

Self-Reinforcing Effects on Trustworthiness of Intelligent Systems in Small Data Problem-Solving

Vladyslav Kuznetsov^{1,2,*}, Alexander Samoylov^{2,†}, Tetiana Karnaukh^{2,†}, Rostyslav Trokhymchuk^{2,†} and Anatoliy Kuliash^{1,†}

¹ Glushkov Institute of Cybernetics, 40, Glushkov ave., Kyiv, 03187, Ukraine

² Taras Shevchenko National University of Kyiv, 63/13, Volodymyrska str., Kyiv, 01601, Ukraine

Abstract

In this work, we propose to investigate such a topic as trustworthiness in large language models in particular, in tasks of knowledge mining. As a part of experimental research, we conducted anonymous study of big data models, whereas analyzed how the size of the context, model memory and number of interactions affect trust estimated value in some quality assessment indicator value. The trust estimate was formed based on the assessment of the quality of the answers given by the models, related to the questions given, using an estimate value, on a scale from 1 to 5, which shows complacency and conciseness of the model answers. As part of the experiments, 11 large language models, with the number of parameters ranging from 1.5 to 13 billion, were studied. For the quality assessment on the task of knowledge mining, the questions were formed on such area, as trustworthiness, using standardized definitions of trust categories from the ISO/IEC TR 24028 standard. In this task, during the experiments, we noticed that each interaction is crucial for trust assessments, because the estimated value of trust either remained the same, increased or decreased. As a result, this showed the complex nature of interactions and a wider range of values of trust in artificial intelligence (AI). We inferred from the experiments that the value of trust is very likely dependent on the previous context and memory of interactions. Thus, the effects of trust reinforcement had shown themselves on nearly all large language models tested in the experiments, whereas the best results were obtained by language models with exact 4 and 8 billion parameters. The study was also covering aspects of model performance and its efficiency based on language encoding. As a result of the experiments, we suggested a number of requirements for personal, as well as public AI systems given the case study example.

Keywords

Trustworthiness, Intelligent systems, Large language models, Self-reinforcing effects, Small data

1. Introduction

Recently, with the rapid development of machine intelligence algorithms such as Large Language Models (LLMs) [1], as well as greater market availability, in particular in area of parallel computing tools, it has been possible to deploy large machine intelligence models on local machines as personal artificial intelligence elements. Opposed to commonly known commercial platforms such as Google Cloud, Amazon AWS, Microsoft Azure or, conditionally free like Hugging Face, one can run a personal AI. This has raised an interest for the developers to make use of AI models that can be run on personal computers locally as personal artificial intelligence for natural language processing, recognition and synthesis [2,3]. The presence of full control over the process of the

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ kuznetsow.wlad@gmail.com (V. Kuznetsov); SamoylovSasha@gmail.com (A.Samoylov); tkarnaukh@unicyb.kiev.ua (T. Karnaukh); trost@knu.ua (R.Trokhymchuk); anatoly016@gmail.com (A. Kuliash)

ORCID 0000-0002-1068-769X (V. Kuznetsov); 0000-0002-7423-5596 (A.Samoylov); 0000-0001-6556-1288 (T. Karnaukh); 0000-0003-3516-9474 (R.Trokhymchuk); 0000-0002-1322-4551 (H. Kudin); 0000-0003-3715-1454 (A. Kuliash)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

language model allows, firstly, increasing the transparency of the process both from the point of view of reliability (algorithmic) and from the point of view of data security, without the risks of transferring sensitive data to third parties or companies.

Currently, there are numerous tools for both language models training, post-training and model deployment, which can range from libraries, command line utilities up to more complex applications such as GPT4All or Nvidia ChatRTX [4]. However, further deployment of the models was limited due to limits of consumer grade hardware [5], and inability to use of high machine precision (like FP16 or FP32), so the developers experimented with memory optimizations - with less precision, trimming weights, creating reduced versions, for instance fp8 GGUF, Q_4, bf16 nf4 (mixed precision FP16) [6,7]. Such optimizations of the models, in fact, led to the appearance of a separate class of models - **small language models**, which show the greatest interest for this study [8,9]. It is worth noting that reducing the size of the language model has its drawbacks, since the number of parameters of the language model as well as its numerical precision directly affects the accuracy of the model and its ability to predict flow of the dialog, the amount of short-term memory, attention, and the ability to retain context.

These limitations pointed to one critical aspect in the testing and evaluation of artificial intelligence models [10]. While, with increase in the number of parameters in classical methods, it is possible to assess how this change will affect the accuracy of the answer, contrary to this, in the evaluation of the performance of language models, a large role is played by the subjective factor. Hence, when evaluating the model in terms of answers, the main role is played by the competence of the user and his or her ability to ask clearly formulated questions in order to receive the same clear and specific answers [11,12]. The accuracy value itself is a very subjective assessment which is difficult to formalize. In order to evaluate the performance of large language models, it is necessary to involve experts who evaluate model responses based on pairwise comparisons of one language model with others, or utilize user surveys that show evaluation of model performance based on satisfaction with answers [13].

One should note, that such approaches have their advantages, however the presence of a person in the evaluation process introduces its own subjective factor (human in the loop), which makes one to ask about the feasibility of such evaluations, as well as their trustworthiness based on selection criterion [14]. In our opinion, a proven way is to assess the trustworthiness of human-machine interaction (human-AI in this case), using a widely accepted industry standard, ISO/IEC TR 24028 [15].

This standard is aimed to formalize the concept of trust in AI systems as well as the ability of AI to provide clear and understandable answers that match the user's expectations, understanding and priorities. Within the ISO/IEC TR 24028 standard, the concept of trustworthiness in AI systems is formalized in the form of a mathematical apparatus based on set theory [16], an algebra of concepts, that use ontologies to clearly define trustworthiness categories. This made possible to obtain a clear and understandable set of tools and methods for assessment of trustworthiness in artificial intelligence.

For a clear understanding of the requirements for trust in AI systems [17], the standard describes a number of requirements that can also be applied to large language models (LLM). However, in the standard, as a drawback, not enough attention paid to ability of LLMs to behave independently, with certain effects of reinforcing or amplification of the interaction over time [18], in contrast to systems that have time-fixed trustworthiness. It is of a great importance to take into account this indicator, as well as other indicators, such as accuracy, stability, reliability (algorithmic) [19]. This creates a number of questions when evaluating LLMs, we decided to ask prior to formulating tasks of the study:

- Can AI be trusted if it gives correct but not complete answers?
- Does training sample size, number of network parameters, or other constraints (e.g., speed, memory, short-term memory size, context size) affect confidence in the model?

- Is trust in AI in the case of LLM a static value or is it variable over time?

It is clear that the difference between systems that do not have short-term memory and attention mechanism, and ones that have it, is quite noticeable when one does practical assessment of their performance [20]. Though, questions may arise in comparing models with different degrees of attention. For example, in AI systems that have no attention mechanism and short-term memory as LLMs have, the errors and their affect on accuracy and trustworthiness can be expressed numerically within given dataset and being valid for this set on small and large data after numerous experiments [21]. In the case of LLMs, there is an effect of uncertainty and ambiguity. This happens because in LLMs each subsequent interaction is attended to a number of previous interactions, which is limited by the ability to retain context and the amount of connectivity of weights (for example, using the mechanism of self-attention) [22]. The presence of spatio-temporal dynamics of trustworthiness implemented as dependency on both the number of interactions and their type, expressed by a certain textual or numerical vector or matrix, creates potential trust-reinforcing effects, when with an increase in the number of interactions, the amount of trust increases or decreases over time, impacting the overall confidence and trustworthiness of the studied model. Because of that, one may require a set of mechanisms to evaluating the self-reinforcing effects of trust over time [23].

In this work, we propose to study the aspects of trustworthiness in AI in human-machine interaction using an example of a sequential dialogue of a LLM and a user. We propose to assess the problem of dynamic trustworthiness of LLMs by fulfilling the following tasks:

- to propose an estimate the self-reinforcing effects in trust in AI systems,
- to conduct practical study the qualitative and quantitative interactions which may or may have influenced on the amount of trust over time,
- to define influence of model limitations on its theoretical accuracy and trust, the scope of interaction and its impact on trust as a result.

This, in turn will allow one to determine trust in AI and allows it to be analyzed and researched. Using case study on small data sets, as well as personal interaction of users with AI, will allow one to determine how the volume of training data and the number of parameters affect the behavior of small and medium-sized LLMs. For this purpose, in the following sections of this work, we proposed to solve the proposed problems by conducting experimental research of AI systems on the example of the problem of knowledge mining and to assess estimated values of confidence in AI during these experiments. The practical recommendations for trust evaluation based on user satisfaction for the practical evaluation of the model would be developed, as well as the proposals from the ISO/IEC TR 24028 standard would be utilized.

2. Experimental Study of Large Language Models

In experimental studies, we proposed to investigate the accuracy and quality of the response of various small LLMs by asking clearly formulated questions related to the topic of trustworthiness. This will make it possible, firstly, to fundamentally assess the abilities of AI to understand trustworthiness, and secondly, to use the ISO / IEC standard TR 24028 not only for the development of the assessment methodology, but also for determining the completeness of the answers to the compliance of the standard and its definitions. Also, as part of the experiments, we proposed to investigate how trust changes in LLM with the number of interactions, to build graphs of the dependence of trust on the number of interactions, with special attention to dependencies and trends, disturbances and fluctuations of trust and the general resistance of LLM to fluctuations in input data (for example, the length of the text, its content, expected user response, etc.). By

analyzing these factors, we will be able to better understand how trust is reinforced during long-term interaction.

As part of the experiments, each of the tested models was asked a number of clear and understandable questions on the subject of trustworthiness in AI. Answers were rated on a scale from 1 to 5, where 1 was incomplete and unclear (irrelevant) answers, and 5 was well-structured answers that contained a combination of ideas, original definitions with interpretations of generally accepted definitions, and retained memory (context) of the conversation to predict questions and type of answers that the user expects to receive. To reinforce this, feedback was introduced into each question, which contained the user's assessment and satisfaction with the answer, in order to reinforce useful actions of the LLM.

Nvidia-based LLM deployment environment Chat RTX and LLM checkpoints downloaded from the official repositories of LLM developers. The Nvidia video accelerator was used to defragment LLM RTX 4060 Ti on a Windows 11 computer with a total memory of 32 GB (system memory and video accelerator memory was used for deployment).

In order to reduce the influence of bias towards LLM, an anonymous study of models was conducted, which included uploading to the Chat RTX system model files (checkpoints) that contained only the number of parameters and the serial number of the model, without specifying the architecture and developer of the model in the name.

The experiment will be conducted in 2 stages:

- At the 1st stage, the effects of trust reinforcement in models with 1.5-4 billion parameters (small LLMs) would be investigated.
- Similarly, at the 2nd stage, these effects would be studied in more detail in models with 7-13 billion parameters (medium-sized LLMs).
- As an additional stage, an experimental study would be conducted to test language encoding density in 7-9 billion parameter monolingual models.

This approach would made possible to focus on the effects that occurred at the micro level, to determine their types and features, while at the 2nd stage, effects at the macro level were studied, when not only quantitative changes occurred (improvement of trust from the number of interactions), but also qualitative ones (the system could anticipate the user's questions and adapt to his requirements and the format of the answers). This made possible to determine the effect of model size on response quality, context retention, and confidence build up over time.

2.1. Stage 1 of the study: testing small LLMs

To conduct the research, an assessment was made about increase or decline of trust over time. For instance, each answer, regardless of its quality, was given 5 percentage points (absolute) relative to the previous answer, so it increased by adding 5 points when trust increased, and subtracted if trust decreased accordingly. All models were given a confidence value of 70 percentage points before asking the first question, as a conservative estimate, regardless of their performance to reduce the influence of model bias or human factor. Thus, if the model gave an answer better than expected (based on a conservative score of 70 points) on 1 step or question, it was given a score of 75 points right away and similarly on subsequent steps (questions) if the confidence changed in one or the other side (positive or negative).

In the first stage, the smallest model with 1.5 billion parameters was initially tested. Overall, this model showed the ability to increase trust over time in the short term, with a sharp decline in trust after the 2nd question. In general, the model showed the quality of answers at the level of 1–2 points, with a peak of 2 points after the first two questions. However, after the 2nd question, the quality began to deteriorate, with a steady decline from the initial 70 points to 60.

It was followed by a model with 3 billion parameters, which showed much better response quality and relatively high stability (with minor fluctuations in response quality and confidence). Overall, the model received approximately 4 points for the best answer, with an average rating of 2.5 points. The quality of the answers began to decline after question 11, reaching a minimum after question 15, which received 2 points per answer. In general, the value of trust reached 80% for a short period, with a decrease to 70% (initial value of trust).

In turn, the 4 billion parameter model consistently scored 4 (lowest 3, high 5) in response quality average, with confidence increasing from an initial 70% to 80%, with subsequent swings around the 80% average with scores of confidences that did not exceed 85%. In general, this model showed a much better understanding of the concept of trustworthiness in AI and was able to generate relatively original answers related to the topic of trustworthiness in AI. Compared to the previous 2 models, there was a clear trend with increasing confidence (trust) over time, however, the presence of trust fluctuations, as well as the inherent uncertainty of the maximum value, allowed us to give conservative estimates based on a short set of questions.

2.2. Stage 2 of the study: testing of medium sized LLMs

Since the general trends and effects of increasing the trust were clear, in this experiment we proposed to focus more on the changes in the number of percentage points of trust with the number of interactions (using the example of 22 questions on the topic of trust in AI). If the model reached a trust plateau from either the upper or lower bound, no further questions were asked. Experiments for the entire question set were conducted separately for models that provided meaningful responses throughout the study to all questions. 8 models with the number of parameters from 7 to 13 billion were analyzed (including 6 models with 7 billion parameters, 1 of 8 and 1 of 13, respectively). In order to compare the models, graphs were constructed that contained the question number horizontally and the trust value (in percentage points) vertically, respectively. The graph (fig. 1) uses the MxPy notation, where x denoted the serial number of the model and y denoted the number of parameters. For example, M1P7 is model number 1 from the subspecies of models with 7 billion parameters.

Let's dwell in more detail on the graph (fig. 1) for models with 7-13 billion parameters, and models that showed an increase in trust indicators over time:

- M1P7 model (1st model out of 7 billion parameters) showed reliable indicators of the quality of the answer at the level of 4.5 points on average. The quality of the answers was kept at this level until at least 15 questions. Trust fluctuations relative to the trend were about 2.5% (absolute). The trend was horizontal, with average trust levels at 78%.
- Model M1P8 (1st model of 8 billion parameters) showed the quality of answers at 5 points on average. At the same time, the relationship of trust on the number of answers had a clear upward trend, which began at the level of 77.5% and ended at the level of 92.5%. The magnitudes of trust swings were minimal at the level of 0.75% (absolute), with a potential asymptote at the level of 95%.
- Model M1P13 (1st model out of 13 billion parameters) showed the results, at the level, approximately in the middle, between the model M1P7 and M1P8. The quality of the answers ranged between 4.5 and 4.75 points. The general trend was upward (not flat as in M1P7, but not as steep in slope as in the M1P8 model), with an asymptote at the level of 82-84% and trust swings at the level of 1.2% (absolute), which showed an ability to improve results over time.

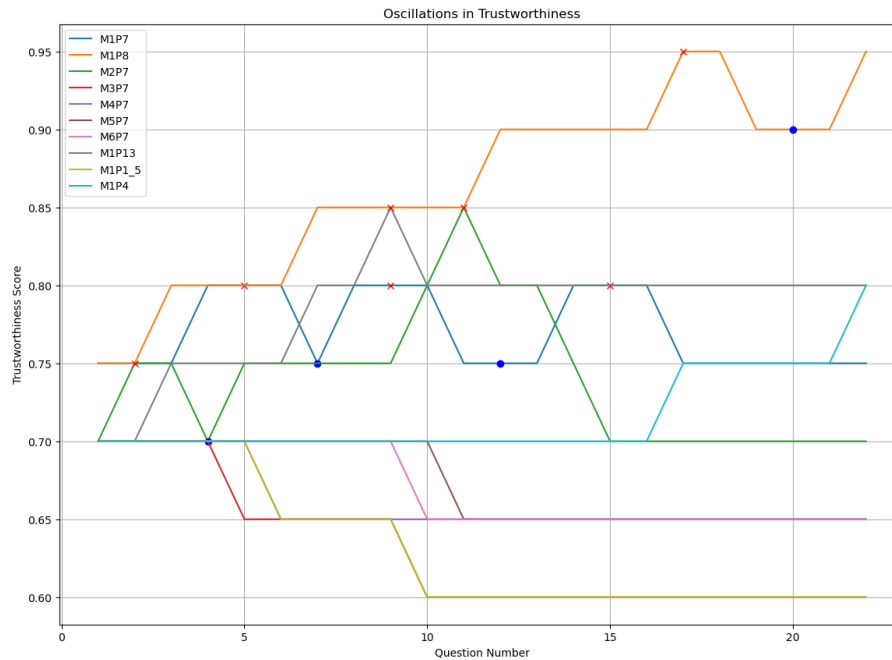


Figure 1: Graph of dependence of trust in large language models on the number of interactions.

The most interesting result was that in all models with each interaction, the effect of interaction improvement decreased with each step and the magnitude of this improvement became non-significant after 21 questions in almost all models, indicating that the effect of interaction improvement has certain limitations and cannot happen indefinitely. This indicates that while model size is important, it is more critical to obtain trustworthy results in the first steps, which anchors the user's trust in the first steps and the trust grows steadily in subsequent steps, allowing for predictable results over time.

2.3. Stage 4 of the study: impact of language encoding effectiveness on the model performance in monolingual context

The third part of the experiment was devoted to the influence of the effectiveness of text encoding in large language models on the completeness of their answers. This experiment was based on the hypothesis of a potentially more efficient encoding of text [24] in languages with hieroglyphic writing, such as Chinese and Japanese, and languages that do not contain vowel sounds in writing (for example, Arabic). The essence of the experiment consisted in a detailed analysis of the model's answers to questions posed with the help of a machine translator, for which the answers were compared with a model that had similar properties, but was English-speaking. These models had the following parameters: the Japanese model - 8 billion parameters, the Chinese model - 9 billion parameters, the Arabic model - 7 billion parameters, the English model - 8 billion parameters (model M1P8 from the previous experiment). The image below (fig. 2) demonstrates part of a dialog done in English, having a question and answer, dedicated specifically to ISO/IEC TR 24028 standard.

In general, the efficiency of coding showed itself differently both in terms of text density and the quality of answers compared to English. To determine the density, the entire dialogue with the language model was translated into English, and the coding accuracy was determined as the ratio of the number of symbols in the text, written in the given non-English language (without space characters) to the number of symbols of this fragment translated into English. The Japanese model had a coding density of 54% (about 2 English characters for each Japanese character or symbol), the Chinese model had a coding density of about 32% (about 3 English characters for each Chinese

character or symbol), the Arabic model had a coding density of about 75% (about 4 English characters per each 3 Arabic characters or symbols).

Prompt: There are many technical standards on the trustworthiness of AI systems, in particular ISO/IEC TR 24028. What do you know about this technical standard and how it defines trustworthiness? What do you know about other standards on the trustworthiness of AI systems in general?

Response: A great question that delves into the world of technical standards!

****ISO/IEC TR 24028: Trustworthiness of Artificial Intelligence****

The ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission) collaborated to develop Technical Report 24028, "Trustworthiness of Artificial Intelligence." This report provides guidelines on how to assess the trustworthiness of AI systems. The standard defines trustworthiness as:

"The degree to which an artificial intelligence system is reliable, consistent, and free from intentional or unintentional biases that could lead to unfair outcomes."

The ISO/IEC TR 24028 framework consists of three main components:

1. ****Trustworthiness criteria****: This includes factors such as transparency, explainability, accountability, fairness, security, robustness, and reliability.

Figure 2: Example of the conversation in English with a 8 billion parameter LLM.

Let's dwell in more detail on the answers of different models based on the assessment of the quality and completeness of the answers translated into English, while the assessment was made on the basis that the quality of the answer is at least 4 points on a 5-point scale in order to exclude the influence of the translator on the quality of the text.

Compared to the English model, the Japanese model demonstrated greater quality and completeness of responses, which was manifested in more expressive, structured responses, which was correspondingly reflected in the initial values of both confidence in the model and its ability to preserve the context of the conversation. The quality of the answers reached 5 points on a 5-point scale, which is the highest indicator among the models that were tested. The model showed an unambiguously good understanding of the concepts of trust in artificial intelligence, for example, a more detailed description and breakdown of definitions into separate categories, lists, highlighting the main structural categories and concepts of trust, their components and connections between these concepts. The answers of this model took into account the wishes of the user (feedback) and contained potential advice on potential questions on the subject of trust in artificial intelligence. Also, a significant feature of the model was the coding of complex concepts with a smaller number of characters. Overall, confidence in the model was on average higher than in the English-language M1P8 model, but reached a plateau after which the differences between English- and Japanese-language responses became smaller. The figure 3 demonstrates a fragment of a conversation with Japanese based LLM.

The result of the text of the Chinese-language model was interesting. Although it had three times higher coding efficiency and a larger model size (9 billion parameters compared to 8 in English), the quality of these responses were lower than in English. This was telling about, firstly, less engagement in the conversation and less initiative in terms of providing leading questions, including prompts for further discussion. Also, the completeness of the answers was clearly lower and was at the level of 3.5 points on average (compared to at least 4 points in the English language model). However, this has not resulted in bad understanding the concepts of trustworthiness - the model understood what trust is and could give a quote from the ISO/IEC TR 24028 standard - but in their formulation, which was more reference than an original, own. This can be caused both by restrictions on the length of the answer, and by the style of the model itself, which aims to give

shorter answers, as opposed to long ones, as, for example, in the Japanese model (fig. 4). Also, another reason for the low response quality score was that the model repeated old definitions and used less text from previous responses to improve results.

Prompt: いいですね。技術的な制限により、翻訳者のおかげで、AIの信頼性の定義を信頼性/信頼性と併用できることに注意してください。AIシステムの信頼性に関する多くの技術標準、特にISO/IEC TR 24028があります。この技術標準について、また信頼性をどのように定義するかについて何を知っていますか?一般的なAIシステムの信頼性に関する他の基準について何を知っていますか?

Response: ISO/IEC TR 24028 (AI: Artificial Intelligence – Guidelines for Trustworthy AI) は、AIシステムの信頼性に関する国際標準であり、特に「信頼できるAI」(Trustworthy AI)の概念を定義しています。この標準は、AIシステムが信頼性、透明性、公平性、説明責任、安全性などの要素を満たすことを目指しています。以下に、ISO/IEC TR 24028に基づく信頼性の定義と、一般的なAIシステムの信頼性に関する他の基準について説明します。

ISO/IEC TR 24028における信頼性の定義

ISO/IEC TR 24028では、信頼性は以下のように定義されています(一部抜粋)：

「信頼性は、AIシステムが特定の期待された性能レベルを維持し、予測可能な方法で動作する能力を指します。これには、システムが一貫して正確な結果を提供し、エラーやバグが最小限に抑えられることが含まれます。また、システムが外部要因や変動に対して耐性を持ち、予期しない状況でも適切に対応できることも信頼性の一部です。」一般的なAIシステムの信頼性に関する基準
AIシステムの信頼性を評価するためには、以下のような基準や指標が広く使用されています(一部抜粋)：1.精度と再現性・精度(Accuracy)：モデルが正しく分類、予測、または予測を行う割合。

Figure 3: Example of the conversation in Japanese with a 8 billion parameter LLM.

Prompt: 关于人工智能系统可信度的技术标准有很多，特别是ISO/IEC TR 24028。您对这项技术标准以及它如何定义可信度了解多少？您对其他关于人工智能系统可信度的标准了解多少？

Response: ISO/IEC TR 24028是一项国际标准，旨在为人工智能系统的可信度提供一个定义框架。标准旨在识别人工智能系统中的信任度概念，并提供一个框架，用于设计、实施和维护这些系统。标准包括了信任度的七个核心特性，即透明度、解释性、安全性、隐私、准确性、可靠性和效率。
关于其他标准，NIST(美国国家标准技术研究院)发布了人工智能系统的可信度框架，该框架提供了设计、开发和部署AI系统的一套建议原则。此外，欧盟机构也在开发一系列标准和指南，以确保开发和部署AI系统时遵守伦理准则和保护个人数据的原则。
这些标准和指南对于企业和组织来说非常重要，因为它们有助于确保AI系统在设计 and 部署时考虑到信任度概念，从而为最终用户提供可靠、安全和负责任的体验。

Figure 4: Example of the conversation in Chinese with a 9 billion parameter LLM.

In the Arabic language model, these differences were less significant (potentially because of high density of encoding characters in written text). A feature of the answers is their shorter length (less than in English), but the quality was not as high as, for example, in Japanese due to the impossibility of structuring concepts into more deep categories (fig. 5). Therefore, on the one hand, the model showed a good understanding of the topic of trust, and showed the ability to predict the course of the conversation (albeit at a level somewhat lower than the Japanese one). Thus, the effectiveness of this model, as well as the conciseness of its answers, can be estimated at a level slightly higher than in English, but lower than in Japanese. In general, some answers were at the level of 4.5 points, which is in itself a good result for a monolingual model.

Prompt: 关于人工智能系统可信度的技术标准有很多，特别是 ISO/IECTR 24028。您对这项技术标准以及它如何定义可信度了解多少？您对其他关于人工智能系统可信度的标准了解多少？

Response: ISO/IECTR

24028是一项国际标准，旨在为人工智能系统的可信度提供一个定义框架。标准旨在识别人工智能系统中的信任度概念，并提供一个框架，用于设计、实施和维护这些系统。标准包括了信任度的七个核心特性，即透明度、解释性、安全性、隐私、准确性、可靠性和效率。

关于其他标准，NIST（美国国家标准技术研究院）发布了人工智能系统的可信度框架，该框架提供了设计、开发和部署AI系统的一套建议原则。此外，欧盟机构也在开发一系列标准和指南，以确保开发和部署AI系统时遵守伦理准则和保护个人数据的原则。

这些标准和指南对于企业和组织来说非常重要，因为它们有助于确保AI系统在设计 and 部署时考虑到信任度概念，从而为最终用户提供可靠、安全和负责任的体验。

Figure 5: Example of the conversation in Chinese with a 9 billion parameter LLM.

3. Discussion

The results obtained as part of the experiments made it possible to formulate several ideas or hypotheses and further directions of research into large language models.

Controllability, stability, and reliability are important indications of language models, which are essentially closed-loop systems with a human in the decision-making loop [25]. Human participation in decision-making, such as feedback or response evaluation, is crucial as feedback. This can be observed in the language models at the level of 7 billion parameters: while the M1P7 model showed excellent performance, with average but predictable responses, it maintained a horizontal trend with minor fluctuations, indicating its controllability and stability. However, as a counter-example, the other models with 7 billion parameters (M3P7, M4P7, M5P7, M6P7) showed unpredictable effects, which were consequently expressed in the decreasing trust with a trend towards decreasing of trust observed in much smaller models like M1P1.5 or M1P3. This meets some observations regarding efficiency degradation in LLMs [26].

In large language models acting as closed-loop automatic control systems with a human in the loop, any actions that do not match the predictions of one or the other participant (AI or human) cause perturbations or disturbances [27]. These perturbations are essentially latent, but are expressed in a misunderstanding (or dissatisfaction) with the user's response or a misunderstanding by the model of the essence of the user's question or feedback, respectively. This causes non-linear dependencies of trust levels on the number of questions. In turn, it shows that the clarity of the answer and the question are decisive in assessing the level of trust. This could be seen on the example of a model with 8 billion parameters, when slight deviations of the response from the expected caused slight fluctuations in user confidence, and disturbances that were transmitted through feedback, steadily increasing trend of confidence.

No less important are the observation horizon and memory [28]. These two properties are as important as the connectivity (attention) and short-term memory of the model. This can be illustrated by an example - for instance, in the M1P1.5 model (1.5 billion parameters), efficiency increased in a short interval, but after the 5th question, due to a small amount of short-term memory, trust and efficiency decreased. In contrast to this model, in the M1P4 model (4 billion parameters) the presence of memory and the ability to retain the context of the conversation allowed the quality of answers to be preserved at least up to 21 questions. As a counter-example, we can mention a model with 13 billion parameters, which, even with a larger short-term memory and observation horizon, received small gain from these indicators.

Also, it is worth noting that the dependence of the value of trust on the number of parameters and interactions does not always meet expectations; for example, the number of interactions is unpredictable (can either increase, decrease or remain the same). Therefore, in such models with

memory, changes are accumulated gradually and are implemented in leaps. These effects were most clearly seen in large language models with 4 and 8 billion parameters.

As in big data models that have many interactions and users, and where the actions of individual users can reinforce beneficial interactions for the whole system in it, self-reinforcing effects can also be observed at the micro level, for example, the interaction of 1 user with 1 AI system. Despite the small volume of data (small data) of interactions, these effects are sufficient to form a data value chain [29] and evaluate data based on context and the decision-making chain "action-response-utility". In essence, the utility (value) of data for the user is formed as a result of the generation of answers that satisfy the user's expectations based on their evaluation, which, in turn, causes the generation of data and reinforcement of trust.

Another result, that was worth mentioning, was an impact of model effectiveness of training and its language encoding quality. In one-language models, like in these case, English, Chinese, Japanese and Arabic, some effects were discovered, whereas efficiency of language encoding can affect the model performance, which shows potential for models for their improvement.

Such results summarize that the user's satisfaction with the answers, and not the number of parameters, is the determining factor when learning a language model. In our opinion, more effective application of architecture optimizations, tokenization of source texts, higher quality training sample are the main criteria for achieving high quality of answers. However, it is worth noting that the advantages, obtained through more efficient text encoding, are in themselves may have been decreased because of the need for translator, and therefore, invisible increase in the number of parameters, so as such monolingual models are more useful likely for users who are native speakers of a given language, in the first place, and secondly for others [30].

So, one can assume that human-machine interaction when communicating with AI (using the example of large language models) has a complex nature. Small bits of data generated at each step affect the model's behavior and increase or decrease its performance over time.

4. Conclusions

This study analyzed how trust shows itself during long-term interaction. As a result of experiments, conducted on the example of 14 large language models with 1.5-13 billion parameters, we have shown that trust is not a constant value over time and indeed has complex dynamics. This dynamic can be traced over time, using such criteria as explainability, predictability and confidence, which form a general assessment of the value of trust.

Each interaction potentially leads to both the growth of trust and its decline, or certain oscillatory processes. This means that each subsequent response can have both a positive effect (user satisfaction and positive feedback) and a negative one, which can lead to unpredictable results as an outcome. The context retention in the conversation is critical because the model can learn patterns of user behavior and reinforce more useful responses for the user. This is due to the presence of self-reinforcing effects, which, in the ideal case, are increasing confidence in the model compared to the initial level of confidence and, as a consequence, confidence grows with time and with each interaction, so trust does too. As in big data models, the presence of feedback forms a data value chain that is linked to trust: the higher the value or utility of the data, the higher the trust.

Thus, summarizing the obtained results, we may assume that the studied features of the large language models show a potential number of problems in the field of trustworthiness in AI, which should be solved when creating new models. Firstly, these models should be stable, understandable, predictable, especially for cases when the models are used as personal intelligence. The relatively small number of interactions and small amount of data (compared to big data) indicates the need to develop large language models that adapt to the user and generate trustworthy responses using the context and memory of previous interactions.

We also want to mention that effects of self reinforcing of trust can be drastically improved by good quality learning and more efficient encoding of the data by the large language models. For instance, by studying models in 7-9 billion parameter range, we discovered, that usage other language that has more informational density, in fact, could (but not always) affect the quality of the answers, their conciseness, structure and context retention. For instance, in Japanese large language model, compared to English one with the same amount of parameters, the trust curve was increasing with higher rate, but, from our observation it started to taper at the same level as the English one, after long conversation, which shows the effect of diminished returns on improvement of model performance. As a counter-example, Chinese large language model performed less efficiently, compared to English one, despite more efficient encoding of language concepts, making effects of such encoding less predictable; it says that more effective application of architecture optimizations, as well as better training data is key for success here.

In the further research we propose to refine our experimental study on other large language models, likely with a larger number of parameters (more than 13 billion), including the online models available at Hugging Face Spaces and other cloud resources as well.

Declaration on Generative AI

During the preparation stage of the study, the authors studied the behaviour of AI models, in particular, Large large language models (LLMs), which were analyzed as a main subject of the study, whereas a set of AI models was benchmarked one against another, in order to establish the formal feasibility of usage of AI models in knowledge mining.

In the paper we conducted research on a large family of language models including:

- Meta Llama instruct 3.0 8b.
- Mistral 2 7b.
- Microsoft Phi 3 mini instruct 4b.

and another large language models with varying number of parameters from 1.5 to 13 billion parameters. In order to establish their utility for our tasks, we tested different LLM deployment software such as Chat RTX, Ollama (Docker image), and different Python libraries for LLM deployment, including llama-cpp, exllama2 and ollama.

Since AI models, such as LLMs were used during our experiments as main subject of the study, we acknowledge usage of such AI models during our experiments as a subject of the study and we take a responsibility of usage of AI tools according to rules of fair use of Generative AI in order to keep the academic integrity of our paper and authorship. Authors reviewed and edited the paper in order to comply with the requirements for publishing the paper with such type of content.

The content generated by LLMs, such as responses of the models, such as on figures 2,3,4,5 serves only the illustrative purposes to prove the feasibility of LLMs usage for knowledge mining. The figures 2-5 were created by transformative work, such as copying the text into an editor, formatting the text, making a screenshot, formatting the image and using image editing software (Ms Paint). Each figure of above include the question posed by the author and the answer by an LLM in 4 languages – English, Japanese, Chinese and Arabic as follows.

References

- [1] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, L. Zhang, W. Han, M. Huang, et al., Pre-Trained models: past, present and future, AI Open (2021). doi:10.1016/j.aiopen.2021.08.002.
- [2] Kuznetsov, V., Krak, I., Barmak, O., Kulas, A. Facial expressions analysis for applications in the study of sign language CEUR Workshop Proceedings, 2353, (2019) . pp. 159-172. DOI:10.32782/cm15/2353-13.

- [3] S. Kondratiuk, I. Krak, V. Kuznetsov, A. Kulias, Using the temporal data and three-dimensional convolutions for sign language alphabet recognition, *CEUR Workshop Proc.* 3137 (2022) 1–10. doi:10.32782/cmisp/3137-7.
- [4] NVIDIA ChatRTX. URL: <https://www.nvidia.com/en-us/ai-on-rtx/chatrtx/>.
- [5] J. Peddie, The GPU environment—software extensions and custom features, in: *The history of the GPU - eras and environment*, Springer International Publishing, Cham, 2022, pp. 251–281. doi:10.1007/978-3-031-13581-1_7.
- [6] Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, Y. Shi, Scaling for edge inference of deep neural networks, *Nat. Electron.* 1.4 (2018) 216–222. doi:10.1038/s41928-018-0059-3.
- [7] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, et al., Few-shot learning with multilingual generative language models, in: *Proceedings of the 2022 conference on empirical methods in natural language processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022. doi:10.18653/v1/2022.emnlp-main.616.
- [8] M. McBride, N. Persson, E. Reichmanis, M. Grover, Solving materials’ small data problem with dynamic experimental databases, *Processes* 6.7 (2018) 79. doi:10.3390/pr6070079.
- [9] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, GPT understands, too, *AI Open* (2023). doi:10.1016/j.aiopen.2023.08.012.
- [10] Jiang, A. Anastasopoulos, J. Araki, H. Ding, G. Neubig, X-FACTR: multilingual factual knowledge retrieval from pretrained language models, in: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020. doi:10.18653/v1/2020.emnlp-main.479.
- [11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2022, p. 27730--27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2022, p. 27730--27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [13] M. Hardalov, P. Atanasova, T. Mihaylov, G. Angelova, K. Simov, P. Osenova, V. Stoyanov, I. Koychev, P. Nakov, D. Radev, BgGLUE: A bulgarian general language understanding evaluation benchmark, in: *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2023. doi:10.18653/v1/2023.acl-long.487.
- [14] P. C. Bauer, Clearing the jungle: Conceptualizing and measuring trust and trustworthiness, *SSRN Electron. J.* (2013). doi:10.2139/ssrn.2325989.
- [15] ISO/IEC TR 24028:2020 Information technology – Artificial intelligence (AI) – Overview of trustworthiness in AI, 2020. URL: <https://www.iso.org/standard/77608.html>.
- [16] B. Brodaric, F. Neuhaus (Eds.), *Formal Ontology in Information Systems*, IOS Press, 2020. doi:10.3233/faia330.
- [17] A. Ferrario, Justifying our Credences in the Trustworthiness of AI Systems: A Reliabilistic Approach, *SSRN Electron. J.* (2023). doi:10.2139/ssrn.4524678..
- [18] E. Katsamakos, O. V. Pavlov, Artificial intelligence feedback loops in mobile platform business models, *Int. J. Wirel. Inf. Netw.* (2022). doi:10.1007/s10776-022-00556-9.
- [19] V. Kuznetsov, S. Kondratiuk, H. Kudin, A. Kulyas, I. Krak, O. Barmak, Development of models of self-reinforcing effects for big data evaluation, *IEEE 18th international conference on*

- computer science and information technologies (CSIT), IEEE, 2023. doi:10.1109/csit61576.2023.10324045.
- [20] Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y. A closer look at memorization in deep networks. 34th International Conference on Machine Learning. 2016. Vol. 70. PP. 2337-2346. URL: <https://arxiv.org/abs/1706.05394>.
- [21] I. Krak, V. Kuznetsov, S. Kondratiuk, L. Azarova, O. Barmak, P. Padiuk, Analysis of deep learning methods in adaptation to the small data problem solving, in: Lecture notes in data engineering, computational intelligence, and decision making, Springer International Publishing, Cham, 2022, pp. 333–352.
- [22] F. Cicirelli, L. Nigro, Control aspects in multiagent systems, y: Studies in big data, Springer International Publishing, Cham, 2015, c. 27–50. doi:10.1007/978-3-319-23742-8_2.
- [23] Assessing and improving AI trustworthiness: Current contexts and concerns, National Academies Press, Washington, D.C., 2021. doi:10.17226/26208.
- [24] Z. Li, Z. Zhang, H. Zhao, R. Wang, K. Chen, M. Utiyama, E. Sumita, Text compression-aided transformer encoding, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1. doi:10.1109/tpami.2021.3058341.
- [25] Automated control and autonomy, y: Systems engineering for ethical autonomous systems, Institution of Engineering and Technology, 2019, c. 41–81. doi:10.1049/sbra517e_ch3.
- [26] X. Feng, X. Han, S. Chen, W. Yang, LLMEffiChecker: Understanding and testing efficiency degradation of large language models, ACM Trans. Softw. Eng. Methodol. (2024). doi:10.1145/3664812..
- [27] K. A. Grasse, Nonlinear perturbations of control-semilinear control systems, SIAM J. Control Optim. 20.3 (1982) 311–327. doi:10.1137/0320024.
- [28] L. Hewing, K. P. Wabersich, M. Menner, M. N. Zeilinger, Learning-Based model predictive control: toward safe learning in control, Annu. Rev. Control, Robot., Auton. Syst. 3.1 (2020) 269–296. doi:10.1146/annurev-control-090419-075625.
- [29] H. Kasim, T. Hung, X. Li, Data value chain as a service framework: For enabling data handling, data security and data analysis in the cloud, in: 2012 IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS), IEEE, 2012. doi:10.1109/icpads.2012.131..
- [30] D. Bahdanau, K. Cho, Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. URL: <https://arxiv.org/abs/1409.0473>.