Barcelona School of Economics

WORKING PAPERS

Uncovering the Semantics of Concepts Using GPT-4 and Other Recent Large Language Models

BSE Working Paper 1394| June 2023

Gaël Le Mens, Balázs Kovács, Michael T. Hannan, Guillem Pros

bse.eu/research

Uncovering the Semantics of Concepts Using GPT-4 and Other Recent Large Language Models

Gaël Le Mens (gael.le-mens@upf.edu)
Department of Economics and Business, BSE, UPF-BSM
Universitat Pompeu Fabra, Barcelona, Spain

Balázs Kovács (balazs.kovacs@yale.edu)
School of Management,
Yale University, New Haven CT, USA

Michael T. Hannan (hannan@stanford.edu)
Graduate School of Business,
Stanford University, Stanford CA, USA

Guillem Pros (guillem.pros@upf.edu)
Department of Economics and Business,
Universitat Pompeu Fabra, Barcelona, Spain

June 08, 2023 [*]

[*] G. Le Mens received financial support from an ICREA Academia grant and the Agencia Estatal de Investigacion (AEI), though the Severo Ochoa Programme for Centres of Excellence in RD (Barcelona School of Economics CEX2019-000915-S) and grant PID2019-105249GB-I00/AEI/10.13039/501100011033. B. Kovács was supported by Yale School of Management. M. Hannan was supported by the Stanford Graduate School of Business. Data, material, analysis code and python notebooks used to interact with the openAI API will be made available online at https://osf.io/emhwc/ upon acceptance of the article for publication.

Abstract

Recently, the world's attention has been captivated by Large Language Models (LLMs) thanks to OpenAI's Chat-GPT, which rapidly proliferated as an app powered by GPT-3 and now its successor, GPT-4. If these LLMs produce human-like text, the semantic spaces they construct likely align with those used by humans for interpreting and generating language. This suggests that social scientists could use these LLMs to construct measures of semantic similarity that match human judgment. In this article, we provide an empirical test of this intuition. We use GPT-4 to construct a new measure of typicality–the similarity of a text document to a concept or category. We evaluate its performance against other model-based typicality measures in terms of their correspondence with human typicality ratings. We conduct this comparative analysis in two domains: the typicality of books in literary genres (using an existing dataset of book descriptions) and the typicality of tweets authored by US Congress members in the Democratic and Republican parties (using a novel dataset). The *GPT-4 Typicality* measure not only meets or exceeds the current state-of-the-art but accomplishes this without any model training. This is a breakthrough because the previous state-of-the-art measure required fine-tuning a model (a BERT text classifier) on hundreds of thousands of text documents to achieve its performance. Our comparative analysis emphasizes the need for systematic empirical validation of measures based on LLMs: several measures based on other recent LLMs achieve at best a moderate correspondence with human judgments.

Keywords: Categories, Concepts, Deep Learning, Typicality, GPT, ChatGPT, BERT, Typicality, Similarity.

JEL classification numbers: C18, C52.

1 Introduction

Do people tend to express opinions that align with the prevalent views in their social groups? Are individuals who express differing opinions penalized or rewarded for sharing such thoughts? Do consumers prefer cultural items that fall into established genres or those that challenge established norms? Do companies hire job candidates with a focused career identity or those with more diverse experiences? Is this negotiator thinking innovatively and creatively? Are interdisciplinary project proposals, which do not fit within a conventional academic discipline, more or less likely to secure funding from a grant agency? Are patent applications that clearly fit within a technology class more likely to be approved by patent reviewers?

Answering these questions requires determining the semantic similarity between text documents and concepts (or categories). Social science research refers to such assessments as “typicality judgments” or “typicality ratings” when they are formulated by humans, and as “measures of typicality” when they are based on computerized text analysis. In this article, we propose an innovative way to construct typicality measures from text data by harnessing the potential of a cutting-edge Large Language Model (LLM), OpenAI’s GPT-4. Our novel *GPT-4 Typicality* measure improves on the state-of-the-art in terms of its correspondence with typicality ratings (by human judges), and requires much less data than other methods. In other words, it is both more valid and more economical in terms of data requirements than prevailing methods. This method is thus potentially applicable to a variety of research settings.

Typicality in a concept refers to the degree to which an object is perceived as representative or prototypical of that concept. For instance, a typical Mystery book may be thought of as a suspenseful and engaging piece of fiction that challenges readers to solve the puzzle with the protagonist. A book with these characteristics will likely be judged as a typical Mystery book, while one lacking these features may not.

Issues of measuring typicality arise routinely in studies of economic organization. In the study of organizations and markets, the focus has been on how agents acting as audience members judge the offers of producers. A crucial part of the evaluation process involves assessing the fit of producers and their products with prevailing categories (Porac, Thomas, Wilson, Paton, and Kanfer, 1995; Zuckerman, 1999; Hannan, Pólos, and Carroll, 2007; Hannan, 2010; Hannan, Le Mens, Hsu, Kovács, Negro, Pólos, Pontikes, and Sharkey, 2019). In other words, concepts such as industry, product category, and cultural genre form the basis for audience expectations, and the extent to which producers and their products fit in established categories indicate whether they warrant attention. A similar process affects how employers evaluate potential job candidates (Hsu, 2006). Early sociological work studying typicality generally did not analyze feature values of objects, only their categorizations (Hsu, 2006; Hsu, Hannan, and Koçak, 2009; Kovács and Hannan, 2015; Porac et al., 1995; Zuckerman, 1999). For example, Kovács and Hannan (2010) studied the of restaurants, and they assumed that a restaurant classified as French and Japanese would have zero (minimal) typicality in all other cuisines such as Mexican or Californian. This approach had two serious limitations: (1) it could not account for the graded nature of typicality in concepts documented by psychologists (e.g. Rosch, 1973), and (2) it could not measure the typicality of items assigned to a single concept.

An approach recently proposed by Le Mens, Kovács, Hannan, and Pros Rius (2023) has significantly improved on previous methods. These authors focused on typicality measures constructed from textual descriptions of objects. However, instead of simple label co-occurrences, it relied on deep learning natural language processing (NLP), which allows for fine-grained measurement of typicality. This is possible because the text data it uses is less coarse than the categorical assignments used in prior research. Specifically, this approach involves a probabilistic text classifier based on BERT (Devlin, Chang, Lee, and Toutanova, 2018), a LLM released in 2018. The parameters of the text classifier are adjusted by training the model on a "training set" made of text documents and label assignments in a process called model "fine-tuning" the model. By applying the fine-tuned classifier on new text documents, the analyst obtains the categorization probability that the document is an instance of the concept. The typicality of the text document in the concept is then obtained by taking the logarithm of this categorization probability. In a systematic comparison with other approaches to typicality measurement, they showed that the resulting *BERT Typicality* measure provides a much better correspondence with human typicality ratings than previous approaches.

In this paper, we conduct a similar comparison, but this time involving *GPT-4 Typicality* and other measures based on GPT-3, GPT-3.5, and GPT text embeddings. We define the performance of a model-based typicality measure in terms of its correspondence with human typicality ratings across two test data sets. The first test set is that used by Le Mens et al. (2023)[1]. The second test set was collected for the purpose of this article. It comprises a random set of Tweets published by US-Congress members between the opening date of the 118th Congress (Jan 3rd, 2023) and May 1st, 2023.

This article reports one of the first systematic efforts to empirically validate a measure based on the most recent class of Large Language Models (LLMs) for use in social science research. As such, it contributes crucial evidence to critically assess the widely shared *intuition* (among social scientists and others) that GPT-4 possesses human-like capabilities. It is important to emphasize that our focus differs from that of the prevailing discussion around ChatGPT. Much of this discussion focuses on generative capabilities, which are utilized by students to write essays, coders to auto-complete code,[2] consultants to summarize information, and more. By contrast, we focus on the properties of *the semantic space* used by LLMs to represent text documents. This article can be seen as a response to the recent call by leading experts in text analysis to go beyond "fitting predictive models to simple counts of text features" and use "richer representations" in social science research (Gentzkow, Kelly, and Taddy, 2019, p. 569). Our contribution is to show empirically that such richer representations not only offer some promise of improved measurement accuracy, but they deliver on this promise provided they are empirically validated.

2 Using GPT-4 to Obtain the Typicality of a Text Document in a Concept

Our approach consists in directly asking GPT-4[3] about the typicality of an object o in a focal concept CONCEPT based on a short textual description of the object o: "TEXT". We use the following prompt, submitted to GPT-4 via the API:

> Here's a DESCRIPTOR: 'TEXT'. How typical is this DESCRIPTOR of the CONCEPT? Pro-

[1] Available at https://osf.io/ta273/.

[2] https://github.com/features/copilot.

[3] We used the most recently published GPT model at the time of this research: GPT-4 (Snapshot from March 14, 2023).

> vide your response as a score between 0 and 100 where 0 means ‘Not typical at all’ and 100 means ‘Extremely typical’.

DESCRIPTOR is an informative label that would apply to all objects from which we aim to obtain typicality measures. For example, if we are interested in obtaining measures of the genre typicality of book descriptions in a literary genre, DESCRIPTOR would be replaced by “book description”. If we are interested in obtaining the typicality of CVs in a job category, DESCRIPTOR would be replaced by “CV”.

We obtain the LLM’s response as a string of characters. The vast majority of the time, the response consisted either in a number of up to 3 digits or such a number and a “%” symbol. In a small percentage of cases, the response did not contain a number, inconsistent with the instructions given in the prompt. We queried ChatGPT again, anticipating that the variability in its responses would yield numerical results. This is what happened. We programmed a loop to submit queries until a numerical response was provided for each text document in the dataset.

This approach is scalable, because researchers can obtain responses to about thousands of text documents in little time by submitting the prompts to GPT-4 using the ChatCompletion command of the OpenAI Chat API. Moreover, using the API allows for automatically submitting multiple queries in a row, each treated as a new chat. This last feature ensures that responses to earlier requests do not influence responses to subsequent requests, which maintains independence across observations.

It is important to note that this approach *does* not *involve any model training on the research data.* As we will see below, other approaches that do not involve training perform (based on the cosine similarity between representations of text documents in feature space) much more poorly than *GPT-4 Typicality* in terms of their correspondence with human typicality judgments. Moreover, approaches that use large amounts of training data, such as those involving model training and fine-tuning, do not perform better than *GPT-4 Typicality*.

3 Validating the *GPT-4 Typicality* Measure with Human Typicality Ratings

We validate the *GPT-4 Typicality* measure in two domains: the typicality of books in literary genres, based on short descriptions of the books, and the typicality of tweets written by US Congress members in the Democratic Party and in the Republican Party.

3.1 Typicality of Books in Literary Genres

The empirical domain is that of the genre typicality of books in literary genres such as Mystery or Romance.

Test Data

For the Mystery genre, these consist of a set of 1,000 book descriptions, 500 Mystery books and 500 books from other genres, used in Le Mens et al. (2023). For each book description, we have about 10 typicality ratings by human coders who responded to the question “How typical is this book to the Mystery genre,”

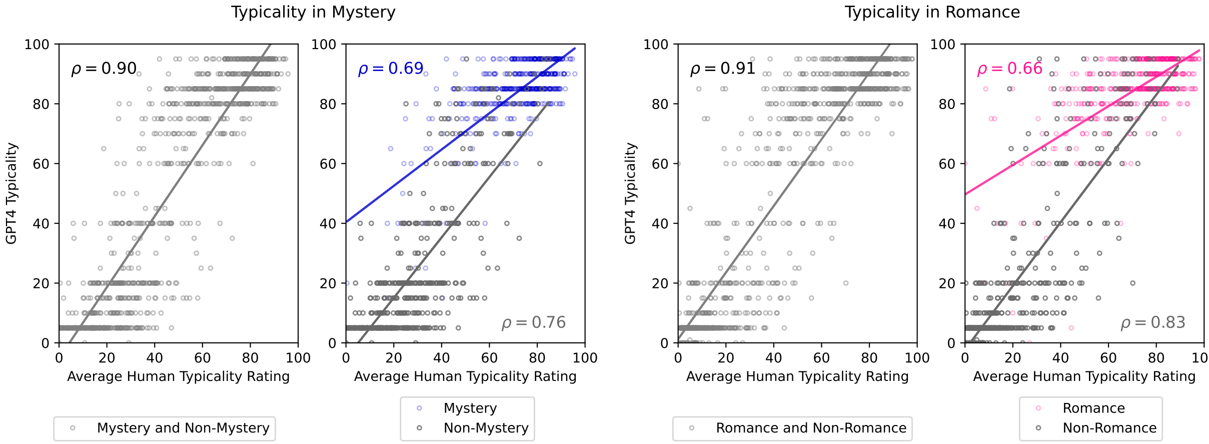


Figure 1: *GPT-4 queried about the genre typicality of the book based on its description:* There is a strong positive association between *GPT-4 Typicality* and *Human Typicality*. *LL panel:* All books in the validation data for the Mystery genre. *L panel:* The positive association holds for Mystery and Non-Mystery books. *R panel:* All books in the test data for the Romance genre. *RR panel:* The positive association holds for Romance and Non-Romance books.

using a 0 to 100 slider. The *Human Typicality* of a book description is the average of the typicality ratings across the participants who rated it. The test data from the Romance genre have the same structure.

Constructing *GPT-4 Typicality*

We directly ask GPT-4 about the typicality of a book in the focal concept based on a short description of the book. Assuming for this example that the focal concept is the Mystery genre, we obtained the *GPT-4 Typicality* by submitting the following prompt using the API:

> Here's a book description: 'BOOK DESCRIPTION'. How typical is this book of the 'Mystery' genre? Provide your response as a score between 0 and 100 where 0 means 'Not typical at all' and 100 means 'Extremely typical'.

Results

GPT-4 Typicality is extremely highly correlated with *Human Typicality*: ρ = .90 for the Mystery genre and ρ = .91 for the Romance genre. Moreover, the association holds among books from the focal category (e.g., Mystery books) and among books which do not belong to the focal category (e.g., Non-Mystery books)(see figure 1). *GPT-4 Typicality* therefore reflects between-book-description differences in *Human Typicality* beyond differences in category membership. In other words, it reflects the graded nature of typicality documented by psychologists that studied it in the laboratory with artificially created stimuli (e.g., Rosch, 1973).

3.2 Typicality of Tweets in Political Parties

The text documents consist of tweets published by members of the US Congress, and the focal concepts are their political parties. We aim to measure the typicality of a tweet in the Republican Party and the Democratic Party.

Test Data

These consist of a random sample ($N = 900$) of the universe of tweets published by members of the 118th US Congress between its opening date (Jan 3rd, 2023) and May 1st, 2023.[4] Importantly, the tweets comprising this test set were published after the collection of the GPT-4 pre-training data (which ended in September 2021). Therefore, cross-contamination between the tweets in the test set and the pre-training data is very unlikely. We obtained about 15 independent ratings of the typicality of each tweet in the Democratic Party and a similar number of independent ratings of the typicality of each tweet in the Republican Party. The *human typicality* rating of a tweet in a political party is the average of the typicality ratings across the participants who rated it. (See data collection details in the Supplementary Appendix.)

Constructing *GPT-4 Typicality*

We used exactly the same approach as with the book descriptions, minimally adapting them to the Tweet context. We use the following prompt (assuming for this example that the focal concept is the Republican Party):

> Here's a tweet written by a member of the US Congress: 'TWEET TEXT'. How typical is this tweet of the Republican Party? Provide your response as a score between 0 and 100 where 0 means 'Not typical at all' and 100 means 'Extremely typical'.

Results

GPT-4 Typicality is highly correlated with *human typicality*, with values of 0.85 and 0.78 for typicality in the Democratic Party and Republican Party, respectively. This is also the case among tweets published by members of a single political party (see figure 2). *GPT-4 Typicality* therefore reflects between-tweet differences in *human typicality* beyond differences in the parties of the politician who published them.

The ability of *GPT-4 Typicality* to capture differences in *human typicality* is excellent, even if the correspondence with human judgments is not as strong as in the context of the book descriptions. We can only conjecture about the sources of the difference. A potential explanation is that tweets contain less diagnostic information about the focal concept (the political party or the literary genre), because they are shorter and are not all written with the intention to convey diagnostic information about the focal concept. To evaluate

[4] These consist of around 91 thousand tweets. We downloaded the tweets on May 2nd, 2023. We used the list of congress members and Twitter usernames made available by `https://github.com/unitedstates/congress-legislators`. We consider only original tweets of more than 2 tokens (not replies, retweets or quote tweets).

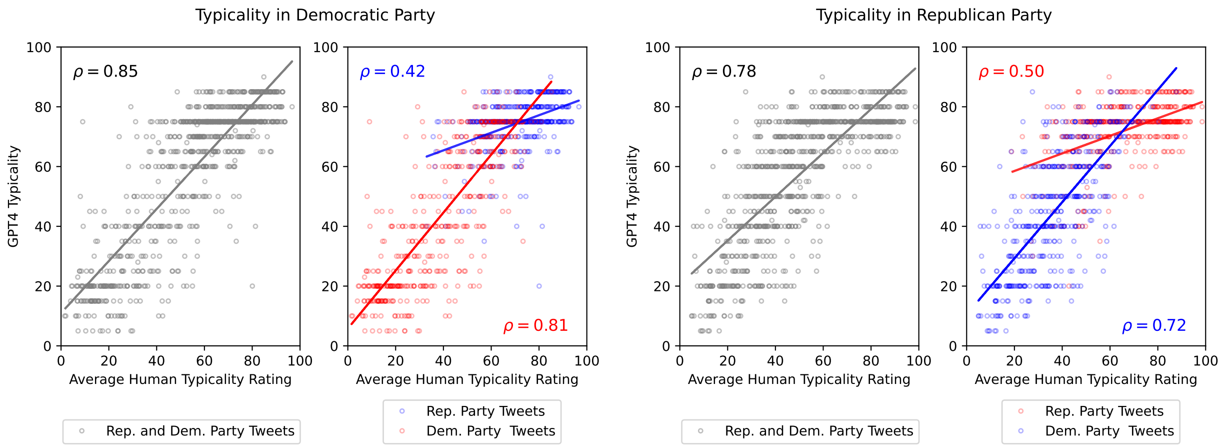


Figure 2: *GPT-4 queried about the typicality of the tweet in the focal party:* There exists a strong positive association between *GPT4 typicality* and *human typicality*. *LL panel:* Typicality in the Democratic Party, all tweets. *L panel:* The positive association holds for tweets published by Democratic and Republican Congress members. *R panel:* Typicality in the Republican Party, all tweets. *RR panel:* The positive association holds for tweets published by Democratic and Republican Congress members.

this conjecture, we transformed the model prediction for both books and tweets in binary classification outcomes, by assuming that if the typicality score produced by GPT-4 was higher than 50%, the book/tweet was predicted to be a member of the focal concept. Doing so yields a classification accuracy for book descriptions of about 90% (both for classification in the Mystery genre and in the Romance genre).[5] This is an excellent classification performance, given that the test sets were made of 50% of instances of the focal concept (so chance prediction would produce an accuracy of 50%). Applying the same approach to the classification of tweets in the political parties results in a classification accuracy close to 80%[6] (in a context in which chance predictions would also produce an accuracy of 50%). The fact that the model is more accurate at classifying book descriptions in literary genres than tweets in the parties of their authors suggests that the text of the tweet contains less diagnostic information than the text of the book description, on average.

3.3 Summary

Overall, the results obtained in the two empirical settings indicate that *GPT-4 Typicality* measure is good enough to substitute for human typicality ratings in empirical studies that require measuring the typicality of text documents in these settings.

[5] The ground truth was defined as follows: a book was considered to be a Mystery book if the “Mystery” label was the most frequent label among the labels given to that book.

[6] The ground truth was defined as follows: the category of the tweet was the party of the Congress member who published it.

Table 1: Typicality of books in the Mystery genre and Romance genres: Comparing the performance of *GPT-4 Typicality* with the previous state-of-the-art (*BERT typicality*) and other model-based typicality measures.

Typicality in the Mystery Genre		Correlations between model-based typicalities and human typicality ratings			Model Training					
Release year	Typicality Measure	Mystery and Not Mystery books	Mystery books	Not Mystery books	Training Sample	Fine-tuning the language representation	Training a probabilistic classifier	Language representation	Sensitivity to word order	Similarity between text document and concept
2023-03	*GPT-4 Typicality*	0.90	0.69	0.76	None	No	No	`gpt-4-0314`	Yes	Typicality rating
2018-10	*BERT Typicality*	0.87	0.67	0.63	680K	Yes	Yes	fine-tuned `bert-base-uncased`	Yes	Log(Cat. Prob.)
2018-10	BERT pre-trained / corr. with. prototype	0.07	0.03	0.07	None	No	No	`bert-base-uncased`	Yes	Cosine
2022-12	ada2 pre-trained / corr. with. prototype	0.76	0.52	0.61	None	No	No	`text-embedding-ada-002`	Yes	Cosine
2022-11	Pre-trained GPT-3.5 Completion	0.81	0.58	0.65	None	No	No	`text-davinci-003`	Yes	Typicality rating
2019	Pre-trained GPT-3 Completion	-0.06	0.00	-0.17	None	No	No	`text-curie-001`	Yes	Typicality rating

Typicality in the Romance Genre		Correlations between model-based typicalities and human typicality ratings			Model Training					
Release year	Typicality Measure	Romance and Not Romance books	Romance books	Not Romance books	Training Sample	Fine-tuning the language representation	Training a probabilistic classifier	Language representation	Sensitivity to word order	Similarity between text document and concept
2023-03	*GPT-4 Typicality*	0.91	0.66	0.83	None	No	No	`gpt-4-0314`	Yes	Typicality rating
2018-10	*BERT Typicality*	0.86	0.54	0.72	680K	Yes	Yes	fine-tuned `bert-base-uncased`	Yes	Log(Cat. Prob.)
2018-10	BERT pre-trained / corr. with. prototype	0.07	0.11	0.04	None	No	No	`bert-base-uncased`	Yes	Cosine
2022-12	ada2 pre-trained / corr. with. prototype	0.79	0.50	0.65	None	No	No	`text-embedding-ada-002`	Yes	Cosine
2022-11	Pre-trained GPT-3.5 Completion	0.87	0.54	0.76	None	No	No	`text-davinci-003`	Yes	Typicality rating
2019	Pre-trained GPT-3 Completion	-0.27	-0.09	-0.16	None	No	No	`text-curie-001`	Yes	Typicality rating

4 Comparing *GPT-4 Typicality* with the Previous State of the Art: *BERT Typicality*

Le Mens et al. (2023) proposed to measure the typicality of an object o in a concept c in terms of the probability that o is a c. This probability is estimated based on the predictions of a text classifier trained on labelled data (of a discrete nature). These authors hypothesized that model training enables the text classifier to become sensitive to features in a manner that mirrors the sensitivity of human *graded* typicality judgments to text features even though it is trained on *discrete* input data. The text classifier uses two distinct components:

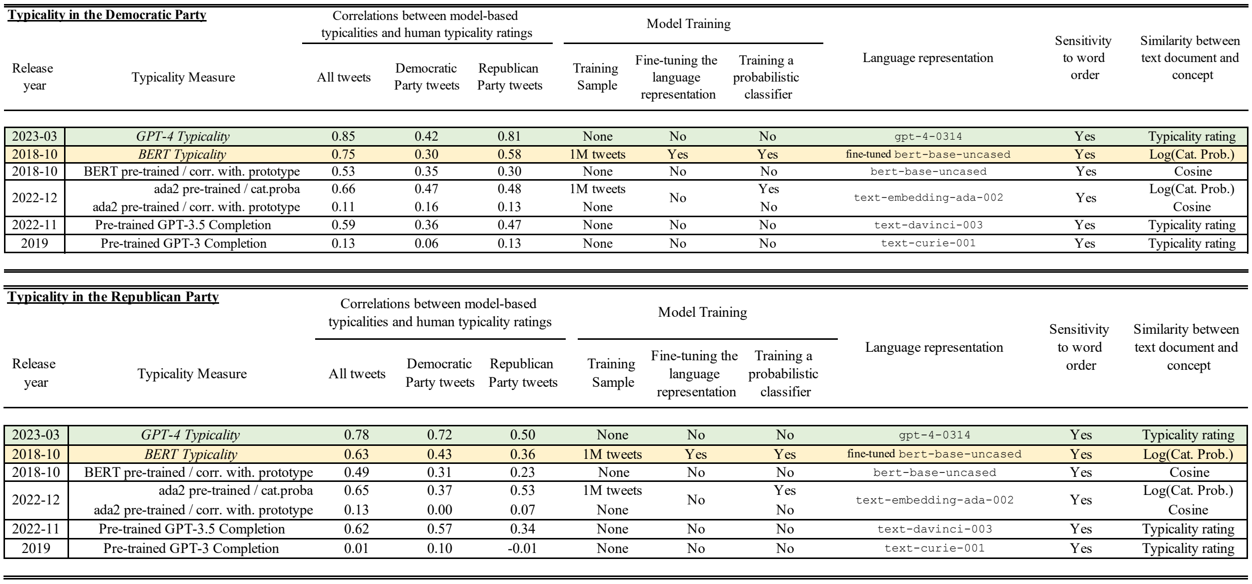
1. a *language model*. It is the *language representation* component that converts a text document into a vector $x = (x_1, \ldots, x_N)$, where N is the number of dimensions of the semantic space used to represent text documents.

2. a *categorization component* that produces categorization probabilities based on the position x of the text document. The typicality measure of a text document in concept c is obtained by taking the logarithm of the categorization probability in c.

4.1 Book Descriptions

In the comparative analysis reported in Le Mens et al. (2023), the model that achieved the best correspondence with *human typicality* used BERT as the language model. They fine-tuned the more than 100 million parameters of the BERT text classifier on a training set comprising 680 thousand book descriptions and their

Table 2: Typicality of tweets in the Democratic Party and the Republican Party: Comparing the performance of *GPT-4 Typicality* with the previous state-of-the-art (*BERT typicality*) and other model-based typicality measures.



Typicality in the Democratic Party		Correlations between model-based typicalities and human typicality ratings			Model Training					
Release year	Typicality Measure	All tweets	Democratic Party tweets	Republican Party tweets	Training Sample	Fine-tuning the language representation	Training a probabilistic classifier	Language representation	Sensitivity to word order	Similarity between text document and concept
2023-03	*GPT-4 Typicality*	0.85	0.42	0.81	None	No	No	`gpt-4-0314`	Yes	Typicality rating
2018-10	*BERT Typicality*	0.75	0.30	0.58	1M tweets	Yes	Yes	fine-tuned `bert-base-uncased`	Yes	Log(Cat. Prob.)
2018-10	BERT pre-trained / corr. with. prototype	0.53	0.35	0.30	None	No	No	`bert-base-uncased`	Yes	Cosine
2022-12	ada2 pre-trained / cat.proba	0.66	0.47	0.48	1M tweets	No	Yes	`text-embedding-ada-002`	Yes	Log(Cat. Prob.)
	ada2 pre-trained / corr. with. prototype	0.11	0.16	0.13	None		No			Cosine
2022-11	Pre-trained GPT-3.5 Completion	0.59	0.36	0.47	None	No	No	`text-davinci-003`	Yes	Typicality rating
2019	Pre-trained GPT-3 Completion	0.13	0.06	0.13	None	No	No	`text-curie-001`	Yes	Typicality rating

Typicality in the Republican Party		Correlations between model-based typicalities and human typicality ratings			Model Training					
Release year	Typicality Measure	All tweets	Democratic Party tweets	Republican Party tweets	Training Sample	Fine-tuning the language representation	Training a probabilistic classifier	Language representation	Sensitivity to word order	Similarity between text document and concept
2023-03	*GPT-4 Typicality*	0.78	0.72	0.50	None	No	No	`gpt-4-0314`	Yes	Typicality rating
2018-10	*BERT Typicality*	0.63	0.43	0.36	1M tweets	Yes	Yes	fine-tuned `bert-base-uncased`	Yes	Log(Cat. Prob.)
2018-10	BERT pre-trained / corr. with. prototype	0.49	0.31	0.23	None	No	No	`bert-base-uncased`	Yes	Cosine
2022-12	ada2 pre-trained / cat.proba	0.65	0.37	0.53	1M tweets	No	Yes	`text-embedding-ada-002`	Yes	Log(Cat. Prob.)
	ada2 pre-trained / corr. with. prototype	0.13	0.00	0.07	None		No			Cosine
2022-11	Pre-trained GPT-3.5 Completion	0.62	0.57	0.34	None	No	No	`text-davinci-003`	Yes	Typicality rating
2019	Pre-trained GPT-3 Completion	0.01	0.10	-0.01	None	No	No	`text-curie-001`	Yes	Typicality rating

genre labels. 'Fine-tuning' means that the parameters of the language models and of the categorization component are adjusted by minimizing a classification loss on the training data using deep learning techniques – this optimizes the language representation to the particular task (categorization) and data. They then applied the fine-tuned BERT text classifier to obtain predicted categorization probabilities in the Mystery genre. The "*BERT Typicality*" is the logarithm of the categorization probability.

This measure performed well. The correlation over all book descriptions and the within-category correlations are very high for both genres (see table 1, yellow panel). Yet, it falls short of matching the performance of *GPT-4 Typicality*. What is important is not so much the performance difference between the novel *GPT-4 Typicality* measure and the *BERT Typicality*, but the fact that obtaining the novel *GPT-4 Typicality* measure involves zero training on research data, whereas the *BERT Typicality* achieves its performance with a large training set of 680K text documents. It is noteworthy that training BERT on a decently sized training set of 10K book descriptions leads to a decrease of 10 percentage points of the intra-class correlation. Additionally, a typicality measure based on pre-trained BERT embeddings, thus involving no use of training data, performs very poorly (see table 1).

These results suggest a qualitative leap in the scope of application of model-based typicality measures. If similar results hold more generally, in other contexts (for other concepts, and other types of text documents), this could eliminate the need for training data.

4.2 Tweets

In this context, the probabilistic text classifier predicts the party of the Congress member who published it. The typicality of a tweet in the focal party is obtained by taking the logarithm of the categorization probability produced by the text classifier.

We assembled training data to fine-tune a BERT text classifier and train the other text classifiers analyzed in Le Mens et al. (2023) (these include classifiers that use Bag-of-Words text representations as well as GloVe word embeddings). The training data consist of the universe of tweets published by the serving members of 117^{th} and 118^{th} US Congresses as available when we downloaded them via the Twitter API (the first week of May 2023): about 1 million tweets published by 537 unique politicians (270 Democrats and 267 Republicans) between Jan 3, 2019 and Jan 3, 2023.

We estimated the performance of various model-based typicality measures discussed in Le Mens et al. (2023) using the code they made accessible with the paper. Results reported in the Supplementary Appendix (table 3) replicate the results obtained by these authors with the typicality of book descriptions in literary genres: the highest performing measure is the *BERT Typicality*. This measure performs well, as can be seen in table 2 (yellow panel). In contrast, relying on the pre-trained BERT representation without any model training leads to poor performance, as with the book descriptions. This confirms the conclusion of this previous paper that the BERT language model provides a text representation that achieves a high correspondence with human typicality judgments, provided it has been fine-tuned on the training data.

The most important finding is that *BERT typicality* fails to match the performance of *GPT-4 Typicality* despite the fact that the BERT LLM was fine-tuned on a large amount of training data (1 million tweets by US politicians). This result reinforces the conclusions drawn from the results obtained with the Book descriptions.

5 Comparing *GPT-4 Typicality* with Measures Based on Other Recent LLMs

We use other LLMs of the GPT family of models to obtain typicality measures.

5.1 GPT Text Embeddings

A frequently used approach to measuring typicality defines the concept prototype as the average position of instances of this concept in semantic space, and takes the cosine similarity between the position of a text document and the prototype as its typicality (e.g., Durand and Kremp, 2016; Pontikes and Hannan, 2014; Smith, 2011). Text representations consist of Bag-of-Words TF-IDF ("Term Frequency–Inverse Document Frequency," Luhn, 1957) or pre-trained word embeddings, generally seen as a cutting-edge method in social science (here implemented with GloVe embeddings (Pennington, Socher, and Manning, 2014)). Le Mens et al. (2023) reported the performance of measures based on these representations on the book description data. We replicated these analyses with the tweet data (see Supplementary Appendix). Results reported in Le Mens et al. (2023) (book descriptions) and in table 3 in the Supplementary Appendix (tweets) show that, with these two representations, the correspondence of model-based typicality measures and human judgment is *at best moderate ($\rho < .3$)*.

We applied the approach described in the previous paragraph to a much more recent embedding model: a text embedding model of the same generation as GPT-3 (the LLM that powered ChatGPT when it was first released): text-embedding-ada-002, released December 15, 2022. Just like BERT, this model transforms text documents in vectors. Results reported in table 1 show that the performance of the cosine similarity measure obtained with this LLM, to measure the typicality of book descriptions in literary genres, is much better than that obtained with the cosine similarity and simpler representations like Bag-of-Words TF-IDF and GloVe Embeddings, or with the pre-trained BERT model. This indicates that the semantic space at the core of the model reflects human judgments of genre typicality (in the Mystery and Romance genres) much better than do the earlier generations of text representation.

At the same time, the performance of this measure with the tweet data is very poor (table 2). This implies that the semantic space at the core of this model does not have a representation of the concepts Democratic Party and Republican Party that matches that of the participants in our survey. This finding should serve as a warning call to researchers who feel tempted to use recent LLMs to construct similarity measures without validating them on empirical data.

To assess whether this text representation could nevertheless be leveraged to produce a typicality measure that corresponds well with the *Human Typicality*, we applied the method advocated by Le Mens et al. (2023): we trained a text classifier based on this representation using the 1 million tweet training dataset.[7] The resulting performance is fairly good. This indicates that model training identified the dimensions in embedding space that matter for typicality judgments. Yet, the performance of this measure remains quite far from that of *GPT-4 Typicality* even though it involved training a model on a large training data set.

5.2 Text Completion with Pre-Trained GPT-3 and GPT-3.5

We adapted the prompt we used with GPT-4 for use with models of the previous generation: GPT-3 and GPT-3.5. The models we consider have not been specifically designed to work as chatbots, but instead to complete text that has been started by the user. Therefore we adapted the prompt by adding the beginning of the response at the end of the prompt. Assuming for this example that we aim to measure the typicality of a tweet in the Republican Party, we used:

> Here's a tweet written by a member of the US Congress: 'TWEET TEXT'. How typical is this tweet of the Republican Party? Provide your response as a score between 0 and 100 where 0 means 'Not typical at all' and 100 means 'Extremely typical'. The tweet typicality score is:

We used a model of the GPT-3 family (text-curie-001, released in 2019) and a model of the GPT-3.5 family (text-davinci-003, released in November 2022).[8] Results reported in tables 1 and 2 indicate that,

[7] We used a random-forest text classifier.

[8] We also intended to use the two most capable original GPT-3 models (curie and davinci) but they did not respond to completion requests with numbers and thus the outputs were unusable.

whereas the typicality ratings provided by the recent GPT-3.5 model have a good level of correspondence with *Human Typicality* (matching that of the *BERT Typicality*, this is not the case for the typicality ratings provided by the GPT-3 model. To the contrary, these have an extremely poor correspondence with human judgment. The correlation is even (slightly) negative in the case of book descriptions. And with the tweets, more than 90% of the responses were a typicality score of "50."

This finding demonstrates that researchers should apply caution when using LLMs to construct typicality measures. In some settings, they fail to match human judgments and relying on such measures in downstream analyses would lead to results of questionable empirical validity. Empirical validation is necessary to avoid falling into such pitfalls.

Discussion and Conclusion

In this paper, we explored how recent advances in LLMs might help social scientists measure the typicality of text documents in concepts. We focused on the most recent LLM powering ChatGPT: GPT-4 (released in March 2023). We tested it in two empirical settings: the typicality of books in literary genres and the typicality of US-Congress members' tweets in the two major US political parties. We measured the performance of these new typicality measures in terms of their correspondence with the average typicality ratings produced by human judges. Our findings demonstrate that the measures derived from querying GPT-4 achieve state-of-the-art performance.

The groundbreaking nature of these findings lies in the fact that this performance is achieved without training the LLM on the research data. To be clear, these findings represent a breakthrough that essentially eliminates the trade-off between size of training data and measurement accuracy, overcoming the limitations of current approaches to measuring typicality with NLP based techniques. This makes the practical applicability of model-based typicality measures much broader, because assembling a training set frequently requires employing research assistants to manually label thousands or tens of thousands of text documents even when pre-trained LLMs such as BERT are used (e.g., Schöll, Gallego, and Le Mens, 2023; Wahman, Frantzeskakis, and Yildirim, 2021). Therefore, the new approach dramatically reduces the financial and logistical costs of obtaining model-based typicality measures.

Our findings suggest that the semantic space constructed by GPT-4 closely matches that of human judges. This opens the door to constructing other measures such as ambiguity, diversity, and polarization.

Additional benefits of the new method are practical. Measuring typicality with GPT-4 is easier than with alternative methods because it requires less programming skills. A significant advantage of GPT-4 is that users can communicate with it using natural language. Although using the API and scaling up the process requires some programming, this is minimal. Moreover, analyzing text typicality with GPT-4 can easily be done similarly across various settings. As demonstrated in this paper, a nearly identical prompt can be effectively applied to query the typicality of both books and tweets in concepts that have little to do with each other. We believe this will enable researchers to conduct studies across a wide variety of settings.

A frequently heard concern regarding the outputs of LLMs is that they lack interpretability. This concern is valid in some use cases, but we do not think it is a damning issue in many other use cases. The reason is that the interpretability concerns that presumably affect the outputs of LLMs are of a similar nature than those that affect human coding (though we often demand more of ‘machines’ than of humans). There is much evidence in psychology that humans are quite poor at explaining how they produce their own judgments (e.g., Nisbett and Wilson, 1977). This implies, for example, that asking research assistants (or experts alike) to explain why they rated the typicality of text documents in a concept the way they did is unlikely to provide much insight about the process that generated their judgments. In other words, human coding is unlikely to be more interpretable than the outputs of LLMs. In settings in which researchers are satisfied with human coding, they should probably be equally satisfied with LLM-based measures, *provided they have collected validation data from human judges and verified that the model-based measures have a sufficiently high correspondence with the measures provided by human judges on the validation data.* This is not to say that there is no trade-off between different model-based measures, such as those that rely on word frequencies (like Topic Models), and the LLM-based measures we propose in this article. The trade-off involves choosing between using NLP methods that highly correspond with human judgments but lack interpretability in a way that also affects human judgment, or using methods that provide measures with a much lower correspondence to human judgment (and are thus noisy estimates of what researchers want to approximate) but are interpretable (e.g., they allow the identification of diagnostic words). One way to resolve this tension is to rely on a combination of methods for the same task to achieve both high measurement accuracy and some level of interpretability. However, such an approach remains imperfect, because it will not provide much insight into the differences in predictions between a LLM-based measure and a measure based on word frequency.

The results presented in this paper align with the emerging sentiment about the advantages of using the newest LLMs. At the same time, we would like to conclude with a few words of caution regarding the use of LLMs as measurement devices for social science. Current and future LLMs are limited by the pre-training process that lead to their construction. This process allows the model to learn semantic associations present in text data written by humans. The models are thus bound to reproduce these associations in the semantic space they construct and in the text they produce as a response to queries (Kearns and Roth, 2019). If the pre-training data contains erroneous semantic associations or associations that correspond to prejudice against particular social groups, ideas, or ideology, these will influence the measures produced by LLMs. It is important to assess the extent to which the situation necessitates the design and application of systematic validation methods. As we saw in the comparative analysis reported above, even quite recent LLMs can produce measures that have a poor correspondence with human judgment. This also suggests that pre-trained LLMs might not perform as desired in domains in which the input data exhibit systematic judgment and evaluative biases. In such domains, it may be necessary to employ model fine-tuning[9] to correct these biases in the measurements derived from LLM outputs. Identifying these domains should be high on the priority list of researchers interested in using LLMs as measurement tools for social science research.

[9] As an illustration note that the performance of the fine-tuned BERT model (at the core of the *BERT Typicality*) is much better than that of the pre-trained BERT model (tables 1 and 2). In ancillary analyses, we also obtained excellent results by fine-tuning curie, a GPT-3 model that could not provide useable outputs without fine-tuning.

References

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Technical report.

Durand, Rodolphe and Pierre-Antoine Kremp. 2016. "Classical Deviation: Organizational and Individual Status as Antecedents of Conformity." *Academy of Management Journal* 59:65–89.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature* 57:535–74.

Hannan, Michael T. 2010. "Partiality of Memberships in Categories and Audiences." *Annual Review of Sociology* 36:159–181.

Hannan, Michael T., Gaël Le Mens, Greta Hsu, Balázs Kovács, Giacomo Negro, Lászlo Pólos, Elizabeth G. Pontikes, and Amanda J. Sharkey. 2019. *Concepts and Categories: Foundations for Sociological and Cultural Analysis*. New York: Columbia University Press.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2007. *Logics of Organization Theory: Audiences, Codes, and Ecologies*. Princeton, N.J.: Princeton University Press.

Hsu, Greta. 2006. "Jacks of All Trades and Masters of None: Audiences' Reactions to Spanning Genres in Feature Film Production." *Administrative Science Quarterly* 51:420–450.

Hsu, Greta, Michael T. Hannan, and Özgeçan Koçak. 2009. "Multiple Category Memberships in Markets: An Integrative Theory and Two Empirical Tests." *American Sociological Review* 74:150–169.

Kearns, Michael and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kovács, Balázs and Michael T. Hannan. 2010. "The Consequences of Category Spanning Depend on Contrast." *Research in the Sociology of Organizations* 31:175–201.

Kovács, Balázs and Michael T. Hannan. 2015. "Conceptual Spaces and the Consequences of Category Spanning." *Sociological Science* 2:252–286.

Le Mens, Gaël, Balázs Kovács, Michael T Hannan, and Guillem Pros Rius. 2023. "Using machine learning to uncover the semantics of concepts: how well do typicality measures extracted from a BERT text classifier match human judgments of genre typicality?" *Sociological Science. 2023 March; 10: 82-117* .

Luhn, Hans Peter. 1957. "A statistical approach to mechanized encoding and searching of literary information." *IBM Journal of research and development* 1:309–317.

Nisbett, Richard E and Timothy D Wilson. 1977. "Telling more than we can know: Verbal reports on mental processes." *Psychological review* 84:231.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pontikes, Elizabeth G. and Michael T. Hannan. 2014. “An Ecology of Social Categories.” *Sociological Science* 1:311–343.

Porac, Joseph F., Howard Thomas, Fiona Wilson, Douglas Paton, and Alaina Kanfer. 1995. “Rivalry and The Industry Model of Scottish Knitwear Producers.” *Administrative Science Quarterly* 40:203–227.

Rosch, Eleanor H. 1973. “On the Internal Structure of Perceptual and Semantic Categories.” In *Cognitive Development and the Acquisition of Language*, edited by T. E. Moore, pp. 111–144. New York: Academic Press.

Schöll, Nikolas, Aina Gallego, and Gaël Le Mens. 2023. “How Politicians Learn from Citizens’ Feedback: The Case of Gender on Twitter.” *American Journal of Political Science, accepted* .

Smith, Edward Bishop. 2011. “Identities as Lenses: How Organizational Identity Affects Audiences’ Evaluation of Organizational Performance.” *Administrative Science Quarterly* 56:61–94.

Wahman, Michael, Nikolaos Frantzeskakis, and Tevfik Murat Yildirim. 2021. “From thin to thick representation: how a female president shapes female parliamentary behavior.” *American Political Science Review* 115:360–378.

Zuckerman, Ezra W. 1999. “The Categorical Imperative: Securities Analysts and the Legitimacy Discount.” *American Journal of Sociology* 104:1398–1438.

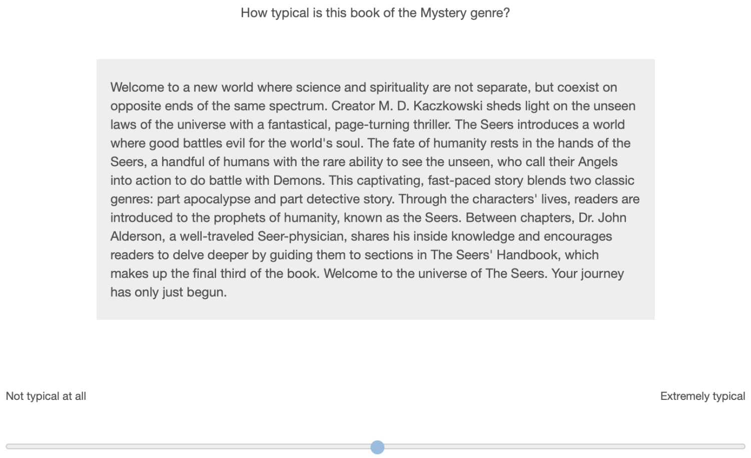
Supplementary Appendix

A Methodological Details

A.1 Survey of Book Typicality Ratings

The survey was conducted by Le Mens et al. (2023), here we replicate their survey prompt to illustrate the Goodreads book descriptions and also the prompt used in the survey.

Figure 3: Example of typicality rating display.



How typical is this book of the Mystery genre?

Welcome to a new world where science and spirituality are not separate, but coexist on opposite ends of the same spectrum. Creator M. D. Kaczkowski sheds light on the unseen laws of the universe with a fantastical, page-turning thriller. The Seers introduces a world where good battles evil for the world's soul. The fate of humanity rests in the hands of the Seers, a handful of humans with the rare ability to see the unseen, who call their Angels into action to do battle with Demons. This captivating, fast-paced story blends two classic genres: part apocalypse and part detective story. Through the characters' lives, readers are introduced to the prophets of humanity, known as the Seers. Between chapters, Dr. John Alderson, a well-traveled Seer-physician, shares his inside knowledge and encourages readers to delve deeper by guiding them to sections in The Seers' Handbook, which makes up the final third of the book. Welcome to the universe of The Seers. Your journey has only just begun.

A.2 Survey of Tweet Typicality Ratings

The online survey was administered in the second week of May 2023.

903 Prolific participants (U.S. residents, average age of 41 years, 57% male, 41% female, 2% others) were randomly assigned to one of two "party conditions." Half of the participants provided typicality ratings of tweets in the Democratic Party. The other half did so for the Republican Party.

The 900 tweets in the test set were rated by participants in the two conditions. We constructed 30 sets of 30 tweets, composed of 15 tweets from each party. Participants were randomly allocated to one of these sets. We obtained between 13 and 18 typicality ratings for each tweet (in each party), with an average of 15 ratings per tweet.

To provide their typicality ratings, participants responded to the question (assuming for this example that the focal concept is the Democratic Party) "How typical is this tweet to the Democratic Party?" using a 0 to 100 slider (centered at 50 when the page appears on the screen). Importantly, participants were not provided with any information about the author of the tweet — just the text of the tweet. Therefore, they were not informed of whether the tweet was written by a Democrat or a Republican.

To ensure response quality, we screened Prolific participants based on the following criteria:

- Fluent languages includes English;

- Approval Rate: Minimum 95%;
- Country of Residence: US;
- Political Spectrum (US): they should have provided a response to this question (the response could be anyone of {Conservative, Moderate, Liberal, Other}),
- U.S. Political Affiliation: they should have provided a response to this question (the response could be anyone of {Democrat, Republican, Independent, Other, None}).

The last two items were included to increase the likelihood that participants have some knowledge of US politics.

After providing informed consent, participants were asked

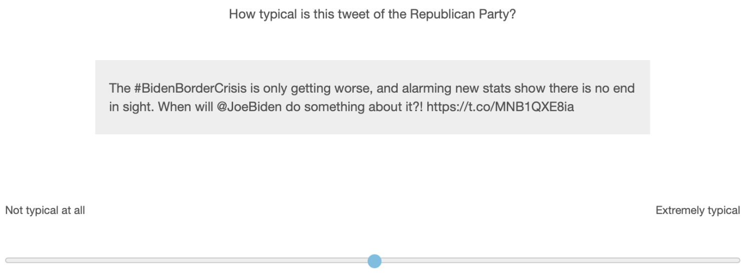
> Please take a moment to write about your expectations regarding tweets by politicians of the [Democratic Party / Republican Party]. You could write about possible topics addressed by such tweets, and or opinions you expect the authors to have about these topics (min. response length: 100 characters).

On the next page, they were provided with short instructions about the typicality rating task.

> For each of the 30 tweets, you will be asked the following question:
>
> How typical is this tweet of the [Democratic Party / Republican Party]?
>
> You will report your response using a continuous slider that goes from 'Not typical at all' (0) to 'Extremely typical' (100).
>
> There is no right or wrong answer. We are interested in your subjective opinion.
>
> Note that the slider will appear on the screen 6 seconds after the text of the tweet.

Then they looped through the 50 tweets and provided their typicality ratings for each of them. See Figure 4 for an example of typicality rating display.

Figure 4: Example of typicality rating display.



How typical is this tweet of the Republican Party?

The #BidenBorderCrisis is only getting worse, and alarming new stats show there is no end in sight. When will @JoeBiden do something about it?! https://t.co/MNB1QXE8ia

Finally, the study concluded with a short demographic survey, and some questions about their political orientation, their political opinions and their frequency of use of Twitter.

B Additional Results: Performance of Model-Based Typicality Measures Proposed in the Prior Literature

In the body of the paper, we put the focus on comparing the performance of the *GPT-4 Typicality* with the *BERT Typicality*. This is justified in the context of the book description context, because prior work found that the *BERT Typicality* had the best correspondence with *Human Typicality*. Here we report analyses that show that the same applies to the typicality of tweets in the two main US parties. Table 3 reports a replication of the analyses performed on the book description data by Le Mens et al. (2023).[10]

What is important to for the present paper is that the highest performing measure is the *BERT Typicality*. Overall, the performance ordering of the previous measures is extremely similar in this setting to the ordering that was obtained with the book description data and therefore we do not discuss it further here (we direct interested readers to the result section of Le Mens et al. (2023).

Table 3: Typicality of tweets in the Democratic Party and the Republican Party: Performance of previous model-based typicality measures. (Replication of the comparative analysis in Le Mens et al. (2023) with the tweet data).

Typicality in the Democratic Party

			Correlations between model-based typicalities and human typicality ratings			Model Training					
		Typicality Measure	All tweets	Democratic Party tweets	Republican Party tweets	Training Sample	Fine-tuning the language representation	Training a probabilistic classifier	Language representation	Sensitivity to word order	Similarity between text document and concept
LLM	BERT	BERT fine-tuned / cat. proba. (*BERT Typicality*)	0.75	0.30	0.58	1M tweets	Yes	Yes	fine-tuned BERT	Yes	Log(Cat. Prob.)
		BERT fine-tuned / corr. with. prototype	0.75	0.31	0.55	1M tweets	Yes	Yes	fine-tuned BERT	Yes	Cosine
		BERT pre-trained / cat. proba.	0.53	0.35	0.30	1M tweets	No	Yes	pre-trained BERT	Yes	Log(Cat. Prob.)
		BERT pre-trained / corr. with. prototype	0.07	0.13	0.11	None	No	No	pre-trained BERT	Yes	Cosine
Methods not sensitive to word order	GloVe Word Embeddings	GloVe fine-tuned / cat. proba.	0.61	0.31	0.36	1M tweets	Yes	Yes	fine-tuned GloVe	No	Log(Cat. Prob.)
		GloVe fine-tuned / corr. with. prototype	0.50	0.26	0.27	1M tweets	Yes	Yes	fine-tuned GloVe	No	Cosine
		GloVe pre-trained / cat. proba.	0.49	0.22	0.32	1M tweets	No	Yes	pre-trained GloVe	No	Log(Cat. Prob.)
		GloVe pre-trained / corr. with. prototype	0.13	0.20	0.01	None	No	No	pre-trained GloVe	No	Cosine
	Word Frequencies	Word Frequencies / cat. proba.	0.60	0.30	0.45	1M tweets	No	Yes	BoW Term Frequencies	No	Log(Cat. Prob.)
		TF-IDF / cat. proba	0.62	0.36	0.47	1M tweets	No	Yes	BoW TF-IDF	No	Log(Cat. Prob.)
		TF-IDF / corr. with. prototype	0.22	0.09	0.21	None	No	No	BoW TF-IDF	No	Cosine

Typicality in the Republican Party

			Correlations between model-based typicalities and human typicality ratings			Model Training					
		Typicality Measure	All tweets	Democratic Party tweets	Republican Party tweets	Training Sample	Fine-tuning the language representation	Training a probabilistic classifier	Language representation	Sensitivity to word order	Similarity between text document and concept
LLM	BERT	BERT fine-tuned / cat. proba. (*BERT Typicality*)	0.63	0.43	0.36	1M tweets	Yes	Yes	fine-tuned BERT	Yes	Log(Cat. Prob.)
		BERT fine-tuned / corr. with. prototype	0.63	0.40	0.26	1M tweets	Yes	Yes	fine-tuned BERT	Yes	Cosine
		BERT pre-trained / cat. proba.	0.49	0.31	0.23	1M tweets	No	Yes	pre-trained BERT	Yes	Log(Cat. Prob.)
		BERT pre-trained / corr. with. prototype	0.05	-0.01	0.01	None	No	No	pre-trained BERT	Yes	Cosine
Methods not sensitive to word order	GloVe Word Embeddings	GloVe fine-tuned / cat. proba.	0.56	0.35	0.33	1M tweets	Yes	Yes	fine-tuned GloVe	No	Log(Cat. Prob.)
		GloVe fine-tuned / corr. with. prototype	0.37	0.27	0.24	1M tweets	Yes	Yes	fine-tuned GloVe	No	Cosine
		GloVe pre-trained / cat. proba.	0.48	0.19	0.36	1M tweets	No	Yes	pre-trained GloVe	No	Log(Cat. Prob.)
		GloVe pre-trained / corr. with. prototype	0.02	0.01	0.19	None	No	No	pre-trained GloVe	No	Cosine
	Word Frequencies	Word Frequencies / cat. proba.	0.53	0.37	0.32	1M tweets	No	Yes	BoW Term Frequencies	No	Log(Cat. Prob.)
		TF-IDF / cat. proba	0.60	0.33	0.49	1M tweets	No	Yes	BoW TF-IDF	No	Log(Cat. Prob.)
		TF-IDF / corr. with. prototype	0.04	0.09	0.09	None	No	No	BoW TF-IDF	No	Cosine

[10] The measures based on label proportions do not have any equivalent to this context; the same applies to the measure based on training a text classifier with 36 target categories.