Improving GPT Generated Synthetic Samples with Sampling-Permutation Algorithm*

Kirill Kalinin†

*Prepared for the 40th Annual meeting of the Society for Political Methodology, Stanford, July 9-11, 2023.

†Researcher at the Hoover Institution, Stanford University (E-mail: kkalinin@stanford.edu)

Abstract

The primary objective of this study is to leverage the capabilities of a large language model (LLM), such as GPT-3, to generate responses from elite individuals who are difficult to access. Similar to the work of Argyle et al. (2023), this study specifically focuses on the domain of multiple-choice questions. To address the issue of instability and hallucinations commonly associated with LLM, a novel algorithm, termed the "sampling-permutation algorithm", has been developed and implemented. The efficacy of this algorithm is assessed by applying it to questions from the *Survey of Russian Elites*(Zimmerman, Rivera and Kalinin 2022). Notably, this study examines the generated responses from synthetic personas representing the Russian President Vladimir Putin and the opposition leader Alexei Navalny by conducting a validation study and exploring the effects of the war context on generated responses. My findings indicate that the proposed approach provides valuable insights, despite the presence of somewhat mixed results.

Keywords: GPT-3, large language model, Survey of Russian Elites, synthetic data.

Introduction

This study introduces a novel "sampling-permutation algorithm" for generating multiple-choice responses from the language model GPT-3 (Brown et al. 2020). The algorithm exhibits promising potential for application in academic research and policy analysis, offering a solution to address the challenges associated with instability and hallucinations that are often encountered in large language models (LLMs).

The algorithm is put to the test in generating responses from hard-to-reach members of the Russian elite, covering a wide range of topics including politics, economics, and culture. The proposed framework builds upon the work of Argyle et al. (2023) and leverages conditional probabilities of tokens in data simulation to ensure more reliable and contextually appropriate responses.

We anticipate the following criteria to be met when generating survey responses using the LLM:

a) The generated responses should demonstrate robustness when faced with semantically similar questions. In other words, the algorithm should be able to provide consistent and coherent answers even when the questions are expressed differently but have the same underlying meaning.

b) The responses should not be influenced by the different orderings of response options. Regardless of the arrangement of the choices, the system should provide consistent answers that are not affected by the positioning or presentation of the options.

c) To ensure the quality and reliability of the generated responses, it is essential to validate them using external data sources or other LLMs that have been trained or fine-tuned on different datasets. This external validation provides an additional measure of confidence in the accuracy and relevance of the generated data.

Guided by the specified criteria, this research presents three distinct code implementations: a) single-factor data generation for closed-ended questions, which entails the creation of a set of prompts and a full set of permuted multiple-choice responses for each question;

b) multi-factor data generation for closed-ended questions utilizing multi-factor crosstabs to build prompts for subgroups based on multiple socio-demographic factors; and c) data generation for open-ended questions that resembles single-factor data generation but is adapted to the format of open-ended questions.

The incorporation of multiple code implementations enables a comprehensive exploration of various data generation scenarios and enhances the research's ability to meet the stipulated criteria. By focusing on both closed-ended and open-ended questions, this research seeks to provide a well-rounded analysis of the language model's performance and reliability in generating responses across different contexts and question formats.

The empirical part of this paper primarily concentrates on the single-factor generation of data for closed-ended questions. Specifically, it focuses on data generation of responses for "Vladimir Putin" and "Alexei Navalny", as well as state and non-state elite members. By adopting this approach, the study aims to ensure consistency and coherence in the generated responses for different synthetic politicians and elite groups. Discussion of other listed approaches can be found in the Appendix.

To generate responses for closed-ended questions, the *text-davinci-003* model is utilized as the core tool for data generation. As a component of the validation process, I incorporated questions from a questionnaire in conjunction with data from the "Survey of Russian Elites" (Zimmerman, Rivera and Kalinin 2022). This study also emphasizes an analysis of contextual effects aimed at simulating diverse responses from synthetic politicians, contingent upon the war context surrounding the Russian-Ukrainian conflict.

The paper is structured as follows: Section 1 provides an overview of the GPT-3 language model and proposed sampling-permutation algorithm. Section 2 delves into the empirical strategy employed for generating responses using a single-factor generation approach. Finally, Section 3 presents main findings of the validation and context-effects analysis. The concluding section summarizes the main findings of this research.

Theory

Overview of the GPT-3 model

The GPT-3, or "generative pre-trained transformer", is an advanced language model developed by OpenAI that boasts 175 billion parameters and has been trained on a vast and diverse corpus of texts totaling 570 gigabytes. The GPT-3 is a standard autoregressive decoder-only language model which given a prompt $x_{1:i}$ produces both contextual embeddings and a distribution over next tokens x_{i+1}, such that $x_{1:i} \Rightarrow \phi(x_{1:i}), p(x_{i+1} \mid x_{1:i})$ (Liang et al. 2022). Simply speaking, instead of looking for the perfect solution each time, the model tries to find the best probabilistic match in the data set on which it has been trained.

In recent literature, there have been attempts to assess the performance of GPT-3 in various domains. One study by Argyle et al. (2023) proposes the use of GPT-3 as proxies for specific human subpopulations in social science research. The authors condition the model on thousands of sociodemographic backstories from real human participants in multiple large surveys conducted in the United States and demonstrate that GPT-3 can closely replicate human responses. In another paper, Kalinin (2023*a*) utilizes GPT-3 generated responses for geopolitical forecasting related to the Russia-Ukraine war. Furthermore, Bommarito and Katz (2022) evaluate the performance of GPT-3 on the NCBE MBE practice exam through an experimental study. Other examples of the application of GPT-3 can be found in OpenAI (2023)'s report. These studies highlight the potential of GPT-3 in various domains and its ability to generate responses that are comparable to those of humans.

Recent work shows that large language models can perform few-shot learning without fine-tuning (Radford et al. 2019), suggesting "in-context" learning can be effectively used without additional fine-tuning parameter updates. And, more importantly, it has numerous practical advantages over the fine-tuning (Radford et al. 2019; Devlin et al. 2019) by allowing to "rapidly prototype" NLP models, simplifying access to users without technical expertise, and reusing the same model for each task. The earlier work by Manakul, Liusie and Gales

(2023), shows that depending on the prompt's format and training examples the accuracy can change from near chance to near state-of-the-art. This instability implies that GPT-3 users, who typically design prompts manually, cannot expect to consistently obtain good accuracy.

This study operates under certain set of assumptions that GPT-3-generated responses must satisfy. The study assumes that GPT-3 has knowledge of potential responses from certain members or subgroups of the elite, resulting in sampled responses that are likely to be similar and contain consistent information. Second, the quality of the generated output depends on the quality of training data: non-relevant data might produce biased generated responses. Third, the model's generation of the most probable responses makes predicting strategic behavior in responses to survey questions quite a challenging task.

When provided a zero-shot prompt for multiple-choice questions GPT-3 can learn to generate letter choice responses to multiple choice questions by identifying the type of question and its semantic meaning. However, due to its autoregressive nature, GPT-3 primarily focuses on the left context when generating predictions. Consequently, information extraction from the model can exhibit instability and may heavily rely on factors such as the format of the prompt and the order in which multiple choice options are presented. This characteristic introduces a potential challenge in ensuring consistent and reliable outputs from the model. Another issue that can arise is LLM hallucination, which occurs when the generated output is nonsensical and does not align with the given information (Ji et al. 2023). LLM hallucination can complicate information retrieval using multiple-choice questions. Researchers and practitioners should be aware of these issues when utilizing GPT-3 for tasks involving accurate information retrieval, and by carefully considering prompt design it is possible to mitigate the impact of this limitation and improve the overall model's performance.

The proposed sampling-permutation algorithm effectively mitigates the problem of instability and hallucinations encountered during the data-generating process by identifying such occurrences as abnormal outliers. This approach not only ensures more reliable and

accurate results but also provides measures of uncertainty that contribute to a deeper understanding of the level of entropy associated with the data-generating process for closed-ended questions. By acknowledging and accounting for these phenomena as outliers, the algorithm enhances the reliability and accuracy of the generated responses, thereby contributing to the overall integrity of the findings.

The GPT-3 model outputs a *log probability* for every known token for both prompts and completions. The reason for this is that it is computationally easier to compute the probability of a sequence of tokens if individual probabilities are expressed in *log probabilities* rather than in probabilities or percentages. Therefore to convert log probabilities to probabilities, we use use the following formula: $prob(x) = 100 \times e^{logprob(x)}$. These probabilities $prob(x)$ can be manipulated using two important parameters that control the randomness of generated response: *temperature* and *Top P* (OpenAI 2022). *Temperature* and *Top P*, sometimes called the "creativity dials", because these parameters control the amount of creativity in response generation. *Temperature* takes a value between 0 and 1: at 0, randomness is removed by boosting the most likely token to 100%. *Top P* ranges from 0 to 1 and controls how many random results the model should consider for completion; it determines the scope of randomness defined by temperature dial.

Other important parameters used for GPT-3 text generation are as follows: *engine* is set to *text-davinci-003* (one of the most powerful and expensive GPT-3's execution engines), *max_tokens* is set to 1 (the maximum number of tokens to generate in the completion is 1, since we only need one letter choice as the answer to a particular question); *logprobs* is set to 10 (the list of log probabilities for 10 most likely tokens). Finally, both *presence_penalty* and *frequency_penalty* are used to penalize new tokens based on their existing frequency in the text by increasing or decreasing the model's likelihood to generate text about new topics or the model's likelihood to repeat the same line verbatim (both parameters are set to their default value 0).

Sampling-Permutation Algorithm for Survey Data Generation

The main premise of this research is centered around the possibility of accurately recovering a true quantity, denoted as θ, through the utilization of prompt sampling and permutations for closed-ended questions in surveys. The objective is to achieve convergence towards the true value by combining these methods in the LLM setting. In this context, let $\bar{X}$ represent the sample mean, θ denote the ground truth, ϵ represent a small positive number indicating the desired level of accuracy, and $\hat{\theta}$ symbolize the estimated value of θ derived from the sample mean. The aim is to establish a scenario where according to LLN the probability tends to zero as the sample size approaches infinity, denoted as $\lim_{n\to\infty} P(|\bar{X} - \theta| > \epsilon) = 0$.

The sampling-permutation algorithm, which has been adapted for utilization in large language models for the purpose of information extraction, can be outlined as follows.

Let $Q_0 = p_0^\frown O_0$ represent the original multiple-choice question, consisting of two concatenated parts: p_0 – the original prompt containing the stem (question or problem), and O_0 – a set of multiple-choice options $\{o_1, o_2, ..., o_n\}$. Using p_0 as an original user query, we can generate a set of semantically identical LLM responses, denoted as $P = \{p_1, p_2, ..., p_l\}$. The LLM's "creativity dials", such as *temperature (t = 0.8)* and *top P ($t_p = 1$)*, are adjusted to produce variations of prompts. Next, given the original set O_0, we obtain a set of new permuted sets, such as $O = \{O_1, O_2, ..., O_{n!}\}$. To generate a set of questions G with permuted options, we concatenate the sets P and O, resulting in set G of size $l \cdot n!$ with $P^\frown O$ questions.

During the computing stage, each question g from a set G is inputted into the generative LLM using OpenAI's completion endpoint. As a result, the LLM generates a single-letter response (represented as a one-letter token) with the highest probability (measured on a logarithmic scale), denoted as $\hat{t} = \arg\max_{g\in G} LLM(g)$. Consequently, for a set of generated questions $G = \{g_1, g_2, ..., g_n\}$, I obtain a set of most probable letter responses $L = \{l_1, l_2, ..., l_n\}$ and a corresponding set of token probabilities for each letter response $T = \{\hat{t}_1, \hat{t}_2, ..., \hat{t}_n\}$. Based on these token probabilities, two quantities of interest are com-

puted. First, I calculate the mean token probability for each winning option category, denoted as c: $\bar{T}_{l=c} = \frac{\sum_{i=1}^{n_{l=c}} \hat{t}_{l=c}}{N_{l=c}}$. By calculating the average score $\bar{T}_c$ for each letter response, we can effectively estimate the true probability of support for each option with reduced noise. Second, the uncertainty estimate, standard deviation σ of option-wise token probabilities, is calculated: $\sigma_{l=c} = \sqrt{\frac{\sum(\hat{t}_{l=c} - \bar{T}_{l=c})^2}{N_{l=c}}}$. Here, $N_{l=c}$ denotes the total count of occurrences where the letter response l matches the category c.

The average probability of a token, denoted as $\bar{T}_c$, can be interpreted as the average text-based generated support for a specific option. This interpretation allows for a meaningful comparison with the survey proportions, facilitating further analysis and evaluation of the findings.

The proposed algorithm, which utilizes token probabilities for estimation, can be extended to incorporate quantities generated directly by the LLM(Lin, Hilton and Evans 2022). This enhancement allows for a broader application of the algorithm beyond token probabilities.

Hence, for each multiple-choice question the algorithm produces two quantities of interest: the average probability of the letter response across all generated prompts and permuted options and the standard deviation of the probability of the letter response.

The permutation algorithm employed in this study entails submitting a query to OpenAI's API for each permuted question. Consequently, the computational cost increases exponentially as the number of multiple choice questions expands. For instance, a question with 2 options results in 2 question-permutations, while 3 options yield 6 question-permutations. This pattern continues with 4 options leading to 24 question-permutations, 5 options resulting in 120 question-permutations, 6 options giving 720 question-permutations, 7 options involving 5040 question-permutations, and 8 options accumulating a staggering 40,320 question-permutations. Hence, we should be mindful of the associated costs and ideally restrict the number of options to a maximum of 3-4.

Empirical Strategy

Data, Prompts, Hypotheses

In this paper I will use the questions from the *Survey of Russian Elites*, which covers the period 1993–2020 (Zimmerman, Rivera and Kalinin 2022). The *Survey* includes members of the Moscow elites working in the major public and private sectors of Russian society. Specifically, the interviews were conducted with high-ranking individuals employed in the media, state-owned enterprises, private businesses, academic institutions with strong international connections, as well as the executive branch, the federal legislature, the armed forces and security agencies.

The data include a wide range of questions related to Russia's national interests, U.S.-Russian relations, the role of military force in international relations, the greatest threats to stability and security, Russia's relations with other countries (e.g., the United States, Ukraine, Belarus, Japan and China), the enlargement of the European Union, NATO expansion, Russia's civilizational path and many other questions related to the international and domestic agenda.

For the purpose of illustrating the functionality of the single-factor data generation algorithm in this particular version of the paper, I selected 129 questions from the 2020 *Survey*. These questions were chosen based on their relevance to the analysis and were subsequently modified to improve clarity and simplicity. Furthermore, placeholders were included in each question to allow for the automated insertion of specific time periods, names, or concepts.

In this study I am mostly interested in generation of responses from two synthetic politicians: Russian President "Vladimir Putin" and Russian opposition leader "Alexei Navalny". As part of the validation process, I utilized data from the "Survey of Russian Elites" (Zimmerman, Rivera and Kalinin 2022) and compared it with the generated responses attributed to state and non-state elite members. The *Davinci* (*text-davinci-003*) model is the central model used for generating responses to closed-ended questions.

The division of elites into two separate categories is based on the rationale developed by Noah Buckley and Joshua Tucker. The aim is to identify members of elites whose views can be closest to either "Vladimir Putin's" or "Alexei Navalny's". Those members of elites working in the executive or legislative branches, the military, or security agencies are classified as "core" or "government" elites, whereas those employed in the media, science and education fields, state-owned enterprises, or private business are "non-core" or non-government elites (Buckley and Tucker 2019). This categorization is intended to partially validate responses generated for Vladimir Putin and Alexei Navalny, with the former's position closely related to that of the government elites and the latter's position closely related to that of the non-government elites.

Consequently, my **validation hypothesis** posits that Vladimir Putin's responses are anticipated to align more closely with state elites, while Alexei Navalny's responses are expected to resonate more closely with non-state elites.

Recall that the probability percentages of responses generated by the *Davinci* model are intended to represent the probabilities of multiple choices across different permutations, so they are not normalized and do not add up to one. The resulting standard deviations of the probabilities for all permutations are used to construct 95% confidence intervals where possible.

The questions for "Vladimir Putin"/state elites and "Alexei Navalny"/nonstate elites are preceded by additional contextual information. First, I generate responses for 2020, which align with the data the *Davinci* model has been trained on, aiming to validate GPT-3 generated responses based on the collected *Survey* data. Second, I also generate data for 2022 responses that the model has not encountered before. This exercise serves to demonstrate the model's capabilities in extrapolating learned information onto the future, particularly within the context of war. The goal is to explore how the GPT-3 model handles novel situations and to assess its performance in unseen scenarios.

The context experiment is based on the concept of the "rally 'round the flag" effect first

explored in U.S. foreign policy crises is explored in Mueller (1985) that refers to a phenomenon observed in politics and international relations where public support for a country's leader or government increases significantly during times of crisis or conflict. This effect suggests that during periods of national threat, such as military conflicts or significant international challenges, people tend to set aside their political differences and unite behind their leaders. This effect can bolster a leader's approval ratings and fortify their authority, often fostering a sentiment of national unity. In accordance with my **war effect hypothesis**, I anticipate that in comparison to 2020, the alignment between "Putin" and state elites, and to a lesser extent between "Putin" and non-state elites, will intensify in 2022. Conversely, for "Navalny", the association between state and non-state elites should tend to weaken.

From a technical standpoint, this war effect experiment for 2020 and 2022 serves the purpose of evaluating GPT-3's capabilities in gauging the context's effects on responses pertaining to "Putin" and "Navalny". The study aims to glean insights into the variations in GPT-3's responses based on the context it encounters and how this context-based priming might affect the two synthetic politicians across diverse policy domains. This part illuminates the significance of comprehending and managing priming when utilizing AI language models like GPT-3, as it underscores the potential to manipulate generated outputs based on the provided context.

For example, for 2020 and 2022 the prompts are as follows:

In 2012 $\underbrace{Placeholder}_{Person=\{VladimirPutin,AlexeiNavalny\}}$ thinks that $\underbrace{A.Option1; B.Option2; C.Option3.}_{Permutations:\{ABC\},\{BAC\},\{CAB\},\{ACB\},\{ACB\},\{BCA\},\{CBA\}}$.
In 2022, Russia's invasion of Ukraine intensified the Russo-Ukrainian War, causing mass casualties and destruction. This led to international sanctions and Russia's isolation. Domestically, Russia prioritized a military economy and effectively quashed political opposition to rally support for the war. In 2022 $\underbrace{Placeholder}_{Person=\{VladimirPutin,AlexeiNavalny\}}$ thinks that $\underbrace{A.Option1; B.Option2; C.Option3.}_{Permutations:\{ABC\},\{BAC\},\{CAB\},\{ACB\},\{ACB\},\{BCA\},\{CBA\}}$.

Another important area of discussion is related to questions for which responses can change during the times of crisis: we can expect that depending on the type of question people specific segments of elites would feel different about different questions. For instance, some authors acknowledge that during a crisis, the more fundamental the threatened values are, the greater the perceived likelihood of war; second, the more dynamic, potent, and fundamental the values under threat are, the longer the perceived time for response will be; third, when time pressure becomes more acute, there is an elevated perception of the likelihood of war and a heightened intensity in the perception of the threat (Brecher and Wilkenfeld 1997). Consequently, according to my **survey questions hypothesis**, it is reasonable to anticipate that survey questions concerning war, values, and perceived threats will be the ones most significantly influenced by the war context.

The Python script that implements automated generation of responses using the GPT-3 model via single-factor data generation algorithm is available on **GitHub**. For simplification purposes, one needs to fill out only the spreadsheet and run the Python code (see all details in the Appendix).

The code was designed to utilize the LLM for generating a collection of three question stems with identical semantic content. This process involves the original prompt along with the three stems, resulting in a list of four stem variants. Subsequently, for each variant of a question stem, the algorithm facilitates the generation of a complete set of permuted options, which are then combined with the list of stems. The full list of concatenated sampled stems and permuted options is used in API queries. Per each API query an algorithm extracts the option with the highest probability and calculates option-wise statistics, such as the mean and standard deviation. The rationale behind focusing solely on the most probable choices is to ensure that the resulting tokens are sensible, given that the options with lower probabilities could be nonsensical. Moreover, certain options may never be chosen by the model and thus are disregarded in the output.

A single API query provides normalized probabilities for generated responses for each

permutation, which sum up to one. However, option-wise aggregate estimates for all permutations do not sum up to one, and thus normalization of the resulting quantities of interest is necessary to ensure consistency. Presently, the script does not implement such normalization. Although computing permutations of questions can be computationally expensive, an increase in the number of permutations can increase confidence in the results and help assess the amount of relevant information in the LLM. Conversely, when the number of options and permutations is limited to two, we may have less confidence in the generated results.

Findings

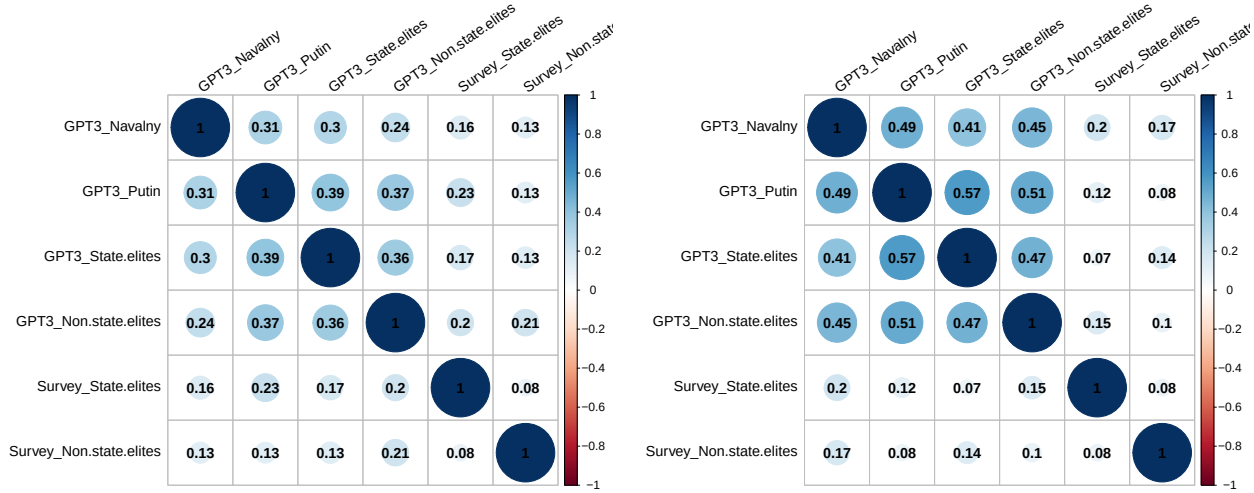
Using the single-factor approach, I implemented the generation of responses from synthetic politicians for the Russian President, Vladimir Putin, and the Russian opposition leader, Alexei Navalny, as well as state and non-state elites for both 2020 (i.e., when the last survey wave was collected) and 2022 when the Russia-Ukraine war occurred. The generated data results are available on **GitHub**: the results for each question contain the mean estimate of token probabilities and confidence intervals for the chosen options.

Validation Experiment

The strength of association, as measured by *Cramer's V*, between the survey and generated letter responses is indicative of complex patterns. Each subfigure in Figure 1 contains associations for six categories: *GPT-3 Putin*, *GPT-3 Navalny*, *GPT-3 state elites*, *GPT-3 non-state elites*, as well as *Survey state elites*, and *Survey non-state elites*. The first four categories represent GPT-3-generated results, while the last two categories draw information from the *Survey of Russian Elites*, 2020 (Zimmerman, Rivera and Kalinin 2022).

In 2020, the *Cramer's V* values indicate a moderate level of association between GPT-3-generated responses attributed to "Putin" and GPT-3-generated responses concerning both types of elites. However, upon comparing the GPT-3-generated responses for "Putin" with

Figure 1: *Cramer's V*



2020 2022

Notes: Figures (a) *Cramer's V* for responses generated for 2020 and (b) *Cramer's V* for responses generated for 2022. *Cramer's V* is a measure of association used to quantify the strength of the relationship between two categorical variables. It ranges from 0 to 1, where 0 indicates no association between the variables, and 1 indicates a perfect association.

the survey results for both types of elites, the associations become notably weaker, failing to offer compelling evidence in support of the validation hypothesis. Furthermore, for "Navalny", these associations are found to be even less substantial compared to those for "Putin".

In comparison to the year 2020, the *Cramer's V* associations in 2022 have displayed an increase in GPT-3 generated results, leading to a transition from weak values to moderate or strong values. This observation is not limited solely to Putin, but also encompasses Navalny and their respective associated elites. This finding demonstrates that the context of war prompts the LLM to generate responses characterized by increased mobilization tendencies and more distinct alignment patterns between leadership and the corresponding elite factions.

Lastly, concerning the validation study, the associations between the generated responses and the 2020 *Survey* results consistently exhibit weak findings, with a marginal decrease

in associations observed for "Putin" and an increase for "Navalny" in 2022 as compared to 2020. This implies that the LLM continues to face a significant challenge in accurately capturing the nuanced intricacies of the survey data. This difficulty is further accentuated by the absence of negative associations between responses from 'synthetic' politicians.

In sum, even though the *Cramer's V* analysis for both 2020 and 2022 suggests limited support for the validation of GPT-3-generated data, the increase in associations in the 2022 matrix highlight the model's ability to adapt and generate contextually relevant responses.

Table 1 presents the *Cramer's V* measures discussed earlier, along with accuracy measurements and correlations among the groups of interest. The table highlights the complex associations between the GPT-3 model's generated responses for the synthetic politicians and their respective associated groups. In the context of 2020, the accuracy measures span a range of 0.30 to 0.60, while for 2022, they exhibit variability from 0.26 to 0.71. It's noteworthy that these observed accuracy measurements appear to align with the trends seen in *Cramer's V*.

Table 1: Validation Study

	2020			2022		
Variable	ϕ_c	a	ρ	ϕ_c	a	ρ
GPT3_Putin vs. GPT3_State elites	0.39	0.60	0.01	0.57	0.71	0.37
GPT3_Putin vs. GPT3_Non-state elites	0.37	0.57	0.23	0.51	0.66	0.46
GPT3_Putin vs. Survey_State elites	0.23	0.32	-0.05	0.12	0.35	0.21
GPT3_Putin vs. Survey_Non-state elites	0.13	0.30	-0.18	0.08	0.36	0.15
GPT3_Navalny vs. GPT3_State elites	0.30	0.50	0.20	0.41	0.61	0.12
GPT3_Navalny vs. GPT3_Non-state elites	0.24	0.49	0.13	0.45	0.64	0.32
GPT3_Navalny vs. Survey_State elites	0.16	0.34	0.04	0.20	0.26	0.36
GPT3_Navalny vs. Survey_Non-state elites	0.13	0.31	-0.03	0.17	0.27	0.24

Notes: ϕ_c – *Cramer's V*, a – accuracy, ρ – Pearson correlation.

In contrast to *Cramer's V* and accuracy metrics reliant on letter-level responses, the computation of Pearson correlations relies on token probabilities, thereby highlighting that these probabilities, on the whole, provide a relatively inadequate representation of survey percentages.

Unlike *Cramer's V* and accuracy measures built on letter responses, the Pearson cor-

relations are computed based on token probabilities, revealing that, on the whole, these probabilities offer a relatively inadequate representation of survey percentages. For example, in the year 2020, the correlation between GPT-3-generated responses attributed to "Putin" and survey non-state elites exhibits a weak negative relationship, aligning with our initial expectations. Contrary to our expectations, the correlation between responses attributed to "Putin" and the attitudes of state elites from the survey shows a weak negative correlation; similarly, the responses attributed to '"Navalny" exhibit effects that contradict our initial expectations.

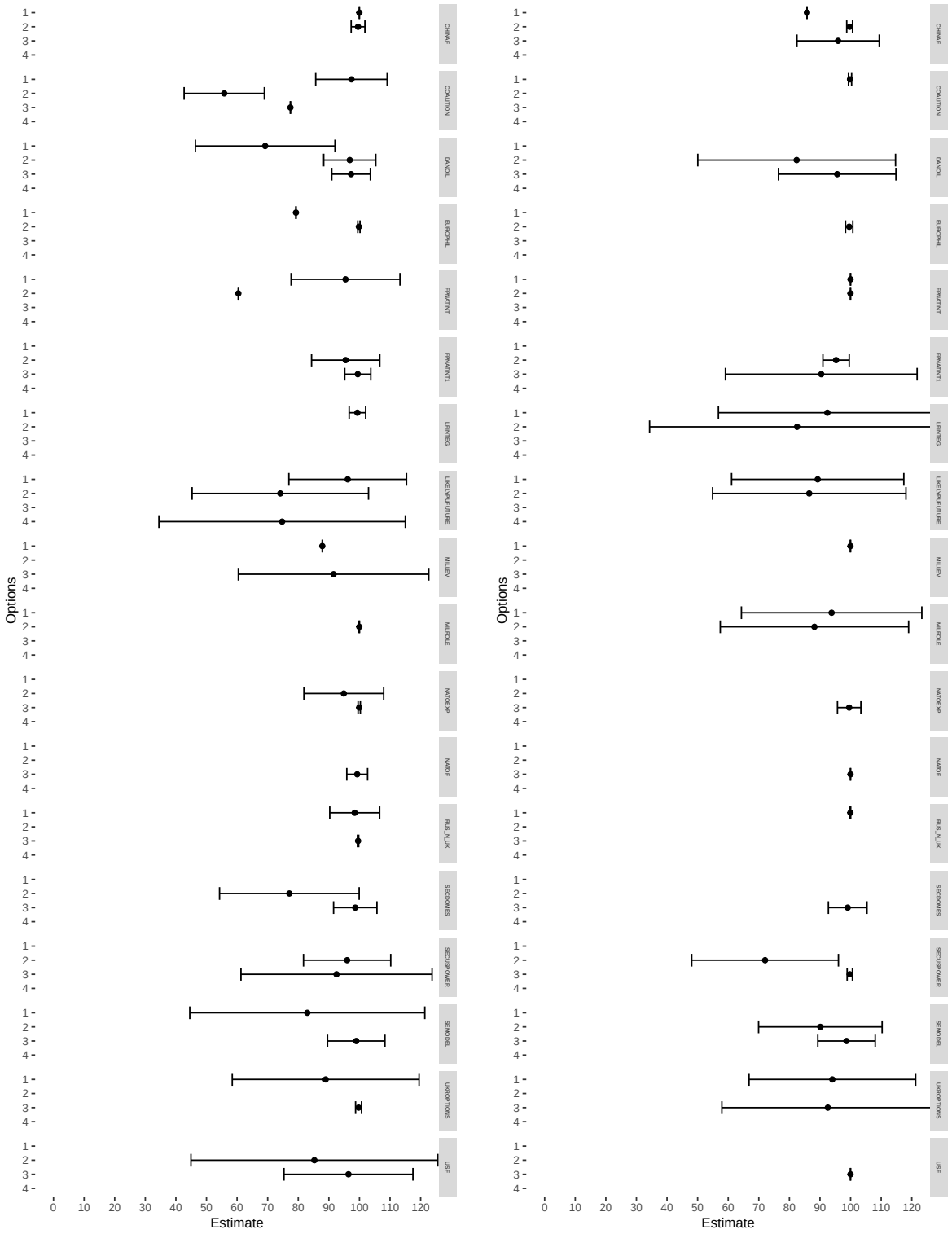
In summary, our validation analysis reveals that the GPT-3 model demonstrates complex associations between synthetic politicians and their associated groups. Both the *Cramer's* V and accuracy measures indicate better validation results in 2022 as compared to 2020. However, upon contrasting GPT-3 measures with survey proportions, we encounter rather mixed findings, indicating that synthetic data does not adequately capture the survey data. Overall, the **validation hypothesis** receives only modest empirical support, shedding light on the limitations of GPT-3-generated data concerning the Russian elites. Our **war effect hypothesis** has been partially confirmed: in comparison to 2020, in 2022, we indeed observe an increased alignment between "Putin" and state elites, and to a lesser extent between "Putin" and non-state elites; conversely, for "Navalny", the association between state and non-state elites has weakened.

Analysis of Context/Priming Effects

Context manipulation within GPT-3 helps to assess its effect on relevant responses attributed to synthetic politicians. This form of contextual priming holds the potential to yield valuable novel insights and facilitate the projection of attitudes and perceptions across diverse scenarios. Figures 2 and 3 illustrate the effects of different prompts on the responses associated with each synthetic politician.

According to Figure 2, there is a noticeable change compared to 2020, where an option

Figure 2: Comparing Context Effects for "Vladimir Putin'



2020 2022

suggesting that China is hostile toward Russia becomes plausible with a non-zero probability. However, it is important to consider that the wide confidence interval for token probabilities does not yield statistically significant results for the friendly and neutral choices (CHINAF). Remarkably, in the war context, the question regarding Russia forming a coalition results in the exclusion of other coalition options, leaving China as the only feasible choice (COALITION). Additionally, the question about the US being friendly toward Russia during the war period elicits a strictly negative response, indicating that "Vladimir Putin" perceives the US as hostile toward Russia, in contrast to his more neutral position in the prewar period (USF). Furthermore, both questions concerning NATO's friendliness toward Russia (NATOF) and the potential further expansion of NATO to countries in the Near Abroad (NATOEXP) logically exhibit the same level of negativity towards NATO in the context of war, or in some cases, an increase in negativity towards NATO.

Another question related to Russia's reliance on oil reveals that, in comparison to the pre-war period, during the war period "Putin's" response would be that a decrease in the price of oil in Russia represents a moderate and utmost danger (DANOIL). This observation logically suggests that Russia becomes more dependent on oil revenues during war times. As anticipated, the model generates an increase in military expenditures during the war, in contrast to the pre-war period (MILLEV). This shift in response aligns with the expectation that countries often allocate more resources to their defense during times of war. Interestingly, the model is indecisive about "Putin's" response in the pre-war period regarding whether Russia should follow the path of developed countries or adopt a unique Russian path. However, in the war setting, the model anticipates selecting only the latter as "Putin's" most feasible response (EUROPHIL). This suggests a strategic preference for a distinct Russian path following the upheaval of war.

The question about the future distribution of power by "Putin" when he leaves the presidency exhibits marked differences between the periods. In the war period, the model excludes the option of free elections and instead considers the options of transferring all

power to a trusted successor or associates as the most viable (LIKELYPUFUTURE). This observation aligns with my earlier findings from the *Predictioneer's Game* (Kalinin 2023*b*), further supporting the model's selection in the post-war context.

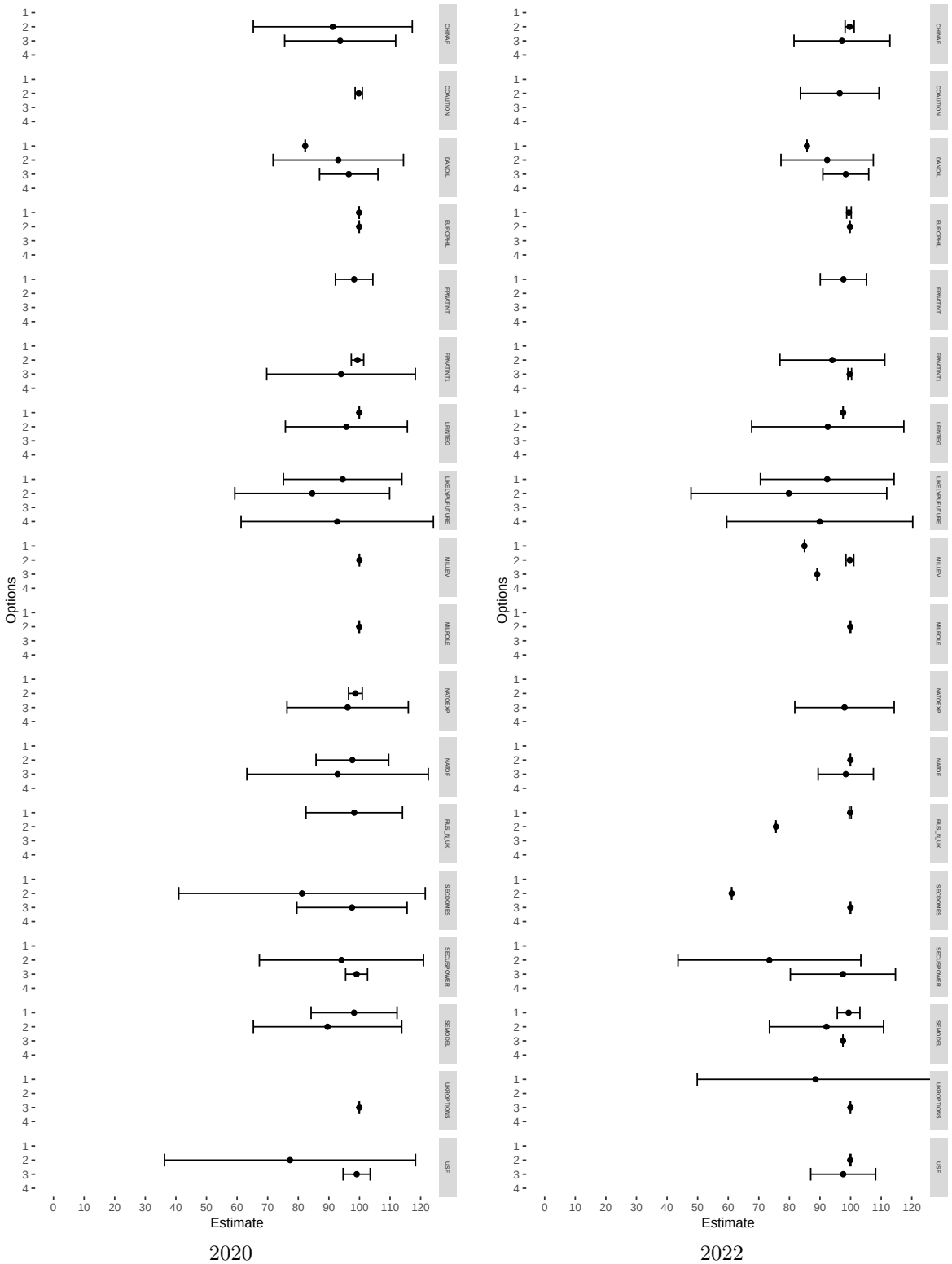
Other intriguing yet conceptually problematic findings suggest, for example, that the permissibility of employing the Russian military for the defense of territorial integrity exhibits greater uncertainty during the war compared to the pre-war period (LFINTEG). This observation raises valid concerns about the model's appropriateness in the war context. Furthermore, the most pivotal question concerning whether Russia and Ukraine should be completely independent countries indicates that the model supports the notion that "Vladimir Putin" would endorse both countries being entirely independent in both periods (RUS_N_UK). However, this generated result obviously contradicts current geopolitical developments.

In Figure 3, constructed for "Alexei Navalny", substantial differences between both studied periods are not observed. This suggests that pronounced contextual effects, as observed in the figure associated with "Vladimir Putin's" responses, are largely absent. In most instances, there is no significant cross-time variation, even though the confidence intervals and estimates of token probabilities exhibit some changes over time.

These findings suggest that the model's responses to prompts related to "Alexei Navalny" remain relatively stable across the periods. However, it is essential to carefully examine the questions and contextual factors to understand why certain variations are present while others are not. Further analysis may shed light on the reasons behind these differences and help elucidate the model's behavior in generating responses for different synthetic politicians.

For instance, during the pre-war period, the results generated by the model for "Navalny" indicate the importance of decreasing military expenditures (MILLEV). However, in the war period, the picture becomes much more uncertain, with all options having non-zero probabilities, signifying that the model cannot confidently extrapolate "Navalny's" view on this topic. Similarly, in the question concerning the future of eastern Ukraine, the model predicted in 2020 that "Navalny" would favor the idea of eastern Ukraine remaining part of

Figure 3: Comparing Context Effects for "Alexei Navalny"



2020 2022

Ukraine. However, in 2022, the model suggested a shift in "Navalny's" stance, indicating that eastern Ukraine should become part of the Russian Federation. These divergent responses highlight the complexity of the model's behavior when addressing certain issues and an obvious lack of training data for making appropriate inferences.

In addition to the cross-time differences, examining disparities in the responses of the two synthetic politicians sheds intriguing light on their their differing policy positions and viewpoints. The most significant contrast between "Navalny" and "Putin" becomes evident in their coalition partner preferences, where "Navalny" favors the European Union, while "Putin" leans towards China. Moreover, the model demonstrates less decisiveness about "Navalny's" attitude towards the US during the war period, with responses fluctuating between being "neutral toward Russia" and "hostile toward Russia" (USF). This nuanced stance stands in contrast to "Putin's" more definitive position. Furthermore, the model reveals an open indecisiveness about "Navalny's" response in the war period regarding whether Russia should follow the path of developed countries or adopt its unique Russian path, which diverges from "Putin's" pro-Russian path stance. Additionally, unlike "Putin", "Navalny" favors the option that suggests Russia's national interests should be limited to its current territory (FPNATINT). Interestingly, according to the model, in times of war, "Navalny" does not exclude the possibility of free and fair elections for power transition, while "Putin" does.

These distinctions in the model's responses for the two synthetic politicians offer valuable insights into their simulated policy preferences and perceptions under varying contexts. It is essential to interpret and contextualize these findings carefully to gain a deeper understanding of the model's behavior and its implications for different policy scenarios.

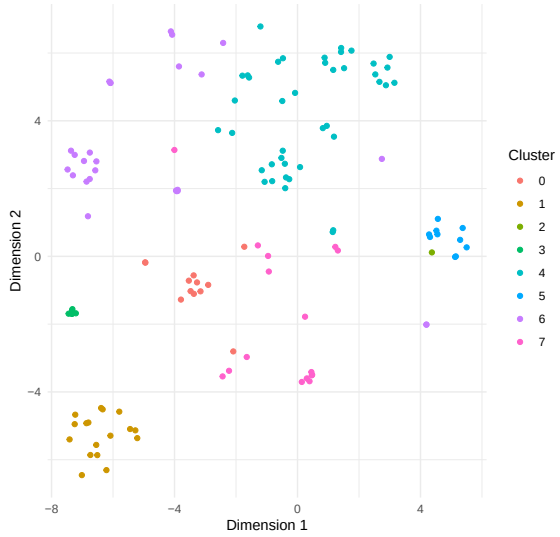
Heterogeneous Context Effects Across Question Clusters

The regression analysis enables us to investigate whether variations in context significantly influence changes in token probabilities for all questions, as observed through point and

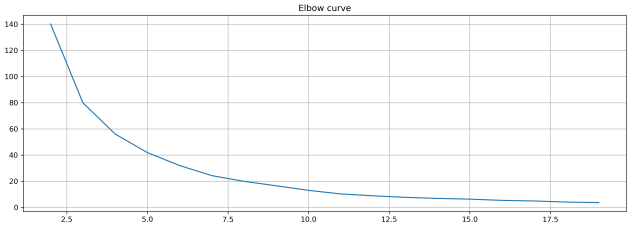
interval estimates.

To conduct the regression analysis, I followed a series of steps. Initially, I generated 2048-dimensional text embeddings for each multiple-choice question using the GPT's *Babbage* model (*text-similarity-babbage-001*). Subsequently, I created a two-dimensional space utilizing *t-SNE* (T-Distributed Stochastic Neighbouring Entities) for the purpose of visualization. Then, I employed the *k-means* clustering algorithm to categorize the *t-SNE* data points into predetermined clusters. This clustering approach helped to group semantically similar questions together, relying on their text embeddings.

Figure 4: K-means Semantic Analysis



(a)



(b)

Notes: (a) Cluster graph (b) Elbow plot.

In this study, the *k-means* clustering algorithm was applied to the entire dataset consisting of 129 questions from the *Survey*. The primary objective was to identify underlying

semantic patterns and associations within the subset of questions. Utilizing the elbow graph method, eight clusters were determined to be the most appropriate representation for the dataset. These clusters were subsequently labeled to signify their dominant thematic areas: *Cluster 0* related to the use of military in international politics, *Cluster 1* represented social values, *Cluster 2* focused on the integration of the post-Soviet space, *Cluster 3* dealt with political behavior, *Cluster 4* addressed sources of dangers, *Cluster 5* pertained to domestic issues, *Cluster 6* explored aspects of the political system, and *Cluster 7* examined individual characteristics. The resulting 2-dimensional graph visually displayed the distribution of elements across these clusters, providing a comprehensive overview of the survey's respondent classifications. This analysis offers valuable implications for understanding the perspectives of Russian elites on a wide range of topics and contributes to a deeper comprehension of their underlying beliefs and attitudes.

I utilized the linear mixed-effects model ($lmer$)[1] to conduct a nested regression analysis. In this model, multiple-choice options are nested within each variable, which accounts for the hierarchical structure of the data. The use of a nested *lmer* model is essential because it allows for the consideration of individual variability within each level of nesting, providing a more accurate representation of the data's inherent dependencies and improving the reliability of the regression estimates.

Table 2 presents results for a series of nested models investigating the effects of context changes on the mean and standard deviations of token probabilities associated with both synthetic politicians. The variable *Context* represents the impact of the war context on token probabilities. In Model M01, the coefficient of -5.59 for the *Context* variable indicates that, on average, the war context leads to a statistically significant decrease of approximately 5.59 units in "Vladimir Putin's" mean token probabilities. Similarly, in Model M05, the coefficient of 0.92 suggests a slight increase in "Alexei Navalny's" mean token probabilities in the war context; however, the p-value for this coefficient is not statistically significant.

[1] *formula*: Estimate $\sim Context + Cluster + (1|Variable/Options) for lmer() in \mathbf{R}$.

Table 2: Regression Analysis of Context Effects

	M01	M02	M03	M04	M05	M06	M07	M08
(Intercept)	68.42	74.74	6.07	6.88	90.78	96.29	7.68	2.91
	(2.24)	(8.21)	(0.44)	(1.5)	(0.7)	(2.8)	(0.48)	(1.77)
Context	-5.59	1.3	-0.35	3.15	0.92	-0.82	-1.78	-0.56
	(2.06)	(7.66)	(0.47)	(1.73)	(0.82)	(3.22)	(0.54)	(2.15)
Cluster 1		-5.61		1.81		-11.33		9.01
		(10.02)		(1.86)		(3.28)		(2.1)
Cluster 2		-12.58		-0.85		-7.84		7.75
		(25.07)		(4.76)		(6.89)		(4.62)
Cluster 3		3.61		-1.18		-4.47		-0.43
		(14.41)		(2.62)		(4.31)		(2.76)
Cluster 4		-7.5		-2.37		-1.78		4.06
		(9.01)		(1.66)		(3.06)		(1.95)
Cluster 5		-5.12		2.03		-5.78		7.14
		(11.11)		(2.07)		(3.48)		(2.25)
Cluster 6		-18.55		-1.5		-8.11		5.28
		(9.53)		(1.76)		(3.15)		(2.01)
Cluster 7		14.04		-0.82		-4.66		3.35
		(10.81)		(1.98)		(3.5)		(2.23)
Context × Cluster 1		-5.82		-4.13		5.07		-2.96
		(9.36)		(2.11)		(3.86)		(2.58)
Context × Cluster 2		15.97		-2.38		-11.27		1.57
		(23.41)		(5.29)		(7.81)		(5.24)
Context × Cluster 3		-3.69		-4.21		6.83		2.34
		(13.46)		(3.04)		(4.88)		(3.27)
Context × Cluster 4		-14.13		-4.04		-1.26		-1.49
		(8.42)		(1.9)		(3.53)		(2.36)
Context × Cluster 5		1.92		-0.94		-2.11		1.57
		(10.38)		(2.34)		(4.09)		(2.73)
Context × Cluster 6		-3.34		-3.89		3.45		-1.22
		(8.9)		(2.01)		(3.7)		(2.47)
Context × Cluster 7		-6.4		-4.49		6.02		-2.31
		(10.1)		(2.28)		(4.08)		(2.73)

Nested random effects model computed using *lmer() in* **R**. Models: M01, M02 – dependent variable "Vladimir Putin's" mean token probabilities; M03, M04 – dependent variable "Vladimir Putin's" standard deviations of token probabilities; M05, M06 -dependent variable "Alexei Navalny's" mean token probabilities; M07, M08 – dependent variable "Alexei Navalny's" standard deviations of token probabilities. Clusters: *Cluster 0*: the use of military in international politics, *Cluster 1*: social values, *Cluster 2*: integration of the post-Soviet space, *Cluster 3*: political behavior, *Cluster 4*: sources of dangers, *Cluster 5*: domestic issues, *Cluster 6*: political system, *Cluster 7*: individual characteristics.

For "Navalny", however, the negative effect of context on standard deviation is statistically significant, demonstrating an increase in the consistency of token probabilities.

The models also incorporate "cluster" variables, representing the main and interactive effects of different clusters of semantically similar questions on token probabilities. The inclusion of these "cluster" variables allows us to explore how the effects of the war context may vary across different clusters. For example, in Model M02 for "Vladimir Putin", the main effect of *Cluster 6* (political system) on the token probability is negative compared to *Cluster 0* (the use of military in international politics), with a coefficient of -18.55. The only statistically significant interactive effect is found in "Context × Cluster 4" (sources of dangers), with a coefficient of -14.13. This indicates that the war context leads to a decrease in the average token probabilities for point estimates related to this specific cluster. According to Model M04, the decrease in the average standard deviations due to the war context is associated with a range of clusters: *Cluster 1* (social values), *Cluster 4* (dangers), *Cluster 6* (political system), and *Cluster 7* (individual characteristics). This finding suggests that, for these types of questions, the generated token probabilities are characterized by less variability and greater consistency in the war context.

For "Alexei Navalny", in Model M06, all the main effects for *Cluster 1* (social values), *Cluster 5* (domestic issues), and *Cluster 6* (political system) are negative and statistically significant when compared to *Cluster 0*. This suggests that, in the war context, token probabilities associated with "Navalny's" responses related to social values, domestic issues, and the political system are consistently lower than those associated with other clusters. Regarding standard deviations in Model M08, all the main effects of clusters demonstrate positive values for several clusters (1, 2, 4, 5, 6), indicating that the model exhibits lower overall confidence in the generated data for these specific clusters. This suggests that the token probabilities for certain questions within these thematic clusters show greater variability, reflecting the model's uncertainty in providing consistent responses in the war context.

Overall, the findings from our analysis suggest that the war context exerts specific and significant effects on the token probabilities associated with the responses of synthetic politicians, with notable impacts on social values, domestic issues, and the political system, thus

partially confirming **survey questions hypothesis** – it appears that clusters related to war, values, and perceived threats are the ones most significantly affected by the war context.

Hence, this section has offered valuable insights into the GPT-generated responses for synthetic politicians within diverse contextual settings. The distinct effects observed across different clusters emphasize the model's sensitivity to thematic nuances and underscore the substantial role of context in shaping its generated responses.

Conclusion

This study aimed to address potential issues related to the design of prompts and option ordering effects by proposing a sampling-permutation algorithm. The main assumption of this algorithm was centered around the possibility of accurately recovering a true quantity through the utilization of prompt sampling and permutations for closed-ended survey questions. The objective was to achieve convergence towards the true value by combining these methods in the LLM setting. The proposed sampling-permutation algorithm for data generation entailed that LLM was capable of generating probable responses that were robust to semantically similar questions. By leveraging this approach, the resulting probabilities for letter choices enabled the calculation of measures of uncertainty and facilitated the assessment of variability in the generated responses for each specific question.

The paper demonstrated how the GPT-3 *Davinci* model could be used to generate responses from hard-to-reach members of the Russian elite in response to multiple-choice questions related to domestic and international politics. Within this study, three different code implementations were proposed: a) single-factor generation of data for closed-ended questions; b) multi-factor data generation for closed-ended questions utilizing the multifactor crosstabs; and, finally, c) data generation for open-ended questions. The code with examples is available on **GitHub**.

In this study, using the single-factor generation of data for closed-ended questions, re-

sponses were generated specifically for two synthetic politicians – President Vladimir Putin and opposition leader Alexei Navalny. The model not only proved useful in identifying the most likely synthetic responses and excluding the least likely ones but also enhanced the reliability of the generated output.

Our validation analysis demonstrated that GPT-generated data yields complex associations between politicians and their associated groups. Specifically, both the *Cramer's V* and accuracy measures indicate enhanced precision in GPT-3-generated responses in 2022 as compared to 2020. However, upon contrasting GPT-3 measures with survey proportions, we encounter rather mixed findings, indicating that synthetic data does not always adequately capture survey data – as a result, our **validation hypothesis** was only partially confirmed.

Furthermore, in comparison to 2020, in 2022, we indeed observe an alignment between Putin and state elites, and to a lesser extent between Putin and non-state elites; conversely, for Navalny, the association between state and non-state elites has weakened. Thus, our findings partially confirmed our **war effect hypothesis**.

The contextual effects analysis demonstrated the effective utility of GPT-3 in simulating various responses based on the given context. Through this analysis, significant differences between the two synthetic politicians, "Vladimir Putin" and "Alexei Navalny", emerged, providing valuable insights into their hypothetical political positions and preferences within distinct contextual scenarios. The study revealed that the war context exerted meaningful effects on token probabilities associated with the responses of both synthetic politicians, particularly with regards to social values, domestic issues, and the political system, thus partially confirming the **survey questions hypothesis**. All these findings shed light on how the model's generated responses can potentially vary across different thematic clusters under the influence of the war context.

Overall, the study expanded our understanding of how contextual variations affect LLM's responses for synthetic politicians and enriched the potential applications of LLMs like GPT-3 in political analysis and survey research. The ability to simulate responses based on context

opens new avenues for exploring policy positions and political stances across diverse scenarios, making such models valuable tools for gaining deeper insights into political dynamics and decision-making processes.

References

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* p. 115.

Bommarito, Michael James and Daniel Martin Katz. 2022. "GPT Takes the Bar Exam." *Date Written: December 29, 2022* .
URL: *https://ssrn.com/abstract=4314839*

Brecher, Michael and Jonathan Wilkenfeld. 1997. *A Study of Crisis*. The University of Michigan Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. 2020. "Language Models are Few-Shot Learners.".
URL: *https://arxiv.org/abs/2005.14165*

Buckley, Noah and Joshua A. Tucker. 2019. "Staring at the West through Kremlin-Tinted Glasses: Russian Mass and Elite Divergence in Attitudes toward the United States, European Union, and Ukraine before and after Crimea." *Post-Soviet Affairs* pp. 365–75.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*

Minneapolis, Minnesota: Association for Computational Linguistics pp. 4171–4186.
URL: *https://aclanthology.org/N19-1423*

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto and Pascale Fung. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Comput. Surv.* 55(12).
URL: *https://doi.org/10.1145/3571730*

Kalinin, Kirill. 2023*a*. "Generation of Synthetic Responses to Survey Questions Using GPT-3: A Case of Hard-to-Reach Members of Russian Elites (based on the Survey of Russian Elites)." *Annual Meeting of the Midwest Political Science Association* .

Kalinin, Kirill. 2023*b*. "Geopolitical Forecasting Analysis of the Russia-Ukraine War Using the Expert's Survey, Predictioneer's Game and GPT-3." *Annual Meeting of the Midwest Political Science Association* .

Liang, Percy, Tatsunori Hashimoto, Christopher Rishi Bommasani and Sang Michael Xie. 2022. "CS324 - Large Language Models." *Course* .
URL: *https://stanford-cs324.github.io/winter2022/lectures/modeling/*

Lin, Stephanie, Jacob Hilton and Owain Evans. 2022. "Teaching Models to Express Their Uncertainty in Words.".

Manakul, Potsawee, Adian Liusie and Mark J. F. Gales. 2023. "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models.".

Mueller, J.E. 1985. *War, Presidents, and Public Opinion*. UPA book University Press of America.
URL: *https://books.google.com/books?id=RENVPgAACAAJ*

OpenAI. 2022. "OpenAI Documentation on-line.".
URL: *https://beta.openai.com/docs/engines*

OpenAI. 2023. "GPT-4 Technical Report.".

URL: *https://cdn.openai.com/papers/gpt-4.pdf*

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Zimmerman, William, Sharon Werning Rivera and Kirill Kalinin. 2022. "Survey of Russian Elites 1993-2020, Moscow, Russia." *Inter-university Consortium for Political and Social Research [distributor]* .

A Appendix. Supplementary Tables and Figures

Table A1: List of Questions to Synthetic Politicians

Variable	Question	Options
CHINAF	[PLACEHOLDER] thinks that	• A. China is friendly toward Russia • B. China is neutral toward Russia • C. China is hostile toward Russia
COALITION	[PLACEHOLDER] would prefer Russia to form a coalition with	• A. China • B. European Union • C. US
DANOIL	[PLACEHOLDER] thinks that a decrease in the price of oil in Russia represents	• A. The absence of danger • B. Moderate danger • C. The utmost danger
EUROPHIL	Which of these statements is closer to [PLACEHOLDER]'s point of view?	• A. Russia should follow the path of developed countries and assimilate the experience and achievements of Western civilization. • B. Taking into account the history and geographic position of Russia at the crossroads of Europe and Asia, it should follow a unique Russian path.
FPNATINT	[PLACEHOLDER] thinks that	• A. The national interests of Russia for the most part should be limited to its current territory. • B. The national interests of Russia for the most part should extend beyond its current territory.

To be continued

Table A1: (*continued*)

Variable	Question	Options
FPNATINT1	[PLACEHOLDER] thinks that	• A. Russia has vital interests in the 'Near Abroad' but not beyond that. • B. Russia has vital interests in the 'Near Abroad' and Eastern Europe, but not beyond that. • C. Russia has vital interests in parts of the world not only in the 'Near Abroad' and Eastern Europe, but also in various parts of the world.
LFINTEG	In [PLACEHOLDER]'s opinion, defending the territorial integrity of the Russian Federation makes the use of the Russian military permissible?	• A. Yes • B. No
FUTURE	In [PLACEHOLDER]'s opinion, how will Vladimir Putin distribute power when he leaves the presidency?	• A. Transfer all power to a trusted successor or like-minded associates • B. Transfer power to at least one like-minded associate, but keep some power for himself well into the future • C. Keep all power in his own hands despite leaving the presidency • C. Let voters decide in fully free and fair elections, even if this allows a true opposition figure to win
MILLEV	[PLACEHOLDER] thinks that Russia should...	• A. Increase its military expenditures • B. Decrease its military expenditures • C. Keep its military expenditures at the same level
MILROLE	[PLACEHOLDER] thinks that	• A. Military force ultimately decides everything in international relations. • B. The economic, and not military, potential of a country determines the place and role of a country in the world today.

To be continued

Table A1: (*continued*)

Variable	Question	Options
NATOEXP	[PLACEHOLDER] thinks that further expansion of NATO to countries in the Near Abroad represents the greatest threat to the security of Russia	• A. The absence of danger • B. Moderate danger • C. The utmost danger
NATOF	[PLACEHOLDER] thinks that	• A. NATO is friendly toward Russia • B. NATO is neutral toward Russia • C. NATO is hostile toward Russia
RUSNUK	[PLACEHOLDER] thinks that	• A. Russia and Ukraine should be completely independent countries. • B. Russia and Ukraine should be partially independent countries. • C. Russia and Ukraine should be united into a single country.
SECDOMES	[PLACEHOLDER] thinks that the inability of Russia to resolve its internal problems represents the greatest threat to the security of Russia	• A. The absence of danger • B. Moderate danger • C. The utmost danger
SECUSPOWER	[PLACEHOLDER] thinks that the growth of US military power compared to that of Russia represents the greatest threat to the security of Russia	• A. The absence of danger • B. Moderate danger • C. The utmost danger
SEMODEL	Which country does [PLACEHOLDER] think can serve as a model of political and economic development for Russia?	• A. Scandinavia • B. Germany • C. China • D. US

To be continued

Table A1: (*continued*)

Variable	Question	Options
UKROPTIONS	[PLACEHOLDER] would prefer	• A. eastern Ukraine to become part of the Russian Federation • B. eastern Ukraine to become an independent government • C. eastern Ukraine to remain part of Ukraine
USF	[PLACEHOLDER] thinks that	• A. US is friendly toward Russia • B. US is neutral toward Russia • C. US is hostile toward Russia

B Appendix. Supplementary Analysis: Survey Data Generation

Guided by the specified criteria, this research presents three distinct code implementations: a) single-factor data generation for closed-ended questions, which entails the creation of a set of prompts and a full set of permuted multiple-choice responses for each question; b) multi-factor data generation for closed-ended questions utilizing multi-factor crosstabs to build prompts for subgroups based on multiple socio-demographic factors; and c) data generation for open-ended questions that resembles single-factor data generation but is adapted to the format of open-ended questions.

Approach I: Single-factor Generation of Data for Closed-Ended Questions

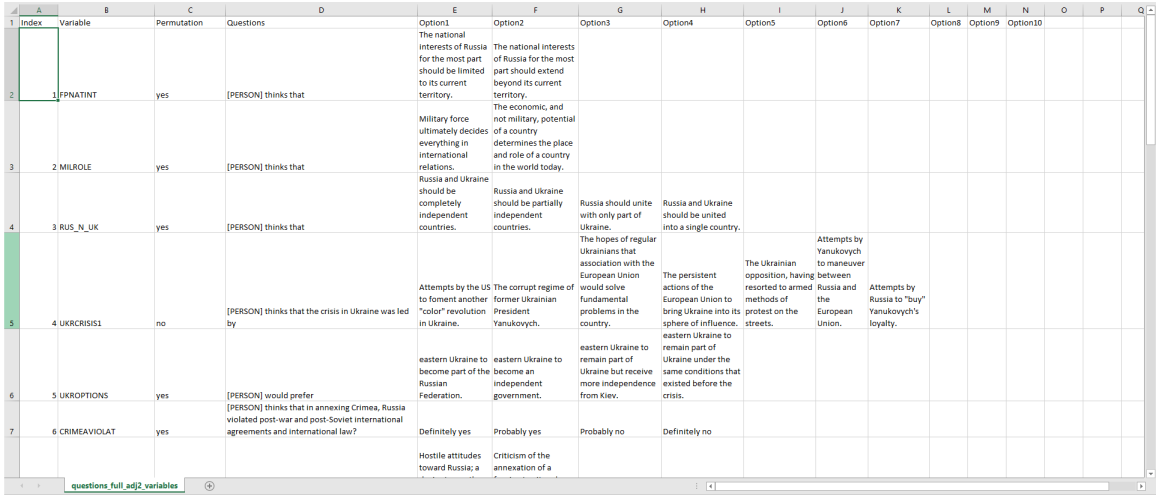
In this single-factor approach, the focus lies in generating responses from synthetic personalities without accounting for various socio-demographic characteristics. Therefore, in this research, the approach is employed for data generation concerning "Vladimir Putin" and "Alexei Navalny".

The construct prompt is as follows:

$$\underbrace{Placeholder}_{Person=\{VladimirPutin,AlexeiNavalny\}} \text{ thinks that } \underbrace{A.Option1; B.Option2; C.Option3.}_{Permutations:\{ABC\},\{BAC\},\{CAB\},\{ACB\},\{ACB\},\{BCA\},\{CBA\}} .$$

The Python script that implements automated generation of responses using the GPT-3 model is available on **GitHub**. For simplification purposes, one needs to fill out only the spreadsheet (see Figure B1) and run the Python code. The spreadsheet contains the following fields or variables: *Index* (question's index 1...N), *Variable name* (can be taken from external data set in case if there is a need to make comparisons between generated responses and survey outputs), *Permutation* ("Yes" if permuted options for specific question are allowed and "No" otherwise); *Questions* (a question can contain a placeholder "[Person]" or "[YEAR]" to be filled with specific name or concept); *Options1...10* (separate fields for

Figure B1: Preparation of Data for Response Generation Using GPT-3's *Davinci* Model



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Index	Variable	Permutation	Questions	Option1	Option2	Option3	Option4	Option5	Option6	Option7	Option8	Option9	Option10			
2	1	FPNATINT	yes	[PERSON] thinks that	The national interests of Russia for the most part should be limited to its current territory.	The national interests of Russia for the most part should extend beyond its current territory.											
3	2	MILROLE	yes	[PERSON] thinks that	Military force ultimately decides everything in international relations.	The economic, and not military, potential of a country determines the place and role of a country in the world today.											
4	3	RUS_N_UK	yes	[PERSON] thinks that	Russia and Ukraine should be completely independent countries.	Russia and Ukraine should be partially independent countries.	Russia should unite with only part of Ukraine.	Russia and Ukraine should be united into a single country.									
5	4	UKRCRISIS1	no	[PERSON] thinks that the crisis in Ukraine was led by	Attempts by the US to foment another "color" revolution in Ukraine.	The corrupt regime of former Ukrainian President Yanukovych.	The hopes of regular Ukrainians that association with the European Union would solve fundamental problems in the country.	The persistent actions of the European Union to bring Ukraine into its sphere of influence.	The Ukrainian opposition, having resorted to armed methods of protest on the streets.	Attempts by Yanukovych to maneuver between Russia and the European Union.	Attempts by Russia to "buy" Yanukovych's loyalty.						
6	5	UKROPTIONS	yes	[PERSON] would prefer	eastern Ukraine to become part of the Russian Federation.	eastern Ukraine to become an independent government.	eastern Ukraine to remain part of Ukraine but receive more independence from Kiev.	eastern Ukraine to remain part of Ukraine under the same conditions that existed before the crisis.									
7	6	CRIMEAVIOLAT	yes	[PERSON] thinks that in annexing Crimea, Russia violated post-war and post-Soviet international agreements and international law?	Definitely yes	Probably yes	Probably no	Definitely no									
					Hostile attitudes toward Russia; a	Criticism of the annexation of a											

questions_full_adj2_variables

Notes: Example of data preparation using an Excel spreadsheet to generate responses.

multiple choice options). The user does not need to number or alphabetize options because the code will automatically insert the appropriate letters during the processing stage.

For each permutation the algorithm extracts the option with the highest probability and calculates option-wise statistics, such as the mean and standard deviation. The rationale behind focusing solely on the most probable choices is to ensure that the resulting tokens are sensible, given that the options with lower probabilities could be nonsensical. Moreover, certain options may never be chosen by the model and thus are disregarded in the output.

A single API request or GPT-3 query provides normalized probabilities for generated responses for each permutation, which sum up to one. However, option-wise aggregate estimates for all permutations do not sum up to one, and thus normalization of the resulting quantities of interest is necessary to ensure consistency. Presently, the script does not implement such normalization.

Although computing permutations of questions can be computationally expensive, an increase in the number of permutations can increase confidence in the results and help assess the amount of relevant information in the large language model. Conversely, when the number of options and permutations is limited to two, we may have less confidence in the

generated results.

Approach II: Multi-Factor Data Generation for Closed-Ended Questions

This approach primarily focuses on data generation utilizing the multifactor crosstabs. The goal is to simulate survey data based on various socio-demographic factors. By constructing a crosstab for socio-demographic groups of interest and utilizing information on both the socio-demographic groups and their corresponding responses to a specific question, complete survey data can be generated using the GPT-3 (*text-davinci-003*) model. Proposed data generation approach enables the exploration of socio-demographic patterns and their relation to the generated responses, shedding light on the attitudes and opinions within different subgroups of the Russian elites. This study specifically aims to investigate whether socio-demographic subgroups divided by age, gender and elite status lean towards specific policy.

Considering the significant costs and time requirements associated with sampling the question component of the prompts, the decision was made to utilize the original wording of the prompts and solely rely on the permutation algorithm as a means of introducing variation. This approach was chosen to strike a balance between computational efficiency and maintaining a sufficient level of diversity in the generated responses.

For validation group-level section of analysis the following template has been used:

$\underbrace{Placeholder}_{Year=2020}$ $\underbrace{Placeholder}_{Age=\{young,old\}}$ member of Russian elite who belongs to $\underbrace{Placeholder}_{Elites=\{state,nonstate\}}$ elites and $\underbrace{Placeholder}_{Gender=\{male,female\}}$ thinks that $\underbrace{A.Option1; B.Option2; C.Option3.}_{Permutations:\{ABC\},\{BAC\},\{CAB\},\{ACB\},\{ACB\},\{BCA\},\{CBA\}}$.

The script is available on **GitHub**. The algorithm fills in the placeholders with the corresponding values retrieved from the crosstab object. Figure (a) demonstrates the output data for a single query and question permutation, presenting the generated data resulting from the application of the GPT-3 model. Figure (b) displays the cross-tabulation, which includes the means and standard deviations of token probabilities that have been generated by GPT-3.

This study aimed to demonstrate the application of the GPT-3 *Davinci* model in gener-

Figure B2: Data Generation: Study 2

(a)

(b)

(a) demonstrates the output data for a single query-permutation; (b) displays the cross-tabulation, which includes the means and standard deviations of token probabilities that have been generated by GPT-3. **Question:** "In {Year} {Age} member of Russian elite who belongs to {Elite Type} and {Gender} thinks that". Options: "Russia has vital interests not only in the 'Near Abroad' and Eastern Europe, but also in various parts of the world","Russia has vitally important interests in the 'Near Abroad,' but not beyond that.", "Russia has vitally important interests in the whole world."

ating multiple-choice responses from various subgroups of the Russian elites. In this study, the group-level data is generated based on socio-demographic factors such as age, gender, and elite type. This approach is adopted due to the high costs associated with generating data at the individual level. It is important to note that the generated group-level data typically exhibits low variation in token probabilities across different groups. This essentially suggests that the model lacks sufficient data to make distinctions between these groups.

To assess the accuracy of the generated group-level data, the token probabilities for the winning option are compared with the 2020 group-level percentages of the winning option. The findings, as depicted in Figure B1, reveal a wide variation in correlation coefficients and differences between the generated estimates and the baseline figures.[2] In terms of correlation analysis, the question on whether Russia should follow the path of developed countries (EUROPHIL), the question on the country-model of socio-economic development for Russia (SEMODEL), and whether the use of the Russian military permissible in defense of the territorial integrity (LFINTEG). Other questions can be found in Table A1 of the Appendix.

[2]The interpretation for the question labels can be found in Table A1 of the Appendix.

Table B1: Group-level Validation Study

	Option 1			Option 2		
Variables	ρ	δ	σ_δ	ρ	δ	σ_δ
FPNATINT	0.05	-0.68	0.19	0.75	-0.40	0.12
FPNATINT1	-0.20	-0.90	0.08			
MILROLE	0.04	-0.47	0.27			
LFINTEG	0.35	-0.89	0.10	0.38	0.11	0.15
SEMODEL	0.45	-0.58	0.36			
EUROPHIL	0.65	-0.21	0.15			
RUSNUK	-0.31	-0.04	0.06			
NATOCIS	-0.13	0.01	0.07	*nan*	-0.82	0.19
UKROPTIONS	0.30	0.12	0.14			
LIKELYPUFUTURE	0.09	-0.72	0.25	0.00	-0.71	0.25

Notes: ρ – Pearson correlation coefficient between the survey and GPT-3 data; δ – prediction error, $(\hat{p} - p)$; σ_δ – standard deviation of prediction error, δ.

The explanation for why these specific questions yield better validation results could be attributed to the fact that these questions were included in the training data.

Approach III: Data Generation for Open-Ended Questions

According to the third approach, the GPT-3 model is fine-tuned using publicly available data related to Vladimir Putin and Alexei Navalny. The data collection process took place in June 2022, sourcing information from two websites: www.kremlin.ru, which provided 8,603 transcripts spanning the period from 2000 to 2022, and http://www.navalny.ru, where 2,954 posts from the years 2009 to 2022 were collected.

Due to the substantial size of the dataset and the preference for a low-cost model, the decision was made to utilize the *Babbage* model, which contains 1.3 billion parameters and is 135 times smaller than the *Davinci* model. Despite its reduced capacity and limited functionalities, such as basic classification and semantic search, this model still produces intriguing results.

Due to the limitations of the *Babbage* model, which lacks the capability to provide letter choices for multiple-choice questions, a specific strategy had to be employed as an alterna-

tive approach. As per this strategy, the responses were generated as text fragments, each consisting of more than 100 tokens, approximately equivalent to 75 words. Additionally, to incorporate time sensitivity in the responses, each question referred to two distinct time periods: "before 2014", representing the period prior to the occupation of Crimea, and "after 2020", during which the political regime exhibited increased repression and hostility towards the opposition. By incorporating these temporal references, the study aimed to generate responses that are contextually relevant and reflective of the changing political landscape over time.

To associate the text fragment with a multiple-choice question, semantic similarity is computed between the generated text fragment and each multiple-choice option. This is accomplished by obtaining embeddings for each generated text and multiple-choice option using the fine-tuned *Babbage* model and then calculating the cosine similarity between them. Cosine similarity is a metric that measures the cosine of the angle between two n-dimensional embedding vectors projected into multidimensional space, given by $cos(\theta) = \frac{A \cdot B}{|\mathbf{A}||\mathbf{B}|}$. The multiple-choice option with the highest similarity to the generated text fragment is selected as the best matching option.

As described earlier, in order to associate a text fragment with a multiple-choice question, the process involves calculating the semantic similarity between the generated text fragment and each available multiple-choice option. This is achieved by obtaining embeddings for both the generated text and the multiple-choice options using the fine-tuned *Babbage* model. By determining the multiple-choice option with the highest similarity to the generated text fragment, the best matching option can be identified.

The quality of the generated results can be evaluated based on several criteria. First, the reasonableness of the generated results can be examined to determine whether they are meaningful and consistent with expectations. Second, a semantic consistency check can be conducted by comparing the outputs generated by the *Davinci* and *Babbage* models using similarity analysis.

Table B2: Comparison of Generated Responses to Closed-Ended and Open-Ended Questions

Question	Vladimir Putin				Alexei Navalny			
	S GE	*Davinci* (general)	*Babbage* (before)	*Babbage* (after)	**S NE**	*Davinci* (general)	*Babbage* (before)	*Babbage* (after)
FPNATINT	**B**	B	B	A	**B**	B	B	B
FPNATINT1	**C**	C	C	C	**C**	C	C	B
MILROLE	**A**	B	B	B	**B**	B	A	B
LFINTEG	**A**	A	A	A	**A**	A	A	B
SEMODEL	**D**	B	D	D	**D**	A	C	C
EUROPHIL	**B**	B	A	A	**B**	A	B	A
RUS_N_UK	**A**	D	D	D	**A**	A	A	A
NATOCIS	**D**	D	D	D	**D**	D	A	A
UKROPTIONS	**B**	A	C	C	**B**	A	D	D
LIKELYPUFUTURE	**E**	B	E	E	**C**	E	E	C
Accuracy		0.5	0.6	0.5		0.6	0.5	0.4
Cramer's V		0.68	0.73	0.76		0.75	0.69	0.47

Notes: See Table A1 in Appendix A for an interpretation of the variable names used. S – stands for the *Survey of Russian Elites*, GE – government elites, NE – nongovernment elites.

Table B2 presents the generated responses obtained from both models.

The main findings of this study are as follows.

First, the strength of association, as measured by *Cramer's V*, between the survey variables and the columns of the generated data indicates a strong relationship between our variables.

Second, the accuracy of predictions for "Vladimir Putin" is higher compared to "Alexei Navalny". Firstly, the dataset available for Alexei Navalny is smaller, with numerous questions from the *Survey of Russian Elites* being irrelevant to the content of his blog. Additionally, the data for Alexei Navalny has undergone automatic translation, and while *Google Translate* is a powerful tool, its output may not always be optimal. Moreover, certain responses from "Alexei Navalny" after the year 2020 bear resemblance to those of Vladimir Putin. This similarity might arise due to the *Babbage* model's attempt to fill gaps in "Alexei Navalny's" data with responses from "Vladimir Putin". This assumption finds support in the correlation analysis of t-SNE scores, which indicates a shift in the correlation between the

responses of the two politicians. The correlation changes from negative before 2012 (-0.32) to positive after 2020 (0.27). Considering the political developments in Russia since 2020, including heightened repression and brutality against political opponents by the regime, the observed positive correlation in responses between the two politicians is rather surprising. The second finding pertains to the strength of association measured by *Cramer's V* between the survey variables and the columns of generated data, indicating a strong association between them.

Third, when it comes to the quality of the generated responses, Table B2 indicates that, despite being fine-tuned, the *Babbage* model does not significantly improve the results. Since *Davinci* model does not require additional fine-tuning it seems to outperform better the *Babbage* model.