Uncovering Deep-Rooted Cultural Differences (UNCOVER)

Aleksey Panasyuk[1], Bryan Li[2], Chris Callison-Burch[2]

[1]Information Fusion Technology Branch, Air Force Research Lab, Rome NY 13440, USA
[2]University of Pennsylvania, Philadelphia PA 19104, USA

ABSTRACT

This study delves into the interconnected realms of Debates, Fake News, and Propaganda, with an emphasis on discerning prominent ideological underpinnings distinguishing Russian from English authors. Leveraging the advanced capabilities of Large Language Models (LLMs), particularly GPT-4, we process and analyze a large corpus of over 80,000 Wikipedia articles to unearth significant insights. Despite the inherent linguistic distinctions between Russian and English texts, our research highlights the adeptness of LLMs in bridging these variances. Our approach, includes translation, question generation and answering, along with emotional analysis, to probe the gathered information. A ranking metric based on the emotional content is used to assess the impact of our approach. Furthermore, our research identifies important limitations within existing data resources for propaganda identification. To address these challenges and foster future research, we present a curated synthetic dataset designed to encompass a diverse spectrum of topics and achieve balance across various propaganda types.

Keywords: Propaganda Detection, Emotional Text Extraction, Cross lingual Question Answering, Large Language Models (LLMs), Synthetic Propaganda Generation

1. INTRODUCTION

In efforts to reshape the global geopolitical arena to their advantage, countries like Russia and China have significantly invested in state-sponsored media initiatives, such as Russia's RT and China's CGTN [1]. Relevant to our investigation, it is believed that Russian propaganda played a crucial role in pivotal events like the 2016 US Presidential Election [2] and the Russian invasion of Ukraine [3].

Combating such influence campaigns necessitates a multi-faceted approach across the interconnected realms of Debates, Fake News, and Propaganda. Although these areas are closely related, each presents unique features and seemingly necessitates the development of specialized analytical models tailored to their specific needs. However, LLMs hold the promise of a unified approach that could span these research fields.

This paper, as an initial inquiry into understanding ideological differences across cultures, studies paired Wikipedia articles, on the same topic but written in either Russian or English. We systematically translate each article pair into a mutual language (Russian or English), employing advanced question-answering techniques to identify emotionally charged texts within this shared language. Subsequently, we analyze and compare the emotional intensity of these texts to categorize and rank the articles accordingly. Key to our methodology is the usage of LLMs, especially GPT-4, to process data and generate insights at scale.

Here is an overview of the sections to follow: Sec 2 presents related research; Sec 3 shows our high-level approach, Sec 4 presents our Wikipedia dataset; Sec 5 expands Wikipedia data via translation, Sec 6 provides LLM zero-shot performance on an existing persuasion dataset; Sec 7 discusses the creation of a question repository and its statistical analysis; Sec 8 aggregates emotional texts for ranking; Sec 9 shows example results between the two Wikipedia; and finally, Sec 10 gives a summary and future steps. The Appendix contains additional experimentation details and information on our synthetically generated propaganda dataset*.

Send correspondence to Aleksey Panasyuk: E-mail: Aleksey.Panasyuk@us.af.mil

*for code and data see: https://github.com/apanasyu/UNCOVER_SPIE

2. RELATED RESEARCH

2.1 Cross-lingual Question Answering

Cross-lingual question answering (QA) is the task of producing an answer, given a question in one language and a context in another. Machine systems implementing this can empower users to ask questions of texts in multiple languages they may not understand. As such, it serves as a useful framing for our work, in that we can ask targeted questions of articles in different languages, then compare the responses.

Wikipedia provides a useful platform for developing cross-lingual Question Answering (QA) systems and datasets, given its wide language availability and range of covered subjects. The MegaWika dataset [4], comprising over 120 million English question-answer pairs, with contexts across 50+ languages. Li et al. [5] release 1m+ high-quality, synthetic cross-lingual QA pairs, with contexts and questions aligned across 4 languages.

These datasets consider fact-based inquiries, where questions are grounded in specific context. In contrast, our work seeks to generate high-level questions which can be asked of any context to evaluates its argumentative techniques. Two articles might convey similar facts, with slight differences in wording and organization leading readers to different conclusions. In other words, we are scrutinizing how information is presented, rather than the information itself.

2.2 Propaganda Detection

The most effective forms of propaganda can manifest subtly and be difficult to detect; for example, by intertwining authentic elements with subtly manipulated components [6, 7, 8]. The methods of propaganda are diverse and multifaceted.

Early studies center on propaganda detection at the entire article-level. Rashkin et al. [9], ranked news articles based on four categories: propaganda, trusted, hoax, or satire. However, recent focus have shifted to more fine-grained analysis (the focus of our work), in both identifying specific text spans within a large document, and for each categorizing the type of propaganda.

The NLP4IF-2019 workshop released 2 shared tasks: multi-class fragment-level classification of propaganda techniques (FLC) and binary sentence-level classification (SLC) [10]. For SLC, a submission by Li et al. [11] utilized Logistic Regression with various features such as TF-IDF, BERT vector, sentence length, readability grade level, emotion, LIWC, and emphatic content; they achieved 66.2 F1. However, FLC proved to be difficult, the top scoring team only achieving 24.9 F1.

The next iteration of the workshop, SemEval-2020 Task 11 [12], revised the FLC task by breaking it down into span identification and a revised 14-way propaganda technique classification. The top performing teams had an F1 of 51.55 for span identification and 62.07 for propaganda technique classification.

The latest iteration at SemEval-2023 Task 3 [13] introduced a multilingual component across eight languages, including Russian, and was conducted at paragraph level. The task provided a representative dataset of contemporary global issues, such as the COVID-19 pandemic, abortion-related legislation, and the Russo-Ukrainian war.

2.3 Synthetic Propaganda Generation

As an alternative to the costly process of labeling data for propaganda, researchers have sought to leverage the latest developments in generative AI and Natural Language Processing (NLP) for automatically generating propaganda. An earlier effort by Zellers et al. [14] showed that from a simple misleading headline, a GPT-2 like model can generate English articles good enough to fool humans.

As discussed earlier, while articles with full-blown propaganda premises exist, it is often more effective to instead subtly embed propaganda in otherwise seemingly factual articles. Researchers are thus looking for more advanced generative methods that can likewise subtly apply various propaganda techniques. Huang et al. [6], developed an approach that introduces loaded language and appeal-to-authority techniques into legitimate articles. Their proposed methodology assigns an 'importance score' to every sentence in the original piece based on its relevance to a generated summary. The sentence with the highest score becomes the insertion point for the propaganda element. For the appeal-to-authority technique, they insert false quotations from experts, which are curated from Wikidata in various fields, and generated using a BART model following a template.

One limitation of such generative techniques is their dependence on clearly defined template structures and lexicons. This may restrict their applicability, as not all propaganda techniques can be expressed using such a template-based approach.

2.4 Communication Practices and Argument Mining

Argument mining is the task of extracting and identifying argumentative techniques in texts. In a survey, Lawrence et al. discuss challenges such as inconsistent data annotation across different domains, conceptual differences in argument understanding, and lack of universally accessible algorithms [15].

Farzam et al. [16] highlighted the shared semantic and logical structures among different argument mining tasks, suggesting a holistic approach to extracting argumentative techniques. Sourati et al. [17] proposes four main logical fallacy categories, including relevance, defective induction, presumption, and ambiguity.

Durmus et al. [18], analyzed 77,655 debates to understand the success factors for debaters, including conduct, spelling and grammar accuracy, persuasiveness of arguments, and reliability of sources. Habernal et al. [19] focused on developing a methodology for systemically reconstructing warrants. This task, essentially connecting the claim and premises of an argument, remains challenging due to the complexity of reconstructing world knowledge and reasoning patterns.

Jin et al. [20] introduces the task of logical fallacy detection and presents new datasets for identifying logical errors, particularly in the context of climate change debates. Findings indicate that a structure-aware classifier performs significantly better at this task than existing pre-trained language models, suggesting promising future work in enhancing reasoning abilities of NLP models to tackle misinformation.

Pauli et al. [21] proposes a new taxonomy for computational persuasion that can detect the misuse of rhetorical appeals. The research, which demonstrates the application of this taxonomy, finds that such misuse is more frequently associated with misinformation than with truthful contexts.

In sum, related research has highlighted a need for realistic labeled data containing propaganda and the integration of diverse features across debate, fake news, and propaganda datasets. These features include gaps in argumentation, fact-checking, and emotional influence.

3. OVERVIEW OF THE APPROACH

Using Wikipedia, our research aims to identify areas of disagreement (and agreement) between English and Russian speaking authors involving propaganda and persuasion techniques. Adding to the current body of knowledge, our research leverages the latest LLMs to address the associated challenges at scale. Here is a high-level overview of our method:

- While Wikipedia is expansive across Russian and English (1m+ paired articles), we narrowed down to a targeted dataset (22046 pairs) which considers substantially-long articles that reference Russian news.
- We study 4 datasets total: the Russian (RU) and English (EN) original articles, and those created through LLM translation: Russian to English (RU2EN) and English to Russian (EN2RU).
- We explore an existing propaganda detection dataset, and as a baseline develop a zero-shot prompt using GPT-4. From the initial exploration, we take a closer look at the baseline's error cases, then automatically generate a range of specific questions that can function as features for an improved classifier. By performing a statistical analysis, we identify the most significant questions, measured against the existing gold annotations.
- Given the most significant questions, we apply them on the four dataset versions, enabling us to gauge the total amount of emotionally charged content. We then rank and visualize the articles.

- Finally, we propose a bilingual synthetic propaganda dataset that prioritizes nation-state propaganda, incorporating examples across numerous politically relevant categories. This dataset addresses the imbalance present in existing resources, providing around 10K examples for each technique.

We anticipated that the articles which contain the largest discrepancies across the paired articles will concern politically-sensitive issues such as historical events and national figures. This is due to nations emphasizing their achievements while minimizing their mishaps–a conventional strategy to gain international leverage. Conversely, we expect neutral topics, such as biology, would have a higher consensus across the two cultures, consequently decreasing the likelihood of propaganda.

4. WIKIPEDIA DATASET: PAIRED ARTICLES IN RUSSIAN AND ENGLISH

WikiData supplies data about specific Wikipedia language links (downloaded as a .bz2 file). We employ these links to center our attention on Wikipedia entities that exist in both English and Russian languages. We target all records featuring 'sitelinks.ruwiki' and 'sitelinks.enwiki'. A total of 1,308,962 English Russian pairs are retrieved.

Wikipedia offers a SQL file that links Wikipedia's and WikiData's ids. We use the SQL mapping to identify the Wikipedia ID for English and Russian articles. For instance, for Belgium, the WikiData ID is Q31, the English Wikipedia ID is 3343, and the Russian Wikipedia ID is 1130.

Russian and English Wikipedia article IDs are used to locate relevant articles within each Wikipedia. Unlike WikiData, Wikipedia is in XML. The SAX XML handler is used to iterate over records matching the Wikipedia ID. We use the mwparserfromhell library to load the associated text for each Wikipedia article and extract the references. All references that start in (i) http://, (ii) https://, (iii) ftp:// and (iv) //www. are processed (9,897,517 vs. 16,650,755 links for Russian and English). Regular expressions (regex) used to focus on site and top level domain name. Unsurprisingly, for the Russian Wikipedia, a larger percent of overall sites are '.ru' (14.86% of all Russian references vs. only 1.65% in English).

Table 1: Top sites with the RU domain as referenced by Russian and English Wikipedia.

RU Wikipedia RU Sites	RU Count	EN Wikipedia RU Sites	EN Count
books.google.ru	62486	mapdata.ru	23075
lenta.ru	28036	allroutes.ru	18277
kommersant.ru	27764	bashstat.gks.ru	4364
ria.ru	26012	kommersant.ru	4148
sports.ru	22665	tass.ru	3518
tass.ru	19584	ria.ru	3286
rg.ru	14661	rg.ru	2364

Table 1 presents the most frequently referenced '.ru' domain sites. Particular attention given to government-controlled news websites that could potentially disseminate propaganda. The top 100 websites with '.ru' domains in Russian Wikipedia processed, isolating all news-related sites and excluding sites devoted to sports, library resources, scientific discourse, and other non-political content. The resultant list comprises 30 news sites outlined under Filter 3. All of the filters that are used to refine our dataset are listed below (filter is followed by number of entries remaining in our dataset after filter applied):

- Filter 1 (1,008,298): remove entries whose WikiData labels that start with 'category:' (258667 instances), 'template:' (34789), 'wikipedia:' (1493), 'portal:' (681), 'module:' (327).
- Filter 2 (947,834): remove entries whose WikiData descriptions start with 'wikimedia'. Example top entries and corresponding entry counts: wikimedia category: 256585, wikimedia disambiguation page: 44562, wikimedia template: 33445, wikimedia list article: 14643, wikimedia set category: 1167, and so on.

- Filter 3 (53,158): articles that in Russian have references to known state sponsored news such as tass.ru[*].
- Filter 4 (22,046): focus on narrative sections of Wikipedia articles, rather than the bibliographic, discographic, or accolade sections[†]. For each section in article remove tables, HTML, and excessive white space. Focus on articles that have at least 2 narrative sections remaining and a total article char of at least 2000.

The richness of Wikipedia also lies in its internal connections, specifically the links between its pages. Such links guide readers towards crucial lists, categorized data, and other concepts found within Wikipedia. After Filter 2, the pages related to the USA, France, Russia, Germany, and the UK were the most linked in Russian Wikipedia. In contrast, the leading links from the English Wikipedia revolved around the 'Living People' category, The New York Times, Russia, France, and the Association Football.

An examination of the top Wikipedia links following the application of Filter 3 (as seen in Table 2) revealed strong political leanings. Topics related to Russia, Moscow, USA, USSR, and Ukraine prominently featured in both the Russian and English Wikipedia entries, attesting to the effectiveness of Filter 3. It's noteworthy that English articles link to Western sources such as The New York Times, BBC News, etc. This draws attention to a clear discrepancy in the types of politically-related sources used in the English and Russian Wikipedia entries.

Table 2: Top links to Wikipedia pages using 53,158 articles post Filter 3.

RU Wikipedia Link (translation)	RU Instances	EN Wikipedia Link	EN Instances
россия (Russia)	20001	the new york times	17136
москва (Moscow)	18019	russia	15170
сша (USA)	14050	soviet union	13805
ссср (USSR)	12563	moscow	12888
риа новости (RIA Novosti)	9963	the guardian	12235
санкт-петербург (Saint Petersburg)	7998	category:living people	12229
коммерсантъ (Kommersant)	7798	the washington post	6821
великобритания (Great Britain)	7255	reuters	6631
франция (France)	6890	bbc news	6618
lenta.ru	6803	united states	6122

5. ADDITIONAL WIKIPEDIA DATA VIA MACHINE TRANSLATION

We chose to use an LLM (GPT-4) for translation, and applied it to larger chunks of text rather than individual sentences on Wikipedia. This follows the findings of Karpinska et al. [22], who found that for Russian-English translation, human translators preferred GPT-3.5's translations at paragraph-level vs. sentence-level 100% of the time.

Translating by paragraphs provides a larger context size that allows for a more accurate translation. By switching to chunk-level translations, we not only increase this context size, but also reduce the number of inference calls and facilitate a more efficient segmentation process.

Each Wikipedia article by default contains sections. Within each section, the content is divided into chunks; where each chunk is not to exceed 3000 characters. For instance, the article on Abraham Lincoln, features chunks

[*] 30 sites of interest lenta.ru, kommersant.ru, ria.ru, tass.ru, rg.ru, gazeta.ru, rbc.ru, kremlin.ru, interfax.ru, demoscope.ru, vedomosti.ru, kp.ru, regnum.ru, vesti.ru, echo.msk.ru, novayagazeta.ru, ng.ru, rian.ru, publication.pravo.gov.ru, 1tv.ru, vz.ru, iz.ru, aif.ru, rosbalt.ru, izvestia.ru, intermedia.ru, top.rbc.ru, polit.ru, fontanka.ru, ntv.ru

[†] The heading section of article is turned to lowercase and split using whitespace. For every word in heading if it contains in Russian [примечания, ссылки, литература, также, см, награды, достижения, ...] or in English [references, links, see, notes, further, sources, awards, ...] (full list on GitHub) then heading is discarded.

as follows[†]: (i) 1: Abraham Lincoln (0), (ii) 1: Abraham Lincoln (1), (iii) 1.1.1: Abraham Lincoln - Early life, ..., (n) 1.11.4.0: Abraham Lincoln - Memory and memorials (where n is the total number of chunks that the Abraham Lincoln article broken up into). Notice bibliographic, discographic, or accolade sections are not utilized (they were removed via Filter 4).

Given the 22,046 paired articles, for RU to EN (RU2EN) 245,778 chunks are obtained and translated, while for EN to RU (EN2RU) 295158 chunks (English articles are longer then Russian counterparts). Each chunk makes up the User Prompt that is processed using the relevant System Prompt[‡]:

System Prompt for RU2EN: Your task is to translate into English the given Russian text.

System Prompt for EN2RU: Ваша задача - перевести на русский язык данный английский текст.

Table 3 shows the resulting number of chunks and total number of characters that needed to be processed. We see that English has more characters (22% more when compared to RU2EN). As to be expected, the total characters between RU and RU2EN is comparable (2.5% more characters for RU2EN) similarly EN and EN2RU are comparable (1.3% more characters when going to EN2RU). Later, we perform QA at the chunk-level.

Table 3: Four Dataset Options and Size

	Type	Number Article	Number Chunks	Total Char Across Chunks
1	RU2EN	22046	245778	334729925
2	EN	22046	295199	408149166
3	RU	22046	245871	326533548
4	EN2RU	22046	295158	413753537

6. EXISTING PROPAGANDA TECHNIQUES AND GPT-4 ZERO SHOT BASELINE

SemEval 2023 Task 3 deals with the classification of persuasion techniques at the paragraph-level [13]. We used the English labeled data to test a zero-shot baseline using GPT-4. GPT-4 contains two prompts: (i) system prompt giving context for the task and (ii) user prompt giving the query. The exact system prompt is shown in Fig. 1, and includes descriptions of the 23 types of persuasion techniques, as well as an additional option for 'None'. The figure also shows an example query and the resulting response.

6.1 Dataset Analysis

The SemEval training and development data sets collectively contained 11,780 unique English texts. Table 4 presents the instance count for each of the 23 propaganda techniques that had a confidence ≥ 50. The table exposes a significant imbalance in the SemEval dataset, with 'loaded language' and 'name calling' being the most commonly used techniques. In contrast, several techniques are sparsely represented or not represented at all.

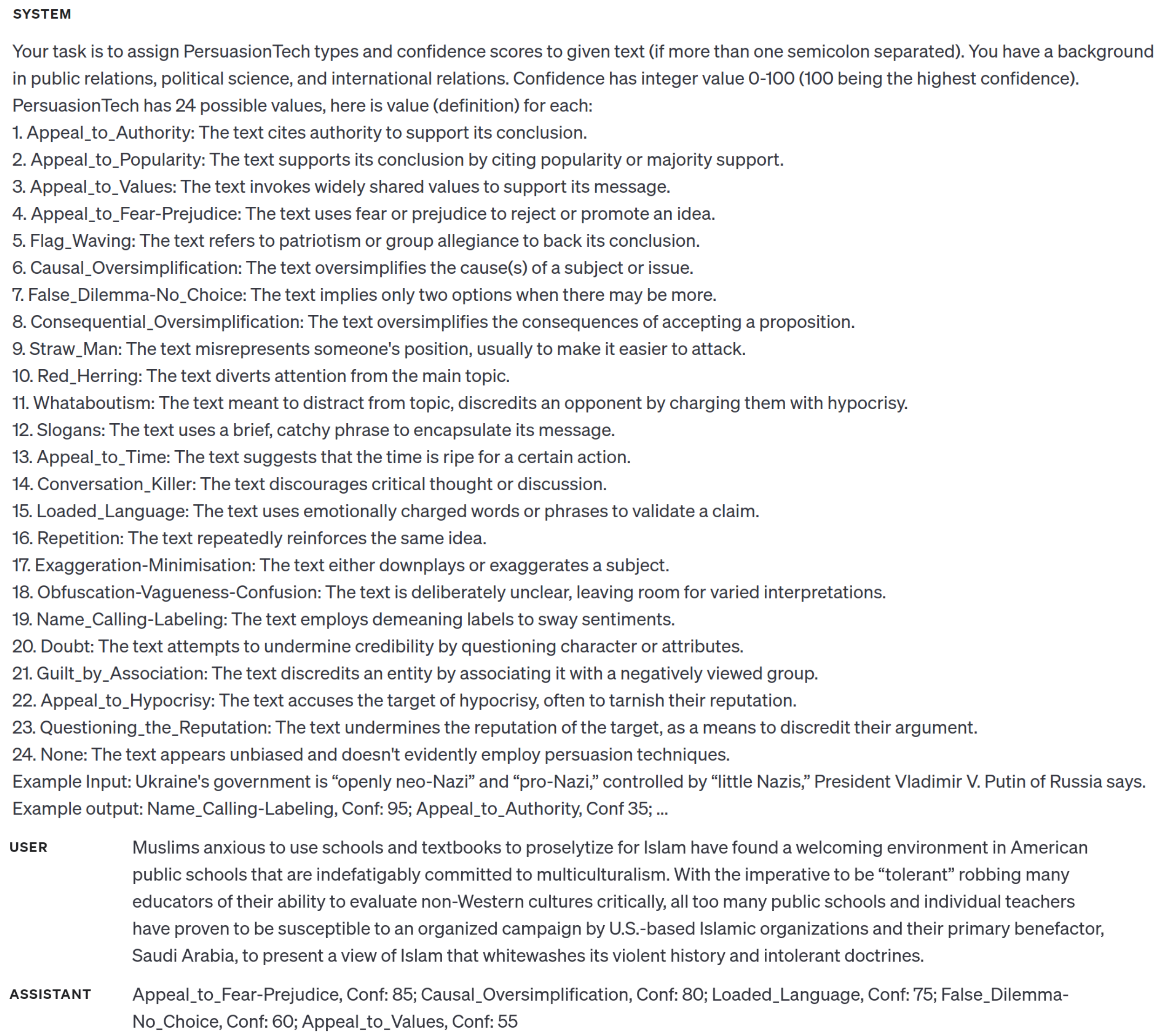
6.2 GPT-4 Zero-shot Baseline using Persuasion Technique Definitions

Predict class propaganda (if None label not predicted and one or more persuasion techniques identified with confidence $\geq x$), else class None[§]. Table 6 shows performance at various confidence x. At $x = 50$, F1 $= 0.447$.

[†] we begin chunk with a unique section count, the article's title is included to give GPT-4 additional context, Wikipedia section name (blank if introduction), and subsection count (given if section over 3000 char)

[‡] RU2EN System Prompt is translated to Russian to give the EN2RU System Prompt

[§] Despite the zero-shot nature, we find the model nearly always adheres to the expected output format: with only 116 errors (12 with no results, 86 with a persuasion technique that is not in original list, and 18 with no confidence score) across 11,780 queries.



SYSTEM

Your task is to assign PersuasionTech types and confidence scores to given text (if more than one semicolon separated). You have a background in public relations, political science, and international relations. Confidence has integer value 0-100 (100 being the highest confidence).
PersuasionTech has 24 possible values, here is value (definition) for each:
1. Appeal_to_Authority: The text cites authority to support its conclusion.
2. Appeal_to_Popularity: The text supports its conclusion by citing popularity or majority support.
3. Appeal_to_Values: The text invokes widely shared values to support its message.
4. Appeal_to_Fear-Prejudice: The text uses fear or prejudice to reject or promote an idea.
5. Flag_Waving: The text refers to patriotism or group allegiance to back its conclusion.
6. Causal_Oversimplification: The text oversimplifies the cause(s) of a subject or issue.
7. False_Dilemma-No_Choice: The text implies only two options when there may be more.
8. Consequential_Oversimplification: The text oversimplifies the consequences of accepting a proposition.
9. Straw_Man: The text misrepresents someone's position, usually to make it easier to attack.
10. Red_Herring: The text diverts attention from the main topic.
11. Whataboutism: The text meant to distract from topic, discredits an opponent by charging them with hypocrisy.
12. Slogans: The text uses a brief, catchy phrase to encapsulate its message.
13. Appeal_to_Time: The text suggests that the time is ripe for a certain action.
14. Conversation_Killer: The text discourages critical thought or discussion.
15. Loaded_Language: The text uses emotionally charged words or phrases to validate a claim.
16. Repetition: The text repeatedly reinforces the same idea.
17. Exaggeration-Minimisation: The text either downplays or exaggerates a subject.
18. Obfuscation-Vagueness-Confusion: The text is deliberately unclear, leaving room for varied interpretations.
19. Name_Calling-Labeling: The text employs demeaning labels to sway sentiments.
20. Doubt: The text attempts to undermine credibility by questioning character or attributes.
21. Guilt_by_Association: The text discredits an entity by associating it with a negatively viewed group.
22. Appeal_to_Hypocrisy: The text accuses the target of hypocrisy, often to tarnish their reputation.
23. Questioning_the_Reputation: The text undermines the reputation of the target, as a means to discredit their argument.
24. None: The text appears unbiased and doesn't evidently employ persuasion techniques.
Example Input: Ukraine's government is “openly neo-Nazi” and “pro-Nazi,” controlled by “little Nazis,” President Vladimir V. Putin of Russia says.
Example output: Name_Calling-Labeling, Conf: 95; Appeal_to_Authority, Conf 35; ...

USER

Muslims anxious to use schools and textbooks to proselytize for Islam have found a welcoming environment in American public schools that are indefatigably committed to multiculturalism. With the imperative to be “tolerant” robbing many educators of their ability to evaluate non-Western cultures critically, all too many public schools and individual teachers have proven to be susceptible to an organized campaign by U.S.-based Islamic organizations and their primary benefactor, Saudi Arabia, to present a view of Islam that whitewashes its violent history and intolerant doctrines.

ASSISTANT

Appeal_to_Fear-Prejudice, Conf: 85; Causal_Oversimplification, Conf: 80; Loaded_Language, Conf: 75; False_Dilemma-No_Choice, Conf: 60; Appeal_to_Values, Conf: 55

Figure 1: The GPT-4 System Prompt, used over SemEval 2023 dataset, assigns confidence values for the persuasion techniques used. An example query from the dataset along with the corresponding outcome is shown.

From Table 4, this baseline identified significantly more persuasion techniques (those excluding ‘None’): 24,209 vs. 7,465[¶]. This shows that the zero-shot GPT-4 approach is overly sensitive to labeling text with persuasive techniques, when they are, according to SemEval annotators, actually neutral statements.

Examples of instances where GPT-4 can be overly sensitive: ‘Email *’ (no persuasion technique in Gold) labeled as an Obfuscation-Vagueness-Confusion technique with 75% confidence. Even though ‘Email *’ does come across as vague, it's not the kind of vagueness or confusion specifically designed to deflect attention away from the main topic (482 such vagueness or confusion instances in Table 4 vs. just 30 for Gold). Similarly, merely mentioning a news source or an individual results in GPT-4 labeling the text as containing appeal to authority (6286 via GPT-4 vs. just 179 for Gold).

[¶] There are 11,780 texts, but 24,209 techniques because each text can include multiple propaganda techniques

Table 4: Propaganda Technique Counts in the SemEval Gold vs. GPT-4 Baseline (confidence ≥ 50)

Persuasion Technique	Gold vs. Base	Persuasion Technique	Gold vs. Base
None	6945 vs. 1450	Conversation_Killer	115 vs. 120
Loaded_Language	2277 vs. 2484	Red_Herring	63 vs. 101
Name_Calling-Labeling	1226 vs. 1871	Guilt_by_Association	63 vs. 339
Doubt	703 vs. 2824	Appeal_to_Popularity	48 vs. 478
Repetition	684 vs. 407	Appeal_to_Hypocrisy	45 vs. 104
Exaggeration-Minimisation	576 vs. 1571	Obfuscation-Vagueness-Confusion	30 vs. 482
Appeal_to_Fear-Prejudice	442 vs. 2260	Straw_Man	24 vs. 19
Flag_Waving	376 vs. 46	Whataboutism	18 vs. 179
Causal_Oversimplification	236 vs. 848	Appeal_to_Values	0 vs. 1938
False_Dilemma-No_Choice	180 vs. 307	Consequential_Oversimplification	0 vs. 361
Slogans	180 vs. 124	Appeal_to_Time	0 vs. 577
Appeal_to_Authority	179 vs. 6286	Questioning_the_Reputation	0 vs. 483

7. PERSUASIVE TECHNIQUE DETECTION THROUGH QUESTION ANSWERING

Given the limited performance of the baseline prompt, we develop an approach based on high-level questioning which better aligns LLM understanding of persuasion to human judgement. A high-level overview of our approach is shown in Fig. 2.

7.1 Comprehensive Set of Questions covering various Persuasion Techniques

The question-answering based approach to persuasion techniques proceeds in several steps. First, for each technique, we create a group of related questions that break down the task into simpler parts. The key insight is that we can use the LLM itself to generate these questions, thereby providing a platform for the LLM's understanding of each technique to be demonstrated. Then, we filter down the large set of questions to a more efficient subset by setting a threshold for high quality questions with reference to the SemEval dataset. We describe this approach below.

We expand the coverage of the questions by augmenting the 23 persuasion techniques from SemEval with 10 communication and 12 presentation of arguments techniques[‖]; which we manually curated ourselves (with assistance from GPT-4). These additional techniques have been incorporated to offer a wider perspective and to ensure a more comprehensive understanding of the task.

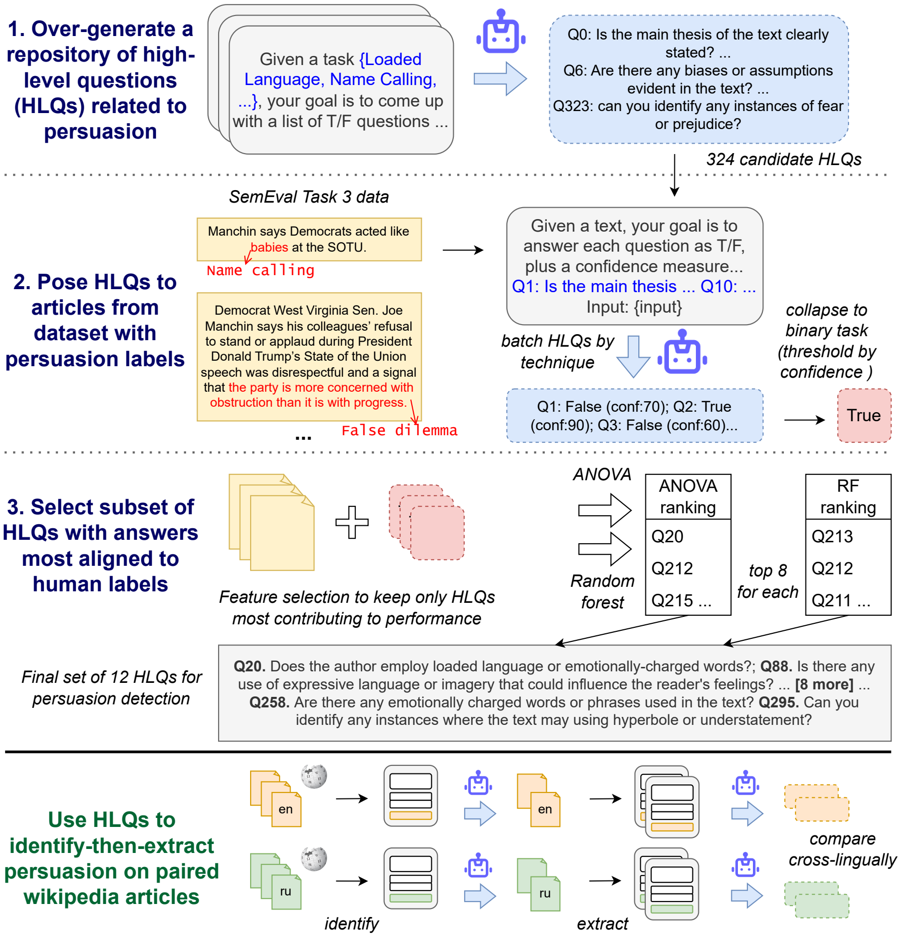
Processing the 46 tasks (23 persuasion + 22 communication + 1 for None) results in 324 Questions[**]. For each task, this GPT-4 prompt is utilized (system prompt remains the same, user prompt specifies each task):

System Prompt *(remains constant)*:

Given a task X, your goal is to come up with a list of questions Y. The list Y contains questions that break the task into simpler components. Questions in list Y should be binomial: True or False. Questions in list Y should be semicolon separated. Avoid questions that rephrase the task, but do not simplify it.

[‖] Communication practices: (1) Respectful Language, (2) Constructive Dialogue, (3) Empathetic Tone, (4) Open-mindedness, (5) Objectivity, ... Presentation of Arguments: (1) a well-defined Thesis, (2) robust Evidence such as statistics and expert opinions, (3) Logical Reasoning with coherent thought progression, (4) Relevance of all points to the overarching thesis, (5) addressing Counterarguments, ... full list and definitions on GitHub.

[**] The full question repository is on GitHub.


1. Over-generate a repository of high-level questions (HLQs) related to persuasion
324 candidate HLQs
SemEval Task 3 data
Name calling
False dilemma
...
2. Pose HLQs to articles from dataset with persuasion labels
batch HLQs by technique
collapse to binary task (threshold by confidence)
True
3. Select subset of HLQs with answers most aligned to human labels
Feature selection to keep only HLQs most contributing to performance
ANOVA
Random forest
ANOVA ranking
Q20
Q212
Q215 ...
top 8 for each
RF ranking
Q213
Q212
Q211 ...
Final set of 12 HLQs for persuasion detection
Use HLQs to identify-then-extract persuasion on paired wikipedia articles
en
ru
identify
extract
compare cross-lingually

Given a task {Loaded Language, Name Calling, ...}, your goal is to come up with a list of T/F questions ...

Q0: Is the main thesis of the text clearly stated? ...
Q6: Are there any biases or assumptions evident in the text? ...
Q323: can you identify any instances of fear or prejudice?

Manchin says Democrats acted like babies at the SOTU.

Democrat West Virginia Sen. Joe Manchin says his colleagues' refusal to stand or applaud during President Donald Trump's State of the Union speech was disrespectful and a signal that the party is more concerned with obstruction than it is with progress.

Given a text, your goal is to answer each question as T/F, plus a confidence measure... Q1: Is the main thesis ... Q10: ... Input: {input}

Q1: False (conf:70); Q2: True (conf:90); Q3: False (conf:60)...

Q20. Does the author employ loaded language or emotionally-charged words?; **Q88.** Is there any use of expressive language or imagery that could influence the reader's feelings? ... **[8 more]** ... **Q258.** Are there any emotionally charged words or phrases used in the text? **Q295.** Can you identify any instances where the text may using hyperbole or understatement?

Figure 2: Overview of our approach for persuasion detection (each robot icon is its own LLM interaction). **Top:** we have an LLM generates a large repository of high-level questions (HLQs), based on its own understanding of persuasion techniques. We then pose these HLQs to articles from an labeled persuasion dataset [13]. We select the subset of questions which are most aligned to the human labels, using feature selection to arrive at 12 HLQs. **Bottom:** on a different dataset of interest, we use HLQs to prompt an LLM to *identify-then-extract* persuasive spans. This is done over paired Wikipedia articles in Russian and English, facilitating cross-lingual comparison.

User Prompt ***(specifies task and task definition)*****:**

Given a piece of text your goal is to identify whether the text has {Task}: {Task Definition}.

Example User Prompt for Obfuscation-Vagueness-Confusion task:

Given a piece of text your goal is to identify whether the text has Obfuscation-Vagueness-Confusion: where the text is deliberately unclear, leaving room for varied interpretations.

Example GPT-4 Response for Obfuscation-Vagueness-Confusion task:

Does the text provide clear and specific details?; Are there phrases or sentences in the text that can be interpreted in more than one way?; Can you identify parts of the text where ambiguity or vagueness have been intentionally used?; . . .

7.2 Question Selection

Next, we identify which of the 324 questions are most effective at detecting persuasive text. Here, we leverage the existing annotations from SemEval 2023 Task 3. The idea is to have the LLM answer each of the questions

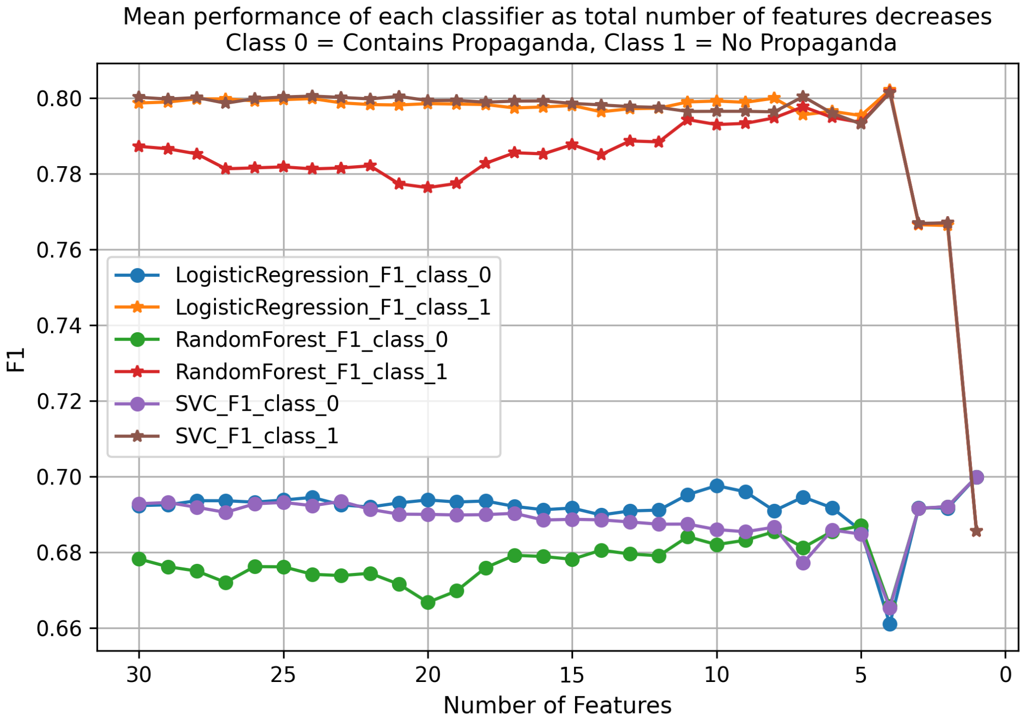


Figure 3: The effectiveness of the classifiers after each feature reduction using ANOVA, showcasing the relationship between the F1-score and the progressively diminished feature sets. The performance of the classifiers at the reduced feature set range from 30 to 1; it starts to deteriorate after less than 10 features remain.

Q0-Q323, for each article text (11,780 total). We then rank the questions based on the performance on the SemEval annotations, and select the top ranked questions.

> **System Prompt:** Given a piece of text your goal is to answer each of the following questions as 'True', 'False', or 'N/A' (if question is not applicable) plus a confidence measure from 0-100. Questions: {list Y associated with task X}. Example output: Q0: True (conf:70); Q1: False (conf:30); Q2: N/A; . . . .

The system prompt will change based on the set of questions making up the task X. For example, for task 'Flag Waving' the corresponding question list Y is Q42: Is the text using patriotic language or symbols? Q43: Does the text appeal to group loyalty or allegiance? Q44: Are there references to national or group pride in the text? Q45: Is the conclusion of the text supported by patriotism or group identity? Q46: Does the text use group affiliation to justify its argument?

User Prompt will be the SemEval text to process. For example, for SemEval text: 'He served in Vietnam with the U.S. Army in Military Intelligence' and system prompt related to task 'Flag Waving', GPT-4 outputs: Q42: True (conf: 60); Q43: False (conf: 90); Q44: False (conf: 90); Q45: N/A; Q46: N/A.

In this way, the 324 questions are answered across 11,780 texts. Each answer is mapped to 1 for True, 0 for N/A, and -1 for False. A separate matrix capturing confidence metric is created (if no confidence supplied by GPT-4, assign a 0). For the gold label 'None' one-hot encoding is applied (Class 0 contains propaganda; Class 1 corresponds to label 'None').

In our feature selection methodology, we embarked on a systematic elimination of features, ranking them using either ANOVA or Random Forest with Gini impurity (RFGini) and removing least important feature one at a time. First, the SemEval data was employed in a 5 fold cross-validation across a suite of classifiers, which included Logistic Regression, Random Forest, and Support Vector Classifier (SVC) (across classifiers kept seed parameter constant). The weighted average F1 across the 5 folds was recorded. Following the reduction of each feature set, we applied 5 fold cross-validation using the suite of classifiers again. For each iteration, we computed the weighted average precision, recall, and F1-score to quantitatively assess the classifiers' performance on the reduced feature set.

Fig. 3 illustrates the impact of feature reduction on the classifiers' effectiveness by plotting the F1-score against the progressively diminished feature sets using ANOVA. Elimination of features for both ANOVA and RFGini reveals a stable performance across classifiers with a sharp drop after top 8 features remain. We therefore take the top 8 selected features from feature reduction using ANOVA, and also the top 8 features from feature reduction using RFGini. As 4 questions overlap, this results in 12 features, which are shown in Table 5.

Table 5: HLQ ID to rank based on analysis using ANOVA and RFGini.

ID	(ANOVA, RFGini Rank): Question
20	0, 3: Does the author employ loaded language or emotionally-charged words?
88	12, 7: Is there any use of expressive language or imagery that could influence the reader's feelings?
92	9, 5: Does the text make use of positive or negative connotations to sway the reader's opinion?
210	6, 9: Does the text contain words or phrases that evoke strong emotions?
211	3, 2: Are there words or phrases in the text that are intended to manipulate the reader's feelings?
212	1, 1: Can you identify any instances where emotionally charged language is used to support a claim?
213	8, 0: Are there parts in the text where the language is used to influence reader's opinion or decision?
215	2, 31: Does the text use language that is intended to provoke a particular reaction from the reader?
216	5, 19: Can you find any instances where the language used is not neutral or objective?
217	7, 12: Does the text use language that is intended to sway the reader's viewpoint?
258	4, 4: Are there any emotionally charged words or phrases used in the text?
295	20, 6: Can you identify any instances where the text may be using hyperbole or understatement?

Each question is related to task X as follows: Q20 pertains to objectivity, Q88 and Q92 address persuasive language to evoke emotional responses, Q210 through Q217 explore aspects of loaded language, Q258 associated with none, and Q295 examines the use of exaggeration or minimization. In this way, 9 out of the 12 questions are related to loaded or emotional language. This was expected given that 'loaded language' is by far the most frequent category in SemEval. These texts may display a positive or negative sentiment with the usage of intensifiers like 'great' or 'horrible'. Nevertheless, the emotional undertone in such texts implies subjective bias. In the rest of the paper we refer to these 324 questions as High-Level Questions (HLQs).

7.3 GPT-4 Zero-Shot 24 Definition Baseline vs. 12 HLQs

We employ the same approach as used for the 24 definition baseline, but now instead of using persuasion technique definitions we are using the 12 HLQs. Predict class propaganda (if one or more HLQs answer affirmatively), else class None. For the 24 definition baseline each label, none and 23 persuasion techniques, was taken into account if its confidence surpasses a distinct threshold, denoted as x. Similarly, for 12 HLQs, we consider those that answer True with a confidence $\geq x$.

From Table 6, when x is close to zero, nearly everything is propaganda. Requiring higher and higher confidence results in fewer and fewer features passing the threshold till nearly everything is classified as having no propaganda (for $x \geq 95$)[††]. For both the baseline and HLQs the maximum GPT-4 performance is seen at $x = 85$. Table

[††] Number of predictions that are None can be obtained by 11,780 - propagadna prediction count – for example, for $x = 20$, for baseline, None count = 1701

demonstrates the useful role that asking the LLM to output confidence scores plays. It also validates the feature elimination methodology and the resulting HLQs. Finally, the approach of predicting text to contain propaganda if one or more HLQs answer affirmatively is shown to be competitive with top classifier scores over all 324 HLQs: 0.738 $x = 85$ vs. Logistic Regression (top F1 = 0.737), Random Forest (top F1 = 0.759), and SVC (top F1 = 0.753). In rest of paper, we use the 12 HLQs for checking whether text has propaganda and for identifying the specific text responsible for affirmative response*. Each HLQ is translated to Russian to support QA over Russian texts.

Table 6: Performance of 24 Definition Baseline vs. HLQs as confidence x increases.

Confidence ≥ x	Baseline F1 (Propaganda Predictions)	12 HLQs F1 (Propaganda Predictions)
x=10	0.472 (10050)	0.641 (8300)
x=20	0.469 (10079)	0.641 (8300)
x=30	0.459 (10165)	0.641 (8300)
x=40	0.454 (10223)	0.641 (8300)
x=50	0.447 (10293)	0.641 (8300)
x=60	0.45 (10266)	0.644 (8258)
x=70	0.509 (9401)	0.696 (7273)
x=80	0.582 (7233)	0.737 (5685)
x=85	0.608 (4264)	0.738 (4002)
x=90	0.573 (2833)	0.723 (3485)
x=95	0.448 (99)	0.648 (1832)

8. EMOTIONAL LANGUAGE IN RUSSIAN VS. ENGLISH WIKIPEDIA DATASET

With the questions and datasets in place, we can now apply the 12 HLQs (*2 languages) to the 22k articles (*4 language settings). Our approach is illustrated in Fig. 4. For each setting, we run its respective language HLQ prompts in two step setting, record specific text spans, and use them for ranking. We illustrate example ranking for English vs. Russian in Sec 9. The following steps are employed:

- Step 1 System prompt (apply 12 HLQs to the input text): Given a piece of text your goal is to answer each of the following questions as 'True', 'False', or 'N/A' (if question is not applicable) plus a confidence measure from 0-100. Questions: {list of 12 questions}. Example output: Q1: True (conf:70); Q2: False (conf:30); Q3: N/A; ...
- Step 2 System prompt (for every HLQ that was answered with 'True', identify the exact part of the text that backs up the answer): Given a piece of text your are tasked with a question: {Question that answered True} Identify specific language instances separated by semicolons.
- Step 3: tally up all the characters in (1) the input text and (2) all the responses to the HLQs.

The above prompts are used for options RU2EN and EN. For options RU and EN2RU, the same steps are followed, but the prompts, and the 12 HLQs within, are translated to Russian (and verified by a native speaker). For example, HLQ20 becomes: 'Использует ли автор насыщенный язык или эмоционально окрашенные слова?'. For Step 2, the following Russian system prompt is used: 'Вам дан текст, и вас просят ответить на

[*] Because the SemEval 2023 did not contain the binary classification and because the 2019 SLC Task featured the same English data, it is therefore the most closely aligned. The top performer on the development set for this task achieved F1 0.6883, Precision 0.6104, and Recall 0.7889 [10]. This suggests that the proposed approach is effective.

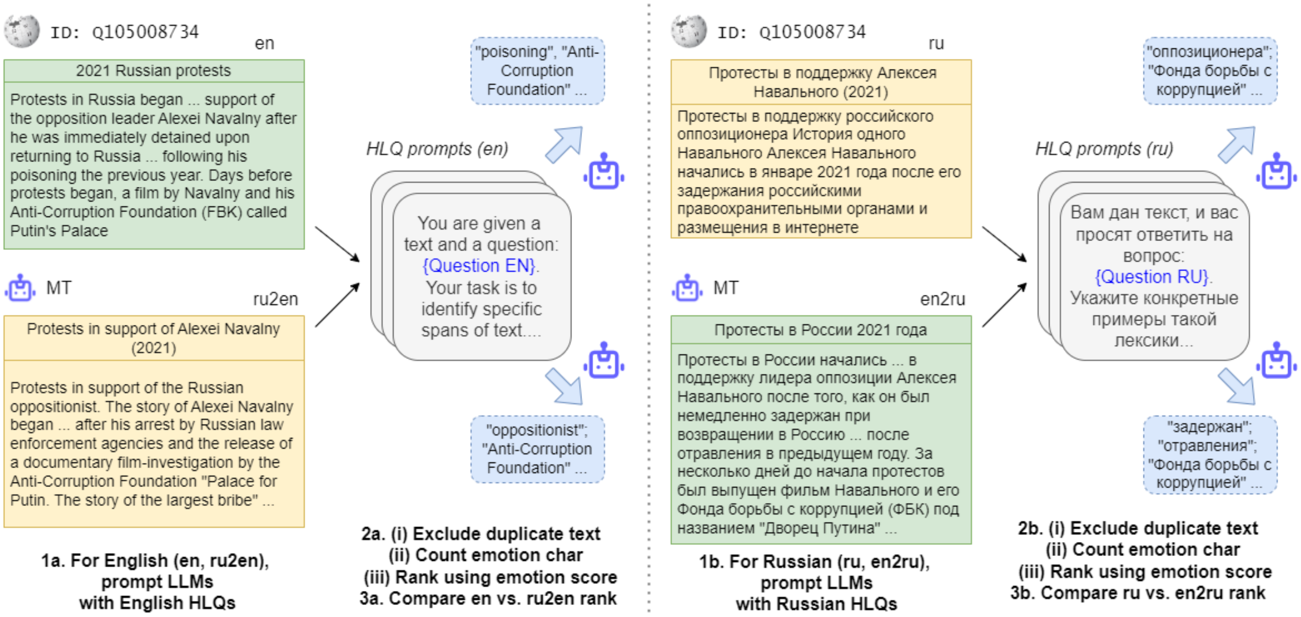

ID: Q105008734
en
2021 Russian protests
Protests in Russia began ... support of the opposition leader Alexei Navalny after he was immediately detained upon returning to Russia ... following his poisoning the previous year. Days before protests began, a film by Navalny and his Anti-Corruption Foundation (FBK) called Putin's Palace
MT
ru2en
Protests in support of Alexei Navalny (2021)
Protests in support of the Russian oppositionist. The story of Alexei Navalny began ... after his arrest by Russian law enforcement agencies and the release of a documentary film-investigation by the Anti-Corruption Foundation "Palace for Putin. The story of the largest bribe" ...
1a. For English (en, ru2en), prompt LLMs with English HLQs
HLQ prompts (en)
You are given a text and a question: {Question EN}. Your task is to identify specific spans of text....
"poisoning", "Anti-Corruption Foundation" ...
"oppositionist"; "Anti-Corruption Foundation" ...
2a. (i) Exclude duplicate text
(ii) Count emotion char
(iii) Rank using emotion score
3a. Compare en vs. ru2en rank
ID: Q105008734
ru
Протесты в поддержку Алексея Навального (2021)
Протесты в поддержку российского оппозиционера История одного Навального Алексея Навального начались в январе 2021 года после его задержания российскими правоохранительными органами и размещения в интернете
MT
en2ru
Протесты в России 2021 года
Протесты в России начались ... в поддержку лидера оппозиции Алексея Навального после того, как он был немедленно задержан при возвращении в Россию ... после отравления в предыдущем году. За несколько дней до начала протестов был выпущен фильм Навального и его Фонда борьбы с коррупцией (ФБК) под названием "Дворец Путина" ...
1b. For Russian (ru, en2ru), prompt LLMs with Russian HLQs
HLQ prompts (ru)
Вам дан текст, и вас просят ответить на вопрос: {Question RU}. Укажите конкретные примеры такой лексики...
"оппозиционера"; "Фонда борьбы с коррупцией" ...
"задержан"; "отравления"; "Фонда борьбы с коррупцией" ...
2b. (i) Exclude duplicate text
(ii) Count emotion char
(iii) Rank using emotion score
3b. Compare ru vs. en2ru rank

Figure 4: High-level Overview of our approach for processing four Wikipedia datasets. **Left:** For en and ru2en datasets, the common language is English, and so the 12 HLQs in English are applied. For each dataset the responses to the 12 HLQs are used to generate an emotional score. Over many articles the emotional score can be used to get a rank. The emotional score and corresponding rank between en and ru2en dataset can be compared. **Right:** For ru and en2ru datasets, the common language is Russian, and so the 12 HLQs in Russian are applied. The other steps are the same with emotional score and corresponding rank between ru and en2ru dataset is compared. Later in Sec 9.2 we also compare how en vs. en2ru and ru vs. ru2en ranks compare in order to gauge how translation affects ranking.

вопрос: {HLQ that answered True in Russian} Укажите конкретные примеры такой лексики, разделенные точкой с запятой.' Further details and outcomes for each step are provided below.

We also ran a small-scale experiment on these 3 steps using an open-source LLM, Llama 2 (13B). In sum, we found that while this showed promise, it required some extra prompt engineering effort and various tricks to work (see Appendix A). Overall performance was worse than for GPT-4, so we use GPT-4 in the rest of paper.

8.1 Step 1: Apply 12 HLQs to the input text

The upper section of Fig. 5 highlights the output from the LLM for each HLQ projected over all text chunks. It shows what percentage of the text chunks provides affirmative answers to each HLQ, denoting the proportion of the chunks where relevant information can be extracted. Interestingly, we find that on average, there's an 85.21% reduction in the number of chunks that need further processing in the subsequent step. This average is calculated across all HLQs over four data types.

As is logically expected, HLQ with a narrower scope, like HLQ295 (presence as hyperbole or understatement), yielded fewer affirmative responses (5.1% on average). On the contrary, broader HLQ like HLQ216 that scrutinize neutrality and objectivity, produced a more substantial volume of affirmative results (16.8% on average).

8.2 Step 2: Identify the exact part of the text that backs up the answer

In step 2, text chunks that had previously responded affirmatively in Step 1 are examined to identify the specific text resulting in this affirmative response. As depicted in the scatter plot at the lower section of Fig. 5, there are minimal queries returning empty text. This graphically reinforces the idea that a significant majority of initially affirmative replies do yield some form of extracted text.

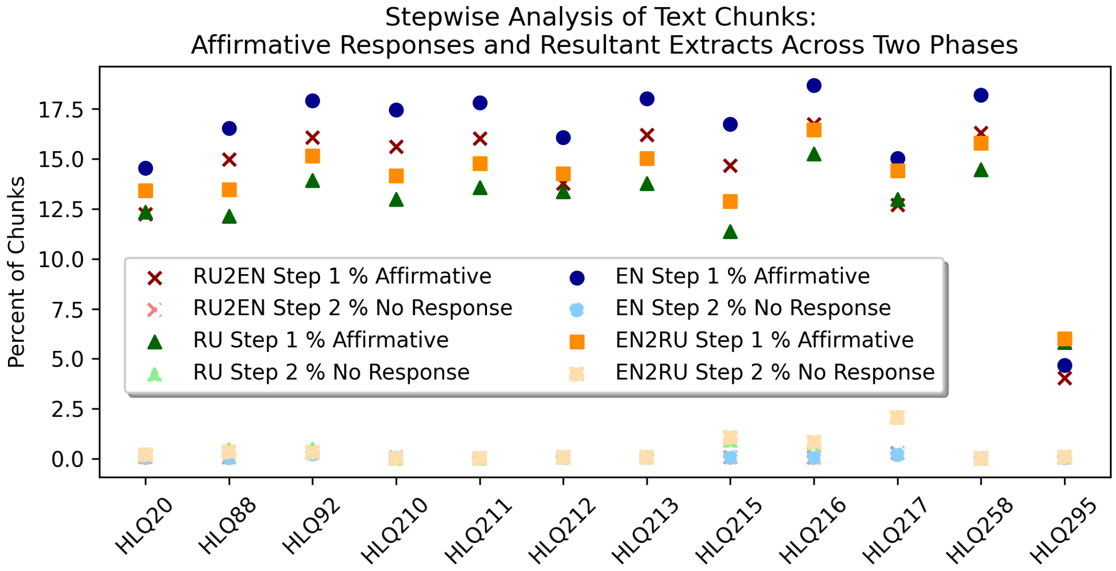


Figure 5: Top scatter plot, shows the percent of chunks that responded affirmatively to each HLQ for step 1. In the subsequent step, all chunks that had an affirmative response are examined to isolate the respective text that led to this answer. The bottom scatter plot shows percent of chunks from step 2 that yielded no result. It effectively portrays that the majority of responses, which were initially affirmative, produced the text responsible for such answers as per expectation.

The frequency of such blank replies in English is comparatively low, averaging at about 0.71% for RU2EN translations and just 0.49% for English-only queries. Nonetheless, this average percentage increases when HLQs are posed in Russian, clocking up to 3.46% for Russian only and 3.03% for EN2RU translations.

The dual stages of text extraction act as an additional check by verifying through two separate queries that the text chunk indeed contains information worth extracting. Table 7 shows example text extracted for the article ‘Fire at the National Museum of Brazil’ (ID Q56441760). Sections Q56441760_2 and Q56441760_3 deal with ‘Fire’ and ‘Consequences’ respectively. The text that supports the affirmative answer is turned to lowercase, as shown for HLQ20 and HLQ88.

Table 7: Example HLQ20 and HLQ88 responses for two chunks from ‘Fire at the National Museum of Brazil’

WikiID_Sec	HLQ_ID	Specific Text Instances Identified
Q56441760_2	HLQ20	‘engulfed’, ‘rapidly destroyed’, ‘tragedy’, ‘repeatedly complained’, ...
Q56441760_2	HLQ88	‘fire engulfed’, ‘rapidly destroyed’, ‘tragedy’, ‘funding cuts’, ‘deteriorating’, ...
Q56441760_3	HLQ20	‘incalculable’, ‘outraged’, ‘cultural tragedy’, ‘lobotomy’
Q56441760_3	HLQ88	‘fire’, ‘loss’, ‘outraged’, ‘tragedy’, ‘destroyed’, ‘ruins’, ‘threat’, ...

Given that many of the HLQs are related, we can expect that many of the same text passages will be quoted. For example, there is overlap between HLQ20 and HLQ88 for chunk Q56441760_2. We want to exclude any duplicate text. For each section we aggregate the responses across the questions into set A and initialize an empty set B. For each string in set A, we check whether or not any other string in the set B contains the current string. If no such string is found, then the string is considered unique and consequently added to set B. This method ensures that the output set is fully reduced, meaning no string in the set B is a substring of any other string within the same set. Thus the responses shown in Table 7 for Q56441760_2 for HLQ20 and HLQ88 should be reduced to [‘fire engulfed’, ‘rapidly destroyed’, ‘tragedy’, ‘repeatedly complained’, ‘funding cuts’, ‘deteriorating’] with duplicates [‘engulfed’, ‘rapidly destroyed’, ‘tragedy’] removed.

8.3 Step 3. Rank Wikipedia articles by Emotional Content

We propose to rank Wikipedia articles by amount of emotional content. This is done by, for each article (aggregated over the chunks) comparing the (1) amount of text detected as emotional to (2) the total amount of text. The advantage of this straightforward approach is that it can be applied to virtually any article.

As an example, for article Q56441760, 231 characters were found to be emotional out of 2979 in total. Table 8 shows the top Wikipedia articles with the most amount of emotional language used. This can be used to quickly gauge the topics that worry each culture the most.

Table 8: Top articles with the most emotion from Russian and English Wikipedia

Top Russian Articles	Top English Articles
Q844787: Organization of Ukrainian Nationalists	Q554482: Persecution of Christians
Q737212: Ukrainian Insurgent Army	Q16335075: War in Donbas (2014–2022)
Q4445396: Bhagavad Gita As It Is trial in Russia	Q17324420: 2014 Gaza War
Q1003: Solidarity (Polish trade union)	Q493302: Hizb ut-Tahrir
Q74365: Timeline of the Syrian civil war	Q83085: Soviet–Afghan War
Q2090117: Russia–Ukraine relations	Q156537: Domestic violence
Q8729: Russian Revolution	Q1800556: Violence against women
Q157280: Stepan Bandera	Q622820: Women in Islam
Q299681: Georgy Gapon	Q1072770: Human rights in China
Q101534: February Revolution	Q20276006: Cultural impact of Madonna

9. RANKED EMOTION BETWEEN RUSSIAN AND ENGLISH WIKIPEDIA

Using the Wikipedia article pairs we can analyze where the Russian and English authors displayed the most and least emotion. For simplicity, we label the Russian as ‘Author 1’ (Author Rus) and the English as ‘Author 2’ (Author Eng). We follow Section 8.3 in calculating the proportion of detected emotion to total characters, terming this the Emotional Frequency (EF). EF is computed for each article written by Author 1 and Author 2.

For those experimented making comparisons between authors, we propose a separate metric. For this, all EF values are combined and normalized from 0 to 1, which we term Normalized Emotional Frequency (NEF). Putting them on a common scale makes comparing the relative emotional content between authors more meaningful[‡‡]. The high-level procedure is captured via the corresponding pseudocode below:

```
author1_ef = calc_emotional_freq(author1_article_length_list, author1_emotional_length_list)
author2_ef = calc_emotional_freq(author2_article_length_list, author2_emotional_length_list)
normalized_emotional_freq_nef = normalize_scores(author1_ef + author2_ef)
```

Where ‘calc_emotional_freq’ returns author_emotional_length_list[i] / author_article_length_list[i] (for i = 0, 1, ... n-1) where n is the number of articles.

[‡‡] When calculating NEF values across the Wikipedia dataset, we utilize the scores across the whole 22K articles as it allows one to see how emotional score of the article compares vs. all other articles in both English and Russian

9.1 Grouping Wikipedia articles into Broader Categories through WikiData

WikiData operates as an RDF triple store, linking nodes via properties. This relationship can be conceptualized as $Subject- > Predicate- > Object$. Predicates are marked with a ‘P’ before them (for instance, ‘P31’ refers to ‘instance of’), while nodes, taking up the role of either ‘Object’ or ‘Subject’, are preceded by a ‘Q’.

We use the property ‘P31’, or ‘instance of’, to identify the high level categories of each Wikipedia page. For instance, every article pertaining to a specific person will be instance of ‘Q5’. As a consequence, we can conduct an extensive analysis on specific subjects or themes across a range of Wikipedia articles that are classified under a particular ‘P31’ category.

Table 9 displays example P31 categories, the number of related articles captured by a count, and the corresponding emotional score for each author. The last column, used for sorting, is the average emotional score for both authors. In this way, the table showcases the top categories with the least and greatest emotional scores.

Table 9: Most and Least Emotional Categories for Russian and English Speakers

WikiData Category (Description)	Count	Author Rus NEF	Author Eng NEF	avg(NEF)
Q180684 (Disagreement Situation)	65	0.303	0.326	0.314
Q47461344 (Written Work)	53	0.301	0.305	0.303
Q178561 (Part of War)	68	0.304	0.284	0.294
Q7278 (Org Influences Gov)	138	0.247	0.296	0.271
Q43229 (Social Entity)	122	0.229	0.257	0.243
...	...	...	...	...
Q23038290 (Fossil Taxon)	52	0.044	0.045	0.045
Q15056993 (Aircraft)	153	0.05	0.035	0.042
Q786820 (Automaker)	52	0.025	0.054	0.04
Q2198484 (Admin Entity)	132	0.038	0.037	0.038
Q14795564 (Date Calculator)	217	0.036	0	0.018

Table 10: WikiData nodes with highest emotional score (E) that are instance of (P31) disagreement (Q180684)

WikiID: Title	Author Rus E	Author Eng E	avg E
Q182865: War in Afghanistan (2001–2021)	37240	66117	51678.5
Q164348: Hungarian Revolution of 1956	44853	43917	44385
Q21083298: 2015–2016 Israeli–Palestinian conflict	38326	35266	36796
Q47465940: Operation Olive Branch	14579	48234	31406.5
Q131297: Rwandan genocide	36647	24134	30390.5
Q4131735: 2009 Russia–Ukraine gas dispute	39230	14665	26947.5
Q169072: Colombian conflict	18440	34921	26680.5
Q82973: Transnistria conflict	29450	4234	16842
Q459282: United States invasion of Panama	13374	14572	13973
Q48150708: Battle of Khasham	15453	11599	13526

The analysis uncovers significant differences in the amount of emotional content found within various categories. The categories that evoke the most emotion include disagreement, friction and discord scenarios

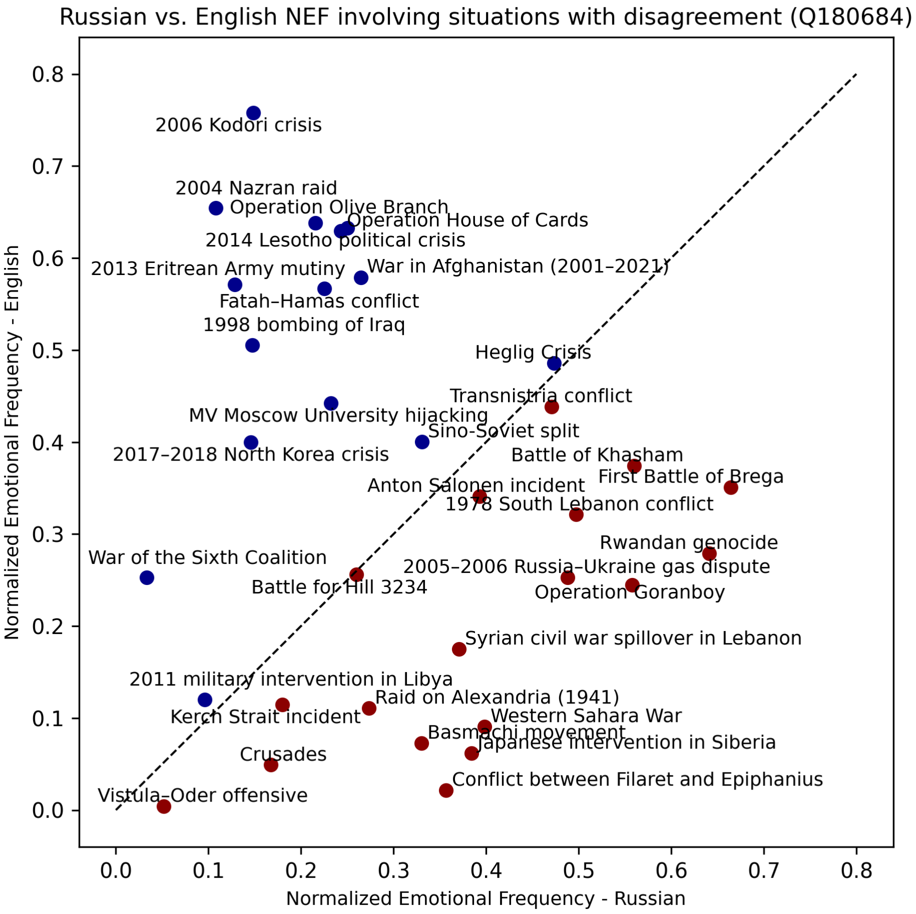


Figure 6: Scatterplot where the x and y positions represent the NEF values of Russian and English articles, respectively. Red points signify articles with a greater use of emotional language in the Russian Wikipedia, while blue points represent a higher level of emotional language use in the English Wikipedia. The dashed line provides a reference, illustrating articles where the amount of emotional language used is approximately equal. The closer points are to this line, the more similar the level of emotional language used in both.

(Q180684), written works (Q47461344), elements of warfare (Q178561), political organizations (Q7278), and formal societal organizations (Q43229). These categories deal with political and societal issues like war. For instance, written works in our dataset predominantly focused on historical atrocities, war, political repression, and contentious analyses of societal structures. This validates our hypothesis that political-related events contain more emotional content. As expected, more neutral categories, such as fossil taxon (Q23038290), aircraft model groups (Q15056993), automakers (Q786820) have the least emotion.

We can spotlight specific articles within each category of interest. For instance, among disagreement scenarios (Q180684), category that featured the highest emotional content, we can examine specific articles. Table 10 demonstrates total characters with emotional content for specific articles. From Table, the “2009 Russia–Ukraine gas dispute” has a lot more emotion in Russian vs. English. Conversely, the article “Operation Olive Branch"– which deals with the Turkish Armed Forces and Syrian National Army’s conflict with the People’s Protection Units of the Syrian Democratic Forces – displays more emotional content in English.

While Table 10 displays total characters with emotional content used, Fig. 6 offers a visual representation of the NEF for Wikipedia articles that fall under disagreement scenario category. This illustrates visually all the entries under the category in our dataset, including all entries from Table 10.

An article's NEF score can be used to assign categories. For example, if the bottom 25th, 25th-75th, and over 75th percentile are assigned as low, medium, and high emotion: the range for medium emotion would be $0.03 \leq \text{NEF} \geq 0.28$ (RU2EN and EN) and $0.02 \leq \text{NEF} \geq 0.27$ (RU and EN2RU).

9.2 Comparing Effectiveness of Translated QA

We have thus far performed monolingual QA across four settings. If the translation and QA methods prove to be successful, we expect the amount of emotional text extracted from the Russian-to-English (RU2EN) translation to be similar to that from the original Russian (RU) text. This is because the text passages with emotional content would essentially be the same in both contexts—they would merely be translations of each other. The same reasoning applies to the English-to-Russian (EN2RU) and English rankings.

Ranking by Emotional Score We centered our analysis on the top 2000 (of 22,046) articles, for each dataset option, ranked by emotional score (denoted as dataset_E). Ranked-Biased Overlap (RBO) [23] was used to assess how similar the top rankings are across these datasets. The RBO metric measures the similarity between two ordered lists by taking into account both the order of items and where the lists overlap. RBO scores range from 0 (no similarity) to 1 (identical rankings).

The findings, illustrated in Fig. 7, reveal a significant overlap in rankings between RU2EN_E and RU_E, with an RBO score of 0.85. Similarly, there is a notable overlap between EN2RU_E and EN_E, evidenced by an RBO score of 0.83. This supports the hypothesis that the questions and answers can be executed equally well in either English or Russian. Achieved by translating the texts into a common language and using translated queries for answering. This insight is particularly compelling as it demonstrates the potential for implementing systems in non-native languages through translation, enabling the extraction of relevant features from the text.

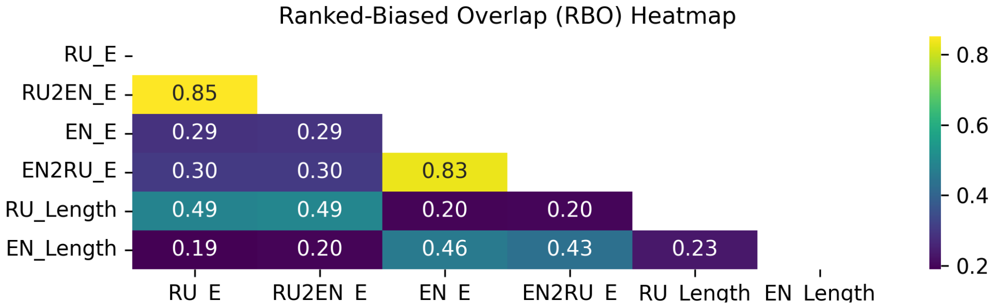


Figure 7: RBO comparisons reveals significant overlap between emotion score rankings in Russian-to-English (RU2EN_E) and Russian (RU_E) with an RBO score of 0.85. A similar high overlap is noted between English-to-Russian (EN2RU_E) and English (EN_E) rankings, with an RBO score of 0.83. These findings validate the notion that question-answering sessions can be conducted just as effectively in either English or Russian.

Ranking by Article Length We also evaluated the list ranked according to the greatest article length in Russian and English, denoted by RU_Length and EN_Length. We compiled the correlation between the Article Character Length and the Emotion Character Length, which resulted in a correlation of 0.96 for Russian (RU_Length to RU_E) and 0.95 for English (EN_Length to EN_E). This analysis indicates that, as expected, articles tend to contain more emotional text as they become longer.

However, the low RBO of 0.43-0.49, when comparing a ranking of articles based on their length vs. the emotion score, demonstrates that the two rankings significantly differ from each other. Reinforcing the point that it's not feasible to replicate a ranking, similar to the one achieved via emotion scores, solely by using article length.

10. SUMMARY AND FUTURE STEPS

This research paper utilizes latest LLMs, specifically OpenAI's GPT-4, to analyze the emotional content communicated by Russian and English speakers. The study leverages a dataset from Wikipedia, utilizing translation, high-level question-answering, and emotional analysis methods to prioritize articles based on their emotional impact. A key insight is the utilization of the LLM itself to create these questions, thereby showcasing the LLM's capability to understand and apply each technique. To refine the vast array of generated questions to a manageable, subset that aligns to human persuasions, we establish a threshold for question quality using an existing propaganda dataset as a benchmark.

Employing a multilingual Question-Answering (QA) approach, the questions are used to summarize emotional responses, revealing significant differences in the emotional content between languages and subjects. Neutral topics typically elicit less emotional reaction for both English and Russian, while political topics generate a range of emotions influenced by cultural relevance. We further show that our approach works just as effectively in either language, as ranking between original and translated articles yield similar outcomes.

The paper concludes by offering a dataset of over 225,000 examples of country-level propaganda (see Appendix B), designed to aid future researchers in refining their questioning techniques for propaganda analysis. This resource aims to enhance the understanding of emotional content across languages and contribute to the advancement of emotional analysis in multilingual settings.

As for future research, we plan to extend our exploration of propaganda techniques. Additional propaganda annotations should facilitate more sophisticated analyses that consider not just overall emotion, but also specific sentiment and employed propaganda techniques.

ACKNOWLEDGMENTS

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

REFERENCES

[1] Moore, M. and Colley, T., "Two international propaganda models: Comparing rt and cgtn's 2020 us election coverage," *Journalism Practice* , 1–23 (2022).

[2] Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., and Tucker, J. A., "Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 us presidential election," *The International Journal of Press/Politics* **25**(3), 357–389 (2020).

[3] Geissler, D., Bär, D., Pröllochs, N., and Feuerriegel, S., "Russian propaganda on social media during the 2022 invasion of ukraine," *EPJ Data Science* **12**(1), 35 (2023).

[4] Barham, S., Weller, O., Yuan, M., Murray, K., Yarmohammadi, M., Jiang, Z., Vashishtha, S., Martin, A., Liu, A., White, A. S., et al., "Megawika: Millions of reports and their sources across 50 diverse languages," *arXiv preprint arXiv:2307.07049* (2023).

[5] Li, B. and Callison-Burch, C., "Paxqa: Generating cross-lingual question answering examples at training scale," *ArXiv* **abs/2304.12206** (2023).

[6] Huang, K.-H., McKeown, K., Nakov, P., Choi, Y., and Ji, H., "Faking fake news for real fake news detection: Propaganda-loaded training data generation," *arXiv preprint arXiv:2203.05386* (2022).

[7] Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y., "Framing and agenda-setting in russian news: a computational analysis of intricate political strategies," *arXiv preprint arXiv:1808.09386* (2018).

[8] Park, C. Y., Mendelsohn, J., Field, A., and Tsvetkov, Y., "Challenges and opportunities in information manipulation detection: An examination of wartime russian media," *Findings of the Association for Computational Linguistics: EMNLP 2022* , 5209–5235 (2022).

[9] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y., "Truth of varying shades: Analyzing language in fake news and political fact-checking," in [*Proceedings of the 2017 conference on empirical methods in natural language processing*], 2931–2937 (2017).

[10] Da San Martino, G., Barron-Cedeno, A., and Nakov, P., "Findings of the nlp4if-2019 shared task on fine-grained propaganda detection," in [*Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda*], 162–170 (2019).

[11] Li, J., Ye, Z., and Xiao, L., "Detection of propaganda using logistic regression," in [*Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*], 119–124 (2019).

[12] Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R., and Nakov, P., "Semeval-2020 task 11: Detection of propaganda techniques in news articles," *arXiv preprint arXiv:2009.02696* (2020).

[13] Piskorski, J., Stefanovitch, N., Da San Martino, G., and Nakov, P., "Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup," in [*Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*], 2343–2361 (2023).

[14] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y., "Defending against neural fake news," *Advances in neural information processing systems* **32** (2019).

[15] Lawrence, J. and Reed, C., "Argument mining: A survey," *Computational Linguistics* **45**(4), 765–818 (2020).

[16] Farzam, A., Shekhar, S., Mehlhaff, I., and Morucci, M., "Multi-task learning improves performance in deep argument mining models," *arXiv preprint arXiv:2307.01401* (2023).

[17] Sourati, Z., Venkatesh, V. P. P., Deshpande, D., Rawlani, H., Ilievski, F., Sandlin, H.-Â., and Mermoud, A., "Robust and explainable identification of logical fallacies in natural language arguments," *Knowledge-Based Systems* **266**, 110418 (2023).

[18] Durmus, E. and Cardie, C., "A corpus for modeling user and language effects in argumentation on online debating," *arXiv preprint arXiv:1906.11310* (2019).

[19] Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B., "The argument reasoning comprehension task: Identification and reconstruction of implicit warrants," *arXiv preprint arXiv:1708.01425* (2017).

[20] Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., and Schölkopf, B., "Logical fallacy detection," *arXiv preprint arXiv:2202.13758* (2022).

[21] Pauli, A., Derczynski, L., and Assent, I., "Modelling persuasion through misuse of rhetorical appeals," in [*Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*], 89–100 (2022).

[22] Karpinska, M. and Iyyer, M., "Large language models effectively leverage document-level context for literary translation, but critical errors persist," *arXiv preprint arXiv:2304.03245* (2023).

[23] Webber, W., Moffat, A., and Zobel, J., "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems (TOIS)* **28**(4), 1–38 (2010).

[24] Zhu, A., Hwang, A., Dugan, L., and Callison-Burch, C., "Fanoutqa: Multi-hop, multi-document question answering for large language models," *arXiv preprint arXiv:2402.14116* (2024).

[25] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., "Language models are few-shot learners," *Advances in neural information processing systems* **33**, 1877–1901 (2020).

APPENDIX A. USING LLAMA FOR EMOTIONAL CONTENT DETECTION

For comparing extracted emotional language in Russian vs. English Wikipedia (Section 8), we also ran a small study using the Llama 2 model.[*] We chose this experiment because it decomposes the larger task into 3 simpler subtasks, which a smaller LLM may handle reasonably. We do expect some performance drop, given the order of magnitude difference in size – 13B vs <1T[†]. Furthermore, given the closed-source nature of GPT-4, a locally-run, open-source model allows for more direct insights and analysis, especially for future work (and classified or proprietary datasets).

As before, we report results for RU2EN only, but have run the steps for all 4 settings. For ease of analysis and our computational budget, we restrict study to a 217 article subset of the original 22,046.

[*] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[†] GPT-4 has been suspected to be over 1T parameters

Implementation details We implement these experiments with the Kani [24] package, which allows for fast development and iteration. Notably, to verify our re-implementation of the prompts taken with GPT-4, we first use the internal Kani class for GPT-4; then easily swap the underlying model to the Llama class in a few lines of code. We further implement on our own batch processing (currently unsupported in Kani) using the `huggingface` package.

Takeaways Overall, we found that several techniques were required to get Llama to adhere to the expected output format: *few-shot* examples, and *pre-generating* the starting tokens of a response. From our analysis of this output, we observed that vs. the gold annotators, the Llama-based approach identified even more persuasion techniques than GPT-4 did. We therefore leave future work to investigate this further. We discuss the specifics and the comparative findings for each step ahead. For future work, we emphasize the importance of these two techniques to elicit proper instruction-following ability from open-source LLMs.

A.1 Step 1: Applying questions

Recall the system prompt from Section 7.2; while GPT-4 successfully adheres to the system prompt, the same cannot be said of Llama. The same zero-shot prompt used for GPT-4 does not work out of the box for Llama. We find that Llama has poor instruction-following capability; for example, for step 1, instead of the proper format of "Q1: True (conf:70); Q2: False", zero-shot prompting Llama will give short answer responses, with the model explaining each decision, and giving substantiating evidence. Thus, Llama's output does not match the expected output format. We use few-shot examples and pre-generation to go from 0% to 90% parsable output.

Few-shot *Few-shot learning* is the process of including example inputs and expected outputs in the prompt [25]. In fact, it was the original paradigm for interacting with LLMs, and it was not until later LLMs developments that led to solid zero-shot, instruction-only abilities.

For step 1, we manually curate 3-shot examples, collecting the contexts and writing outputs ourselves. The contexts are copy-edited paragraph-long excerpts from Wikipedia: we select a political article where most answers are True (on Augusto Pinochet), a scientific article with a handful of emotional terms (on Cobalt), and a scientific article with no emotional terms (on Bananas).

We tested the effect of increasing the number of examples over the responses over 2 articles ($\approx$50 chunks). 0-shot prompts nearly never ($<$5%) output parsable responses, 1-shot prompts improve to $\approx$25%, and 2-shot and 3-shot prompts achieve to 45-50%.

While above we describe an English setting, for the 2 Russian settings, we use a machine translate and verify approach on the few-shot examples.

Pregeneration Considering the 50% of properly formatted responses, we observe that they are always prefixed with `Q1:`; in contrast, unparsable responses nearly always have different prefixes. This leads to the intuition that we can *pre-generate* the proper prefix by concatenating it to the input sequence. Afterwards, the model will continue generations in this modified distribution space; we found that with pre-generation (and few-shot), instruction-following improves to $>$90%.

Mathematically-speaking, consider the language modeling task as given a sequence of words $x_1^n = x_1, ..., x_n$, predict the likelihood distribution over the next word $y_1 - P(y_1|x_1^n)$. Pre-generation is equivalent to setting the probabilities for $y_1, ..., y_m$ to 1 for each prefix token (and 0 elsewhere), then continuing language modeling as normal for y_{m+1}. This is a simple albeit effective technique; we are not aware of prior work that formalizes pre-generation, as we have done, but similar techniques have been used in the prompt engineering literature; for example, by concatenating the prefix "Answer: " for QA tasks.

Step 1 error analysis Depsite the correct outputs, as for the actual task, Llama seems to underperform GPT-4. We characterize this by observing for Llama: it mostly outputs True, has much higher confidence scores (most are 95-100), and gives answers out of order (e.g. `Q0...Q1...Q9...Q4...`).

A.2 Step 2: Extract text spans

As with step 1, we observed that the technique of few-shot prompts and pre-generation were required for the model response format to be parsable. For few-shot prompts, we use the same contexts, and write our few-shot examples for all questions. For pre-generation, we set the prefix to be a single quotation mark ".

Table 11 compares model responses (same example as Table 7). The bolded spans indicate where the LLMs' concur. From this example, we see while the models are largely making differing decisions, the task itself is quite subjective, and the spans are semantically related. Anecdotally speaking, we observe that GPT-4 is able to extract longer clauses (though not demonstrated in this example), while Llama can only extract short phrases.

Table 11: GPT-4 (black) and Llama (blue) responses to two questions for the article "Fire at the National Museum of Brazil"

Section	QID	Extracted Spans
2	Q20	'engulfed', 'rapidly destroyed', '**tragedy**', 'repeatedly complained', ... 'negligence', '**tragedy**', 'could have been avoided'
2	Q88	'fire engulfed', 'rapidly destroyed', '**tragedy**', 'funding cuts', 'deteriorating', ... 'negligence', '**tragedy**', 'could have been avoided'
3	Q20	'**incalculable**', 'outraged', 'cultural tragedy', '**lobotomy**' 'cultural tragedy', '"**incalculable**" loss', '**lobotomy** of Brazilian memory', ...
3	Q88	'fire', 'loss', 'outraged', '**tragedy**', 'destroyed', 'ruins', 'threat', ... 'cultural **tragedy**', '"incalculable" loss', 'lobotomy of Brazilian memory', ...

A.3 Step 3: Rank articles by Emotional Content

For the 217 article subset considered for the Llama experiments, we ran the same ranking procedure as was done for GPT-4 on the full set. Recall that for GPT-4, Table 8 shows that of the top articles for either language, most follow our intuition – the top articles in Russia concerned Russian figures and more relevant events to Russia (e.g. Ukraine conflict, Stepan Bandera), while the top articles for English concern Western figures and more relevant events to the USA (e.g. 2014 Gaza war, Madonna). As for ranking with Llama, we found that the top articles in this way were overall still intuitive, but there were more neutral-sounding topics in the list as well. This worse performance led us to stick with GPT-4 for the main analysis, though we have shown the promise in using additional techniques to enable far smaller, open-source LLMs, to achieve reasonable performance.

APPENDIX B. GENERATING SYNTHETIC PROPAGANDA

Recall that despite the quality of the SemEval dataset, and other datasets we reviewed, the coverage of persuasion techniques is highly imbalanced, due to real-world usage in the wild. As we found in our experiments above, this imbalance lends itself towards our own study, where 9 of the 12 top-ranked HLQs were related to the top class, Loaded Language. The other persuasion techniques have been underrepresented, and therefore, we are motivated to release a synthetic dataset which covers all 23 SemEval techniques in a balanced way (≈10k each).

Our focus is on nation-state propaganda. An added advantage of our dataset is that it includes a repository of propaganda examples shaped on a DIME-like ontology for factors that a country wants to promote (DIME stands for diplomatic, informational, military, and economic factors). This guarantees a broad coverage of topics.

Example DIME like high-level categories: Judiciary, Military, Healthcare System, Education System, Infrastructure, Financial Institutions, Communication Networks, Technological Advancements, Educational Excellence, Trade Policies, Human Rights Record, and others[‡] .

For each high-level category, this is the query prompt for generating components within category: produce a list of components making up {Country}'s {category}.

[‡] categories are based on GPT-4 query with 'items a nation-state might wish to promote to gain advantage over other nation-states'

For example for country 'Russia' and high-level category 'Military' LLM produces: Russian Ground Forces, Russian Aerospace Forces, Russian Navy, Strategic Missile Troops, Russian Airborne Troops, Federal Security Service (FSB), ..., Russian Electronic Warfare Troops, Russian Engineer Troops, Russian NBC Protection Troops, Russian Signal Troops, Russian Rear of the Armed Forces, Russian Logistics Support.

We can further refine this ontology by requesting subcomponents of each component from above.

We've found that using the term 'institutions' instead of 'components' often leads to the generation of educational institutions. In a military context, this would imply military academies and universities, not military organizations or units. Thus, proper keywords and prompt engineering is important in this context.

This DIME-like ontology that is specific to each considered country is used to generate synthetic examples. The query to GPT-4 has this structure (where Action can be 'promote' or 'minimize'):

> **System Prompt:** You are a helpful assistant with a background in political science, economics, and international relations.
>
> **User Prompt:** Produce a paragraph that {Action} {Country}'s {Component} using {Propaganda Technique}: {Technique Definition}

At this point the dataset has examples for top 20 countries based on population. The dataset is a JSON file containing 225092 items. Some queries fail to perform specified action, i.e. can actually be promoting vs. minimizing. GPT-4 might find some queries offensive and refuse to generate a response. Thus additional verification is required over this dataset.

The existing literature details numerous propaganda techniques, but until now, it hasn't been possible to create a dataset that encompasses all of them. Our goal is to continue expanding this dataset[§].

[§] dataset and updates will be available at: https://github.com/apanasyu/UNCOVER_SPIE