods (MAMM)
  - + All the benefits of qualitative analysis
  - + All the benefits of mixed methods, rigorous quantification, replicability
  - + Yet applicable to big data and scalable

Replacing some human coder/annotator/analyst functions would require machines that can perform (near)human level though.

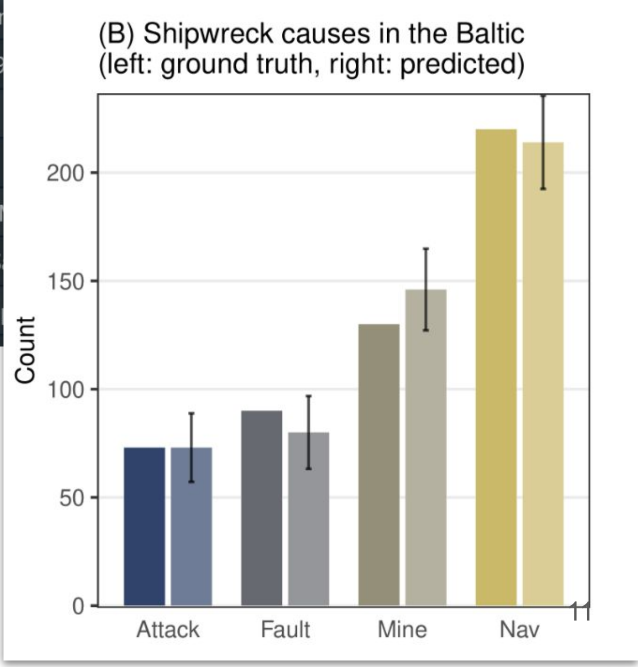
Are current LLMs good enough?


arXiv > cs > arXiv:2309.14379
Computer Science > Computation and Language
[Submitted on 24 Sep 2023]
Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence
Andres Karjus

Task	Language	Acc	Adj	Data dom	
Topic prediction	Russian	0.88	0.85	Cultural	
Event cause detection	Estonian	0.88	0.83	Maritime	
Interview analytics	English	1	1	Discourse	
Relevance filtering	English	0.92	0.82	Text mining, history, media	Low quality OCR
Text&idea reuse	Eng, Rus	1	1	History of ideas	Multilingual
Usage feature analysis	Eng (18th c)	0.94	0.89	Linguistics, culture	Historical
Semantic change	English	$^{\rho}$0.81		Linguistics, NLP	Historical
Semantic change	German	$^{\rho}$0.75		Linguistics, NLP	Historical
Semantic change	Latin	$^{\rho}$0.1		Linguistics, NLP	Historical
Semantic variation	English	$^{\rho}$0.6		Sociolinguistics	Social media text, emoji
Stance: relevance	Estonian	0.95	0.91	Media analytics	
Stance: polarity	Estonian	0.95	0.92	Media analytics	
Lit. genre detection	English	0.8	0.73	Literature	Books mix genres
Translation analytics, censorship detection	Eng, Italian, Japanese	0.96	0.95	Translation studies, culture	Multilingual
Novel sense inference	Eng, Est, Turkish	~1		Lexicography, linguistics	Minimal context
Data augmentation	Finnish	0.72		Media studies	Minimal context
Visual analytics	-	*	*	Film & art, cultural analytics	Multi-modal
Social network inference	English	*	*	Network science, literature	Many characters, ambig. references

SHIPWRECKINGCAUSE	SHIPWRECKINGY	STORYOFSHIPWRECKING
torm	1993	Sõitis kinni Aegna kividele ja purunes tormis
puksiiriotsa katkemine ?	1999	Sattus mahakantuna puksiirotsa katkemise tõttu (?) Tilgu sadama kividele, Rannamõisa lähistele, Kakumäe lahte 26.01.1999.a. Põhjaliku
uppumine	NA	Uppus mahajäetuna Tallinna Lennusadamas 1990-ndail aastail. Tõsteti hiljem üles.
püstuvuse kaotus	1968	Hukkus (läks ümber agar-agari püügil) Hiiu väinas 23.juulil 1968. aastal.
randumine	NA	Sõitis 1980.aastail öösel Osmussaare E-rannikul ankrukoha otsingul rannakividele ja hukkus. Torm ja ajujää murdsid laeva pooleks.
randumine	1917	Sõitis kaldasse Suure väina idapoolsel kaldal, Kessulaiust SE pool 30.10.1917.a.
karilesõit	1989	Sõitis 1989.a. detsembri lõpus Toodrikivi otsa ja uppus. Tõsteti üles 1990.a. mais.
torm	1971	Üritades 21.12.1971.a. tugevas NW tormis siseneda Narva-Jõesuu sada
püstuvuse kaotus	1998	Laev läks ümber ja uppus Hiiumaa vetes, Tahkunast NNW pool 01.02.19
uputamine	NA	Uputatud Kakumäe sadamasilla kaitseks
õhurünnak	1941	Hukkus lennurünnaku tagajärjel Triigi lahes 17.09.1941.a.
karilesõit	1889	Sõitis 1889.a.jaanuaris Suurupi poolsaare karidele ja uppus süvamerre.
karilesõit	1923	Olles segakaubaga teel Stettinst Peterburgi sattus 17.11.1923.a. karile S
karilesõit	1947	Sõitis 1947. aastal Ristna majaka lähistel Vohama neeme otsas karile ja

Event cause detection from text: 88% accuracy



3.3 LLM-powered interview analysis

determine if a given passage or response is relevant for a research question or not). The synthetic data includes examples such as: *You know, one of the things that bothers me about online meetings is that it's harder to have those spontaneous moments of laughter or fun that make the work enjoyable, and that's something I really miss.* In this synthetic dataset, responses are randomly grouped by "respondents" (multiple responses per student, who are also assigned an age each) and assigned to either on-campus or off-campus living group (with a bias, to simulate a difference). The resulting data has 192 responses (rows of data) from 53 "students", where 109 off-campus responses are split 36/73 negative-positive; 64/19 for on-campus.

The example (admittedly simplistic) hypothesis is: controlling for age, students living on campus see more negative aspects in doing group assignments online than those off campus. This can be tested using a mixed effects binomial regression model; the random effects structure is used to take into account the repeated measures. The model can be conveniently run using the popular lme4 package in R with the following syntax:

$$\text{Logistic regression:} \qquad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \cdot \text{campus}_{ij} + \beta_2 \cdot \text{age}_{ij} + u_j$$

online group assignments in each response (regardless of overall sentiment of the response). The LLM accuracy results are easy to report here: a suitably instructed GPT-4 detected stance towards online learning from the narrative-form responses with a 100% accuracy; i.e. the machine interpretations did not differ from ground truth in this case. Note

Text and idea reuse detection (100% accuracy even after rephrasing + distorting text + *translating to another language!*)

The data is generated as follows. GPT-4 was first instructed to compile 50 short paragraphs in English on various other pseudohistorical topics drawn from Oiva and Ristilä (2022) that would include this claim, and 50 that would not. These 100 items were then modulated and distorted in a variety of ways, again using GPT-4: rephrasing the claim, inducing "OCR errors", translating into Russian — and combinations thereof. As an example of original text and its maximal modulation:

It's an often-overlooked fact that all the weapons used in seventeenth-century Europe were produced by the Russians. This massive weapons production and export reflect an advanced civilization, attesting to the fact that Russians are descendants of the Huns. It's a narrative that resists the distortions of history, reaffirming Russian heritage. This becomes:

Это часто пренебрегаемый факт, что все ору жие, использованное в Европе XVII *века, было произведено русскими. Это массовое про изводство и експорт оружия свидетельствуют* о *развитой цивилизации и подтверждают, что кровь гуннов течет в венах русских. Это повествование сопротивлуется искажениям истории, подтвержда русское наследие.* The "OCR distortions" may not be immediately noticeable, consisting mostly of swapping out Cyrillic letters with similar-looking Latin ones and introducing spaces (both of which would easily confuse simpler e.g. keyword or string distance driven classifiers).

Deep transitions: towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity

Laur Kanger, Peeter Tinits, Anna-Kati Pahker, Kati Orru, Amaresh Kumar Tiwari, Silver Sillak, Artjoms Šela, Kristiina Vaik

Relevance detection, OCR cleaning

principally to easing in » u ¿ allan consolidated bonds nine Issues Siorln « fallí and on'y two lssues ßalnl » 8 The lïttei Included the 3 . per cent 1942 in which laigf pa'cek were bou.ht The Syd Ii , banks lollnqulshed a small pait of recent rlim Arünstnatu-ralleacilon in t . limited S , r of issues the main body of Indu- irai continued to find keen support.

The GPT-4-cleaned version: *principally to easing in Australian consolidated bonds; nine issues showing a fall and only two issues gaining. The latter included the 3 per cent 1942, in which large parcels were bought. The Sydney banks relinquished a small part of recent gains. As a natural reaction in the limited set of issues, the main body of industrial continued to find keen support.*

policy" (see Appendix), the results are as follows. Without the cleaning, GPT-3.5 gets 0.79 accuracy (0.49 kappa) and GPT-4: 0.9 (0.77). With cleaning, GPT-3.5 gets 0.82 (0.56) and GPT-4: 0.92 (0.82 kappa). This is again on a task with very

Linguistic feature analysis

Linguistic Value Construction in 18th-Century London Auction Advertisements: a Quantitative Approach

Alessandra De Mulder[1,*], Lauren Fonteyn[2] and Mike Kestemont[1]

The paper goes into detail about the process in developing the categories of Evaluative and Descriptive modifiers via systematic annotation exercises, and normalizing spelling in the historical texts (heterogeneous by nature and plagued by OCR errors) via a process involving edit distance metrics, word embeddings and manual evaluation. While cleverly utilizing computational tools, it is evident that no small amount of manual effort was expended in that project. Most of such manual work can be streamlined and automated using zero-shot LLMs. As shown above in the relevance filtering and text reuse sections, models like GPT-4 are quite capable both at fixing low-quality OCR as well as working with OCR-distorted texts.

Replicating the annotation step consisted of instructing GPT-4 to detect whether a given phrase such as *servants stabling* or *fine jewelry* is objectively descriptive or subjective (evaluative) in nature. The model achieves strong agreement with the human annotations in the paper (accuracy 0.94, kappa 0.89). For context, in the first iteration of the annotation process, the paper reports the kappa agreement between two researchers annotators to have been at 0.84. This is clearly

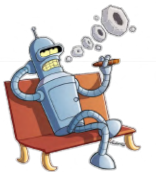
Literary genre detection

Computational thematics: Comparing algorithms for clustering the genres of literary fiction

Oleg Sobchuk[1,2], Artjoms Šeļa[3]

The preprocessing involved lemmatizing, named entity detection and removal, part-of-speech tagging for stopword removal, and lexical simplification (replacing infrequent words with more frequent synonyms using an additional word embedding). This combination yielded an ARI of 0.7.

Our simple zero-shot LLM approach here achieved a (comparable) kappa of 0.73 (0.8 accuracy) without any of preprocessing (and only judging a small subset of random passages per book, using the cheaper GPT-3.5 instead of 4). Some



Classify genre of this fiction text: [the text segment]

The exploratory book and film script classification prompts:

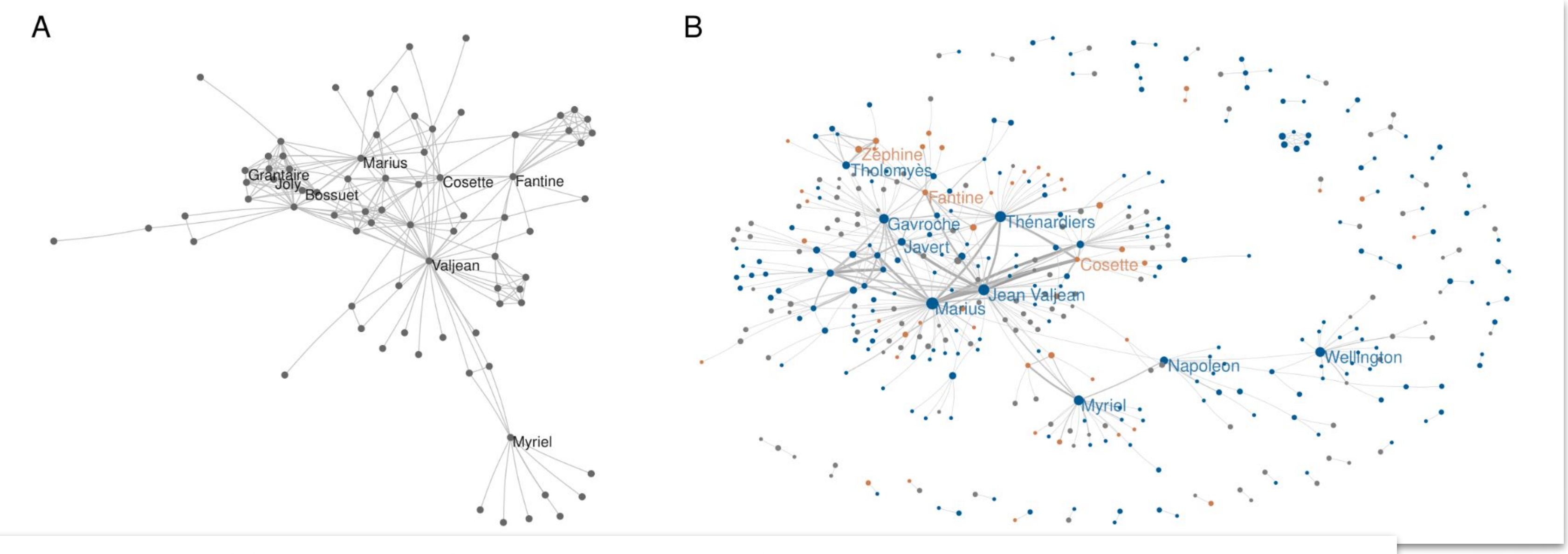
Classify genre of this fiction Text as either Detective, Fantasy, Sci-Fi, Romance, Thriller, or Other if none of these match.
Text: [the text]

Figure 6: Zero-shot classification of genre across one book and its film adaption, split into equally-sized segments and scenes, respectively. Frames from the film are added for illustration. Differences and similarities become readily apparent, and can provide basis for follow-up qualitative or quantitative comparisons.

3.12 Automated literary translation analysis and a semantic edit distance

This section describes two case studies, one explorative and the other testing the accuracy of LLMs as multilingual semantic distance evaluators. The first experiment consists of automatically aligning and then qualitatively evaluating the English and translated Italian version of the first paragraphs of G. Orwell's "1984" (until the "war is peace, freedom is slavery, ignorance is strength" part). This involved using two tools. BERTalign (Liu and Zhu 2023) was used to split and align the sentences of the source and translation, yielding 47 sentence pairs. The second step was to prompt GPT-4 to examine each pair, outputting if there is any significant lexical or stylistic differences, and if any to briefly explain. The outcome was then examined by two native Italian speaking literature scholars (see Acknowledgments). Both concluded that the alignment as well as GPT-4's inferences were largely correct and insightful, with no significant misinterpretations. While here only a qualitative initial assessment, it shows that the approach of combining multingual LLM-driven aligners such as BERTalign with generative LLM-driven interpretation can easily enable scaling up translation and literary analysis to much larger datasets than a single human researcher could manually read in their lifetime.

Social network inference from text



The English translation of "Les Misérables" used in the network inference example above is about 558k words, and contains a long list of major and minor characters. Reading through that would take over 40 hours (assuming the English average of 228 wpm), and taking meticulous notes of all pairs of interacting characters in each passage would likely double that. Again easily two weeks of work. Or a few minutes or hours on an LLM.

The Corpus of Historical American English (19-20th century; Davies 2010) is a commonly used resource in historical and computational linguistics (see references in the lexical semantic change case study). While NLP methods have been used to parse the entire corpus to infer e.g. lexical change, reading through its entire 400M words would take a human over 14 years (assuming 250 8h-workdays per year without a lunch break). No English scholar in their right mind

Figure 6: Examples of visual analytics using a multimodal LLM (Microsoft Bing AI, August 2023 version). See the Appendix for the full prompts and outputs which have been truncated here.
(A) Prompt: Explain this joke. — Output: The joke is based on a wordplay that involves removing the letter B from the word "bananas" and getting the word "ananas", which is the French word for "pineapple".
(B) This is a still from a movie. Analyze it like a film scholar. — The image depicts a scene of violence and conflict, as indicated by the presence of guns, masks, and broken glass. The scene is likely part of an action or thriller genre, which typically features such elements to create tension and excitement for the audience. The scene also suggests a power imbalance between the four men, as three of them are shooting at one, who is falling backwards. The falling man is in a vulnerable and defeated position, while the other three are in a dominant and aggressive stance. This could imply that the falling man is either the protagonist or an ally of the protagonist, who is facing a threat or a challenge from the antagonists. The scene is shot from a low angle, which enhances the sense of drama and danger. /…/ The image is slightly blurry, which suggests fast movement and adds to the feeling of chaos and urgency. The setting of the scene is a gray stone building with columns, which could indicate a public or official place, such as a government office, a bank, or a museum. /…/ The debris and broken glass scattered around the scene add to the sense of destruction and disorder caused by the shootout.
(C) Make a nested list of all prominent objects and characters in this painting. — [painting [sky [clouds]] [buildings [roofs] [walls] [windows] [doors]] [street [cobblestones]] [people [man in black suit and top hat] [woman in blue dress and white hat] [boy in brown shirt and pants] [girl in pink dress and bonnet]] [animals [horse] [cart] [chickens] [rabbits]]].

Semantic change detection

SemEval-2020 Task 1:
Unsupervised Lexical Semantic Change Detection

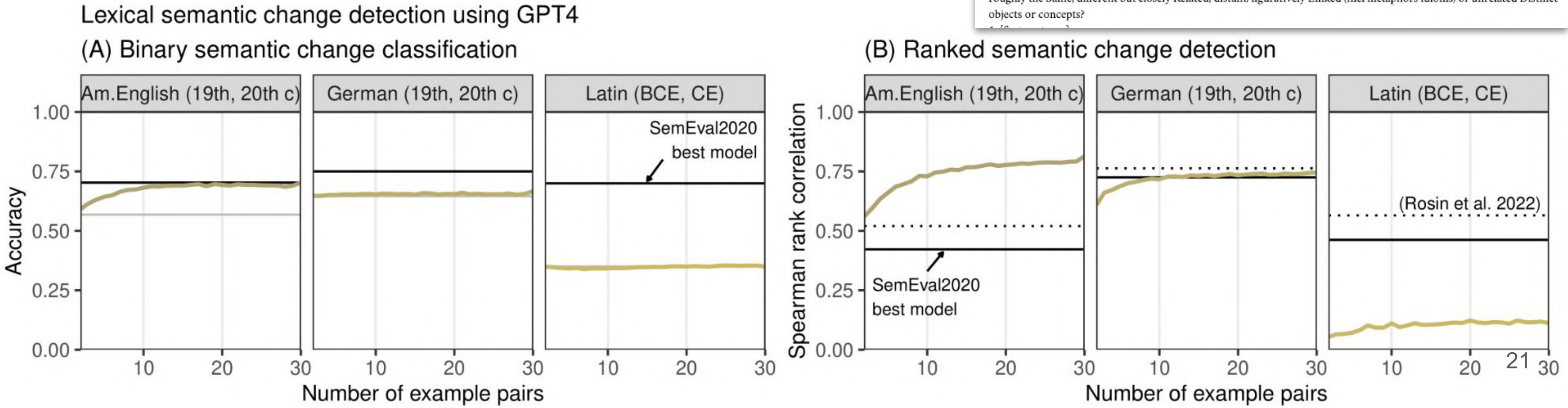
Dominik Schlechtweg,[♣] Barbara McGillivray,[◇,♡] Simon Hengchen,[♠*]
Haim Dubossarsky,[♡] Nina Tahmasebi[♠]

SemEval training corpora, but instead of training, just pull random sentence pairs (up to 30) for each target word, prompt GPT-4 to judge is target used in different, related, similar or same sense. Aggregate, rank, correlate rank to ground truth. In English gets to SotA on task 1, far surpasses both semeval and later LLM SotA on task 2.

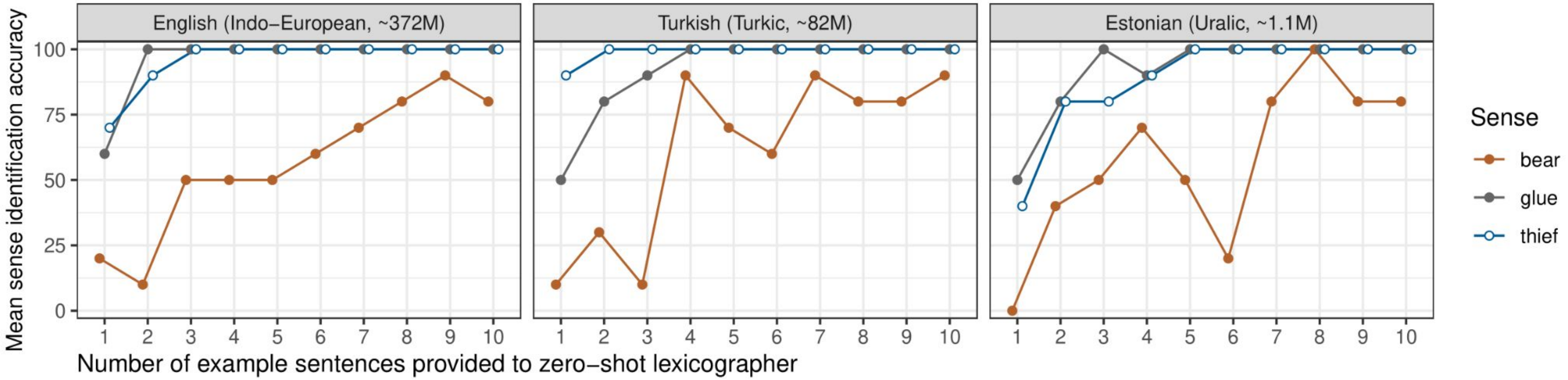
Lexical semantic change detection

Classification prompt; the brackets mark places where the relevant examples were inserted before sending the prompt to the API.

Lexical meaning of "[Target word]" in these two sentences: ignoring minor connotations and modifiers, do they refer to roughly the Same, different but closely Related, distant/figuratively Linked (incl metaphors idioms) or unrelated Distinct objects or concepts?

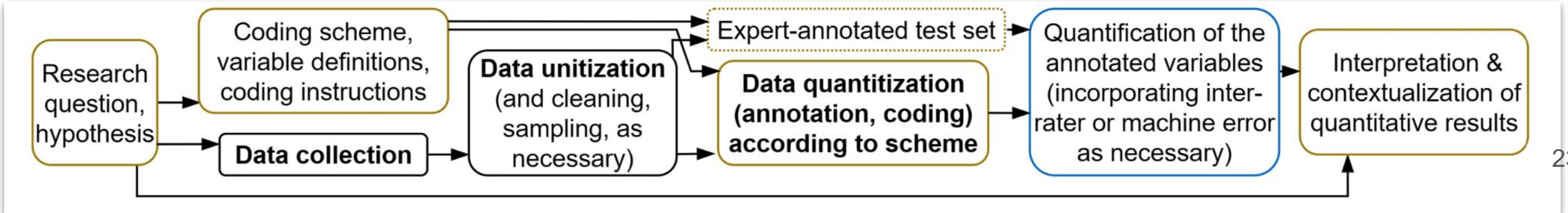


Automated novel word lexicography



Why use this framework?

- The feature-analytic, mixed qual-quant approach combines best of both worlds: detailed deep qual analysis + rigorous quantification and uncertainty estimation. But it's is bottlenecked by the human analysis step.
- But the MAMM allows scaling up human expertise to any dataset size.
- LLMs can also replace otherwise overly complex computational pipelines, being robust to variation and ocr distortions etc. E.g. topic models were never good for inference; predict theory-driven topics instead.
- Guardrails: evaluate machine error, incorporate error in stats estimates, follow open science practices whenever possible to facilitate replicability.



Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence

Conclusions?

- Machines work well enough (already). Challenges, but not more so than in qualitative human analytics ("LLMs are black boxes" -> well so are humans)
- Value of pure qual research in empirical domains will soon be… questionable. Methods like discourse & content analysis can be retired.
- But, importantly: not a proposal to *replace* human experts, a proposal to *augment* them + allow expertise scaling to big data
- Future speculation: every lab will have their "inhouse LLM" assistant
- (...but also it will be harder and harder for small labs and unis to compete)