This study's object is the process that formalizes text prompts to large language models of artificial intelligence for the purpose of automatically generating action cards for hexagonal tabletop wargames. The task relates to the ambiguity (42% of terms are misinterpreted), contextual incompleteness (37%), and syntactic variability (21%) of natural language prompts, resulting in unhelpful and unpredictable responses.

To address this problem, a conceptual-practical model has been proposed, which combines structured prompt templates, a localized glossary of key terms, as well as clear instructions on response format.

Practical verification was carried out by generating a set of action cards for solo wargames on modern large language models by generating over 100 prompts and ana-lyzing more than 300 responses. The experiments demonstrated that the formalized prompts reduced the total error rate by 58%, as well as increased the relevance of re-sponses from 55–65% to 88–92%. The average time for preparing prompts was reduced by 25–40%. The "d6-table" templates ensured the stability of the output format in 90–95% of cases while JSON structures provided stability in 85–90% of cases. The glossary and structure definition integrated into the prompts minimized semantic discrepancies and syntactic errors.

A special feature of the proposed template structures for prompts is adaptability to different subject areas through the use of a description language specific to each of these areas. The research results have practical value for automating game content de-velopment processes and could be adapted for other subject areas where accuracy, con-sistency, and structure of language model responses are important.

The proposed systematic approach facilitates the automation of complex content development with a guaranteed increase in the quality and predictability of responses from large language models

Keywords: prompt formalization, large language models, artificial intelligence, structured templates, wargames

UDC 004.8:004.9, 004.932
DOI: 10.15587/1729-4061.2025.335473

FORMALIZATION OF TEXT PROMPTS TO ARTIFICIAL INTELLIGENCE SYSTEMS

Vladyslav Oliinyk
Corresponding author
PhD Student*
E-mail: vladyslav.oliinyk3@nure.ua
Andrii Biziuk
PhD, Associate Professor*
Zhanna Deineko
PhD*
Viktor Chelombitko
PhD*
*Department of Media Systems and Technologies
Kharkiv National University of Radio Electronics
Nauky ave., 14, Kharkiv, Ukraine, 61166

Received 15.07.2025
Received in revised form 15.09.2025
Accepted date 22.09.2025
Published date 31.10.2025

How to Cite: *Oliinyk, V., Biziuk, A., Deineko, Z., Chelombitko, V. (2025). Formalization of text prompts to artificial intelligence systems. Eastern-European Journal of Enter-prise Technologies, 5 (2 (137)), 84–97. https://doi.org/10.15587/1729-4061.2025.335473*

1. Introduction

Given the evolution of artificial intelligence (AI), people increasingly interact with it through text prompts (prompts). This can be information search, text generation, data analysis or automation of certain processes. However, the quality of the AI response directly depends on how correctly and clearly formulated the prompt is. One can say that AI understands only what was written to it, not what was meant. Only the quality of the prompts used determines how well language models generate answers and how easy they are to use [1, 2].

A key factor in effective interaction with AI is the ability to ask questions that accurately reflect the user's needs. The formulation should be clear, logical, without excessive ambiguity or contextual assumptions. If the user enters an incorrect or unclear prompt, the AI may give an incomplete, inaccurate, or even unhelpful answer. For example, the prompt "Tell me about Python" is too general and does not provide the system with enough information because it can refer to both a programming language and a snake. In this case, it is worth clarifying: "Tell me about the Python programming language, its features and applications". This is the prompt that contains a clear task, genre, and understanding of the target audience of the source text. Such detailing helps get a result that really meets expectations.

In this context, the so-called "prompt engineering", i.e., the process of constructing prompts, becomes important. This is a new skill that is becoming increasingly in demand in various industries: from education and marketing to programming and design. The ability to formulate precise, context-rich, and structured prompts makes it possible to unlock the full potential of Large Language Models (LLMs) of AI and significantly increases user productivity [3].

Despite the rapid development of Natural Language Processing (NLP) technologies, AI still has certain limitations:

– ambiguity: many words and phrases in natural language may have multiple meanings that can only be understood through context;

– contextual dependency: the meaning of a word or sentence can change depending on the context;

– implicit assumptions: people often do not mention details that are obvious to them at first glance, but the absence of these details can make it difficult, and sometimes impossible, to correctly understand an AI prompt;

– grammatical and stylistic variations: the same meaning can be expressed in different ways, which may complicate the interpretation of the prompt.

To avoid these problems, it is necessary to formalize prompts using clear, structured formulations. Despite the universal nature of the issue of formalizing prompts, this task is especially acute in applied areas where it is necessary to receive not just text answers from AI but structured scenarios, algorithms, or specialized game solutions. In such cases, obtaining "correct" answers from AI is critically important.

One such area is tabletop wargames – games that simulate conflicts and scenarios on playing fields where the automation of scenario development using AI critically depends on the quality of text prompts.

Thus, the formalization of text prompts for AI is a key stage in achieving accuracy and relevance of results, especially in complex application areas such as the development of scenarios for tabletop wargames. High-quality prompt design makes it possible to minimize ambiguity, take into account the context, and ensure that the result obtained meets the set goals.

Therefore, it is a relevant task to carry out studies aimed at devising methods and technologies to formalize text prompts for automated generation of wargame scenarios using AI. They have significant practical potential for the further development of the field of tabletop games design engineering, in particular wargames, as a component of the interactive entertainment and simulation modeling industry.

2. Literature review and problem statement

In recent years, the scientific community has become increasingly interested in the process of formalizing textual prompts for artificial intelligence systems. This is confirmed by systematic reviews in ACM Computing Surveys, where prompt engineering is identified as a critical tool for managing large language models. In [4], the results of studies on the impact of correctly formulated prompts on the accuracy of responses of large language models are reported. It is shown that the correct formalization of prompts improves the accuracy of responses by 40–78% for mathematical reasoning tasks. Similar results are confirmed in study [5], which shows that even adding a simple phrase "Let's think step by step" to a prompt can significantly improve the accuracy and quality of artificial intelligence responses.

In [6], the results of studies on transformative architecture with a self-attention mechanism are reported, which has become the foundation of modern research in the field of natural language processing and prompt engineering. The self-attention mechanism underlying this architecture has provided a revolutionary leap in the quality of interpretation of complex textual prompts by large language models. This technology has become a catalyst for the development of prompt formalization methods, as it has allowed AI systems to more effectively analyze context, detect implicit conditions, and generate structured responses.

In [7], a systematic review of various methods for improving the performance of large language models, including few-shot and zero-shot prompting techniques, is given. It is shown that these methods allow models to better adapt to tasks with limited or no training examples, which significantly expands the possibilities of practical application of LLMs. In particular, the chain-of-thought prompting method allows models to decompose complex tasks into intermediate steps, which significantly improves the quality of reasoning, evidenced by the results of the study reported in [8].

In [9], a systematic analysis of prompt engineering approaches in the medical field was carried out, emphasizing the key importance of structured instructions for improving the accuracy of results. At the same time, a limitation of this study is its focus on a narrow subject context – medical applications, so the conclusions cannot always be directly transferred to other areas, in particular gaming or education.

Work [10] covers a wider range of user interaction with language models and demonstrates the potential of structured "prompting" to improve the quality of output responses. Despite this, the authors' conclusions are not sufficiently supported by long-term empirical data, which limits their applied validity in specific application areas. The official guide [11] offers a set of practical recommendations for clearly formulating prompts and determining the desired response format.

In [12], the results of research on the importance of clearly specifying the output data format for large language models when working with structured tables are reported. It is shown that the specification of the format and structure of the prompt plays a significant role in how AI interprets user prompts. At the same time, problems related to the adaptation of models to various types of tables with different complexity of structure and content remain unresolved, which limits the universality of the application of existing approaches. One of the possible reasons is the limited ability of models to adequately recognize and combine contextual information with formal features of tables, which leads to variability in the quality of interpretation and response generation. In addition, the lack of uniform prompt formatting standards complicates the scaling of methods and reduces their representativeness in various practical application scenarios.

Similar conclusions were drawn in study [13], which emphasized the role of format specification in increasing productivity when working with different types of prompts. However, the issues of adapting the considered methods to complex prompts and ensuring comprehensibility for non-technical users remain unresolved. Possible reasons include the lack of unified standards, which complicates the creation of universal templates, as well as the variability of language models, which interpret the same instructions differently depending on the wording or context.

In [14], a mathematically based method for constructing prompts to large language models was proposed, which showed a 14% improvement in the quality of answers. Again, the issues of the stability of the results with different complexity of the tasks and the ability to scale the method to different subject areas remain unresolved. Perhaps this is due to the limitations of the formal model, which does not take into account all linguistic and contextual variations.

Summarizing the results of the above studies, we can state that the efficiency of working with AI-based systems largely depends on the clarity of the instructions and the completeness of the prompt context. As shown in [12–14], a structured approach and the selection of key language constructs significantly increase the accuracy and stability of answers. At the same time, the studies noted that there is a need to transition from intuitive formulation of prompts to a systematic, engineering approach. This makes it possible to avoid ambiguities and increase the suitability of the results for further use. However, challenges remain in adapting such approaches to different subject areas and heterogeneous language models, which requires further development of generalized methods and unified standards.

In [15], the results of research on the SLOT methodology for structuring the output data of large language models are reported. It is shown that the Mistral-7B model with limited decoding achieves 99.5% structuring accuracy and 94.0% semantic content similarity, which significantly exceeds the performance of competing solutions. The issues of scalability of the method to tasks with a more complex output data struc-

ture and the challenges of integrating SLOT into real-world operating environments with high requirements for speed and flexibility remain unresolved. Possible reasons include limitations in algorithmic support for adaptation to different output formats, as well as insufficient consistency between structural representation and semantic content in individual application contexts.

Study [16] demonstrates the results of using the "f-String" and "Follow the Format" methods to generate JSON structures, showing an average success rate of 82.55%, varying from 0% to 100% depending on the complexity of the task. It was found that the main unresolved problems are the instability of results when working with complex or multidimensional structures, as well as the difficulty in maintaining format consistency with dynamic changes in input data. The reasons for this may be the limitations of the current model architecture in accurately understanding and repeating complex formats, as well as the imperfection of output data quality control algorithms.

The authors of [17] report the results of a comprehensive assessment of the capabilities of large language models to generate complex structured data through the FormatCoT methodology. It is shown that the LLaMA-7B model, after targeted training, is able to outperform more powerful LLMs in the tasks of generating structured output data. Despite the success of the FormatCoT methodology and the LLaMA-7B training, there are still issues with generalizing the methodology to various types of structured data and supporting complex multi-step formats. These limitations can be explained both by shortcomings in the training data and by technical difficulties in integrating extended structural instructions into the model, which also requires further research and improvements. Issues related to the instability and unpredictability of results when formulating prompts to large language models remain unresolved. The reason for this may be objective difficulties associated with linguistic ambiguity, the lack of unified standards for different types of tasks, and the empirical nature of existing methods.

According to [18], the taxonomy of prompt engineering patterns covers only 58% of software engineering cases, which emphasizes the need for industry adaptations. The study noted that existing prompt templates do not fully capture the specificity of different scenarios in the field of software engineering, leaving the issues of adaptation and universality unresolved. This is due to the complexity of the tasks in the industry, the diversity of requirements, and the lack of unified standards for different types of tasks, which creates challenges for the development of effective prompt engineering templates.

Paper [19] shows that 73% of the participants used the "trial and error" method to achieve the desired results, which indicates a global lack of systematic tools. Despite the growing understanding of the potential and limitations of generative AI tools, challenges remain in the formation of clear methodologies and practices that would ensure more effective use of technologies in creativity. The main reasons are the lack of standardized approaches to developing prompts and the complexity of integrating innovative technologies into curricula, which requires further research and the development of tools to support the creative process. Despite the growing interest in formalizing prompts, the considered methods face a number of systematic limitations that hinder their widespread use.

Analyzing the above studies, one can see that most of the proposed solutions are either too abstract or overly specialized. For example, such "solutions" are often rather general recommendations for basic prompt structuring, or they are overly specialized instructions for rather narrow and niche areas, such as medical diagnostics or software code generation. As noted in [18], key challenges in the field of prompt engineering include:

– linguistic ambiguity, which leads to incorrect interpretation of prompts by the model. Experiments with GPT-4 have shown that even minor syntactic changes in prompts noticeably reduce the accuracy of answers;

– lack of unified standards for different types of tasks, such as data generation, analysis, or verification (the taxonomy of prompt engineering patterns, considered, covers only 58% of cases from software engineering, which emphasizes the need for industry adaptations);

– the empirical nature of the methods, which are often based on observations rather than formal algorithms (for example, in a study on digital art, 73% of participants used the "trial and error" method to achieve the desired results, which indicates a lack of systematic tools).

Study [20] on LLM integration in Android applications found that 68% of developers face difficulties in reconciling output formats with backend systems due to the instability of structured responses from models [20]. This forces developers to spend up to 40% of their time on additional validation and data transformation. A promising direction to overcome these limitations is the development of specialized prompt description languages (DSLs) that combine the flexibility of a natural language interface with a strict structure. As described in [21], experiments with the System Prompt Meta Language (SPML) demonstrate that the use of DSL reduces the number of syntax errors in responses by 37% and ensures compatibility with external APIs. Similar results were reported in study [22], in which a two-stage template system increased the accuracy of responses by 3.5 times. At the same time, the problems of linking language to complex, variable usage scenarios, as well as ensuring the flexibility of DSL while maintaining strict standards remain unresolved. This may be due to the high complexity of natural language instructions, the interdisciplinary nature of applications, and the limited capabilities of adaptive processing within the framework of rigid formalization.

The reviewed studies in the field of prompt engineering demonstrate a critical shortage of unified standards for formalizing prompts to language models. The lack of a common structural model or classification makes it difficult to devise universal approaches that would cover various types of tasks – from analytical to generative. This problem makes interaction with AI inefficient for non-technical users and limits the scalability of solutions in applied areas.

A promising way to overcome these limitations is to develop specialized prompt languages that combine the rigor of formal standards (such as SQL) with the flexibility of a natural language interface. Such tools could provide clear rules for describing the desired output, reducing the unpredictability of results. However, their effectiveness will depend on the ability to take into account contextual nuances and adapt to different types of models, which remains a key challenge for future research.

Most of existing recommendations focus on a technical or highly specialized audience. Work [23] emphasizes that existing prompt engineering tools and techniques remain complex for non-professionals. The issues of adapting prompt engineering to a wide range of users with different levels of

digital literacy, as well as the development of systems that facilitate collaborative work and knowledge sharing in communities, remain unresolved. This is due to the insufficient consideration of social and behavioral factors in the design of existing solutions, as well as the limited involvement of non-technical users in the process of creating and optimizing prompts. There is a need to devise such formalization methods that would be understandable and applicable in everyday interaction with AI.

Paper [24] shows that prompt engineering is especially difficult for non-technical users who are less familiar with AI technologies. It was found that creating effective prompts requires significant effort, skills and multi-stage iteration, which complicates the process for non-professionals. The issues of designing intuitive tools to support users with different levels of expertise and increasing the transparency of interaction with language models remain unresolved. The problem is especially noticeable in the field of education, small businesses, and among users who have no experience with IT or analytics.

Although language models such as Claude or GPT-4 can theoretically automate routine tasks (planning, text analysis), their practical application requires a semi-technical style of formulation. For example, attempts to build comparison tables or step-by-step instructions often fail due to the inability of users to accurately express requirements. As shown in study [25], the situation is aggravated by the fact that technical documentation is often too complicated for non-technical audiences and creates barriers due to its conceptual complexity and the prevalence of niche jargon. The reasons for these problems are indicated by the low adaptation of documentation to the needs of different audiences, the insufficient level of pedagogical strategies in teaching technical writing, as well as the lack of implementation of effective user testing methods to improve understandability. The authors emphasize the need to use simpler language, visualizations and real-life examples, emphasizing that existing methods still do not fully solve the problem, leaving the need to improve training programs and further research in this area.

Attempts to teach a wide range of users to use language models are accompanied by the publication of short instructions and example prompts. However, some of them do not take into account the digital literacy of the user and real-world use cases but focus on narrow examples that are not always suitable for a wide range of people [26]. The issues of low digital and AI literacy among people with limited digital skills, as well as fears and anxiety caused by negative media stories, remain unresolved. The excessive complexity of AI systems for beginners also makes it difficult for them to engage. For example, recommendations for creating tables or prompts that generate algorithms are often built in complex technical language or require basic programming knowledge, which makes their widespread use impossible.

Paper [27] confirms that problem, demonstrating that only 45% of engineering students were able to complete tasks using generative AI and structured prompts even after special training. This indicates that AI systems are not sufficiently adapted to their level of knowledge and needs. Among the reasons for this are the lack of accessible representation language, limited training time frames, and insufficiently developed prompt engineering techniques that could facilitate the form of interaction with AI. Thus, to overcome barriers to the implementation and use of generative AI, more inclusive, easier-to-use educational approaches should be devised that take into account the different levels of digital and technological preparation of users.

Most existing solutions do not take into account the contextual variability of LLMs: the same prompt in different styles can give opposite results. To overcome this, we propose to design an interface-friendly recommendation system that will combine:

– adaptive templates for popular tasks (“comparisons”, “instructions”, “algorithms”);

– visual prompts for real-time prompt correction;

– automatic quality checking of formulations based on NLP metrics.

It is assumed that such a system will not only reduce the burden on a user who is not a technical specialist in this field but will also help in forming a critical understanding of the principles of interaction with AI.

Unlike existing tools, it will integrate contextual adaptation mechanisms that take into account the linguistic features of the Ukrainian audience and the specificity of local tasks.

Thus, prompt formalization is considered a critical component for enabling effective interaction between the user and AI technologies. Our work is aimed at bridging the gap between the technical capabilities of large language models and the real needs of users by designing an intuitive interface that transforms abstract concepts of prompt engineering into practical actions.

Given the above, the main problem of the study is the inconsistency and lack of methodological tools for formalizing text prompts that would ensure the accuracy, structure, and relevance of the responses of large language models.

This problem is especially acute in applied areas where it is necessary to obtain not general considerations but structured objects: tables, algorithms, scenarios or simulations (for example, in the field of wargames, military modeling, data analysis, training systems).

The practical context of our study is predetermined by the needs of a modern publishing company specializing in the release of board wargames in the “base game” format with the possibility of systematic expansion through additional modules. Each supplement includes a set of character miniatures and accompanying game documentation. These supplements are created in a short time – approximately within 1–2 months from the conceptual sketch stage to the release of the final product.

One of the key elements of the accompanying documentation is a set of game cards for solo mode under which the user independently controls one of the parties, while the opponent’s actions are implemented based on a programmed algorithm built into the structure of the cards.

All this allows us to assert that further development of a systemic approach to the formalization of text prompts to artificial intelligence systems is advisable. The approach should be based on clear methods of structured prompt engineering, integration of localized glossaries, and automated prompt refinement tools. Another problem is also the difficulty of adapting the methods for non-technical users, which requires the development of intuitive interfaces with visual prompts and automatic quality checks of formulations. In view of this, it is advisable to construct a scientifically sound formalization model that would combine linguistic rules, algorithmic structures, localization of terminology, and flexibility of the user interface. This will ensure efficient, structured, and scalable generation of answers for various application areas,

including automated scenario generation in tabletop wargames and related interactive systems.

3. The aim and objectives of the study

The aim of our study is to build a conceptual and practical model for formalizing text prompts to artificial intelligence systems, specifically adapted for the automated development of AI opponent action cards in hexagonal tabletop wargames. This will provide an opportunity to enable structured, accurate, and relevant generation of responses by large language models, optimized for solo-mode game scenarios, where the opponent's actions are determined by a set of pre-prepared cards.

To achieve the goal, the following tasks were set:

– to analyze linguistic challenges in user interaction with AI language models;

– to determine the principles of constructing effective prompts to AI language models to increase the accuracy and relevance of responses;

– to compare the indicators of relevance, structuring, completeness, and compliance with the format of responses of AI language models between formalized and non-formalized prompts;

– to devise typical prompt structures for AI language models, suitable for use in gaming and software environments, and test them on the example of generating action cards for an AI opponent in hexagonal wargames.

4. The study materials and methods

The object of our study is the process that formalizes text prompts to artificial intelligence systems, in particular large language models, in the context of automated generation of AI opponent behavior scenarios in hexagonal tabletop wargames. The main hypothesis of the study assumes that a clear structure, templatization, and the use of a localized glossary in text prompts could significantly increase the accuracy, structuring, and compliance of the results with the in-game logic.

The study adopted the following simplifications:

– opponent behavior modeling is implemented based on a limited number of styles (e.g., aggressive, cautious, or balanced);

– the prompt structure involves the generation of a fixed number of actions (e.g., 6 options for rolling a d6 die).

The study was conducted within the framework of a systems approach, but the main attention is paid to the application of specific methods for analysis and verification of the effectiveness of formalized prompts. This makes it possible to combine the study of the theoretical principles of formalization with their practical testing under game modeling conditions.

The research is based on a mixed methodology that combines theoretical modeling, experimental verification, and qualitative expert assessment. At the initial level of the research, a comparative analysis of existing approaches to prompt engineering was conducted in order to systematize known formalization methods, identify their strengths and weaknesses, as well as the possibilities of adaptation to the specifics of game scenario generation. In parallel, structural-functional analysis was applied to identify key components of prompts – context, glossary, response format, and constraints – which allowed the development of typical templates, in particular d6 tables and JSON structures, which serve as the basis for further experimental testing.

To empirically test the effectiveness of the proposed model, a controlled experiment was conducted to generate over 100 prompts – half of which were formalized according to the developed templates, and the other half were unstructured, natural language. Over 300 responses were generated and analyzed from large language models, in particular the GPT, Claude, and Gemini series of models. The comparative analysis of the results was carried out using quantitative criteria: the level of relevance, the structure of the responses, semantic accuracy, and compliance with the given format, which allowed us to objectively assess the reduction in the number of errors and the increase in the predictability of the responses.

The final component of the methodology was a qualitative expert assessment of the generated scenarios with the participation of at least five board game designers. The experts evaluated the materials based on the criteria of comprehensibility, gameplay value, compliance with wargame mechanics, and reproducibility in the gameplay. The feedback received was used to iteratively improve the model, which increased its practical applicability and adaptability to the needs of real users.

The experimental part of the study is divided into several stages, which is schematically depicted in Fig. 1.

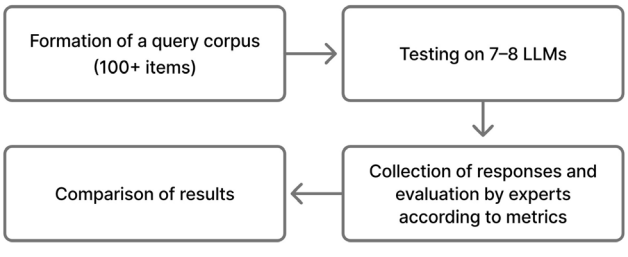

Fig. 1. Flowchart of the experiment

At the first stage, a corpus of prompts was formed, which included both formalized (structured according to templates) and non-formalized variants. Formalized prompts contained clearly defined instructions, limited use of multi-valued terms, specified the expected response format (table, pseudocode, JSON, markdown) and the use of key terms from the developed localized glossary. More than 100 prompt variants were devised for various game situations.

Example of non-formalized prompts:

– "How should a general act who must restrain an enemy attack and organize active defense on the battlefield?"

– "Generate 6 actions for an infantry unit in an aggressive style."

– "We need variants of tactical decisions for a cavalry commander who acts cautiously and tries to preserve his/her forces, dealing with enemy artillery nearby."

Examples of prompts after formalization are given in Table 1.

In the second stage of the study, testing was conducted using several modern large language models with different access methods. The most common and accessible large language models from leading developers were selected for testing, providing a wide range of capabilities – from closed APIs to open source models. Below is a list of the models used, distributed by developer companies:

– OpenAI (USA): o3-mini, gpt-4.1 models. Testing was carried out via a standalone web interface with a paid API. Without access to the source code;

– Google (USA): Gemini 2.5 Pro model – via a standalone web interface using a free API; Gemma 3 12B model – free via interactive chat on the Google AI Studio platform. Source code availability: Gemini 2.5 Pro – without access to the source code; Gemma 3 12B – with open source;

– Anthropic (USA): Claude 4.0 Sonnet model – via a paid subscription to the Perplexity Pro analytics and search service. Without access to source code;

– xAI (USA): Grok 3 – via paid subscription to the analytics and search service Perplexity Pro. Source code availability: the base model is closed, but lighter open source versions are planned for release in late 2025;

– Mistral AI (France): Mistral 7B – free via open chat on the Hugging Face platform. Open source model;

– Alibaba Group (China): Qwen3 8B – via free local in the Ollama application. Open source model;

– DeepSeek (China): DeepSeek-R1 model – free via open chat on the company's official platform. Open source model;

– Perplexity (USA): Sonar model – via paid subscription to the analytics and search service Perplexity Pro. Source code availability: the model is based on the open LlaMa 3.1 70B architecture, but Perplexity's own improvements (fine-tuning and training on its own data) are closed.

Table 1

Examples of prompts after formalization

No.	Prompt text
1	CONTEXT: An infantry battalion is in a defensive position on a hill. Enemy cavalry is approaching from the north. Distance 300 meters. Available: muskets, bayonets, cannon. OBJECT: Infantry battalion (300 men) STYLE: Defensive FORMAT: d6-action table GLOSSARY: – Square – square battle formation against cavalry – Volley – simultaneous firing of all units – Bayonet charge – close-in with melee weapons Generate a d6 action table for an infantry battalion: | d6 | Action | Expected effect | |----|-----|------------------| | 1 | [action] | [effect] |

2	CONTEXT: Artillery battery in an open position. Enemy infantry attacking from 400 meters. Available: 4 guns, 20 gunners, limited supply of gunpowder.
	OBJECT: Artillery Battery
	STYLE: Aggressive
	FORMAT: JSON
	Create JSON action structure:
	{
	«id»: «unique_action_id»,
	«action_name»: «назва_дії»,
	«instructions»: «покрокові_інструкції»,
	«expected_outcome»: «очікуваний_результат»,
	«resources_required»: [«ресурс1», «ресурс2»],
	«tactical_effect»: «вплив_на_бій»
	}

This allowed us to compare the functionality of the models under different access and integration modes, taking into account both local solutions with private hosting and cloud services with public web interfaces available for free.

All models were used under a zero-shot mode without prior training. In each case, the prompt was sent without intermediate editing, and the responses were saved for further analysis.

The third stage involved a qualitative and quantitative assessment of the responses received.

– structuredness (compliance with the specified format);

– contextual correspondence to the specified command style;

– contextual relevance (taking into account the described game situation);

– completeness (completeness and validity of all elements of the response).

The expert panel consisted of three game design experts who did not participate in the formation of the prompts. A five-point rating scale was used (1 – minimum level, 5 – maximum). The final score was calculated as the arithmetic mean of the experts' ratings. The statistical analysis method was used to calculate the average values, deviations and percentage changes between formalized and non-formalized prompts. Errors in the interpretation of prompts or violations of the format were also recorded.

To confirm the reliability of the results, a triangulation method was used, combining:

– quantitative assessment (statistical analysis of metrics);

– qualitative assessment (expert evaluation);

– practical verification (testing in a gaming environment).

To study the linguistic problems of interaction with AI, more than 50 informal prompts and responses to them from different AI models were analyzed. The analysis covered applied prompts in the field of hexagonal tabletop wargames that involved the generation of AI opponent actions. For each prompt, the presence or absence of a specific error (term ambiguity, contextual incompleteness, syntactic variability) was recorded. One response could contain several errors; the share of the category was calculated as follows: (number of recorded errors of this type ÷ total number of all recorded errors) × 100%.

5. Results of investigating the formalization of text prompts to AI systems in the context of generating AI opponent actions in wargames

5. 1. Results of analyzing linguistic challenges in interaction with large language models

Our study revealed three main categories of difficulties that negatively affect the accuracy and relevance of answers:

– ambiguity of terms: 42% of cases of incorrect or partially incorrect answers. The most problematic were terms such as "contain an attack" or "active defense", which can be interpreted as defensive or offensive actions depending on the context;

– contextual incompleteness: 37% of cases. The model sometimes ignored the current state of military units or scenario conditions if this data was not explicitly specified in the prompt;

– syntactic variability: 21% of cases. The same content could be formulated in different ways, which made it difficult for the AI to interpret and led to inconsistent responses.

The largest proportion of errors occurred when the prompt did not specify the context or used terms that were not defined in the internal terminology system (glossary).

In a further experiment, formalized prompts were used. Each prompt contained the following:

– a clear description of the tactical situation;

– terms from the local glossary;

– an indication of the desired response format (table, algorithm, or JSON).

The results showed that the number of linguistic errors decreased by 20–30% while the share of structured and relevant responses increased.

The analysis of results, given in Table 2, reveals that the use of formalized prompts makes it possible to significantly reduce the number of linguistic errors and change their structure by categories.

Table 2

Comparison of AI response analysis results

The category of difficulty	Error rate in unformalized prompts	Error rate in formalized prompts	Change (%)
Ambivalence of terms	42%	18%	–24%
Contextual imperfection	37%	14%	–23%
Syntactic variation	21%	10%	–11%
Cumulative error rate	100%	42%	–58%

A visualization of this dynamics is shown in Fig. 2, where both the overall decrease in the number of errors and their distribution are clearly visible.

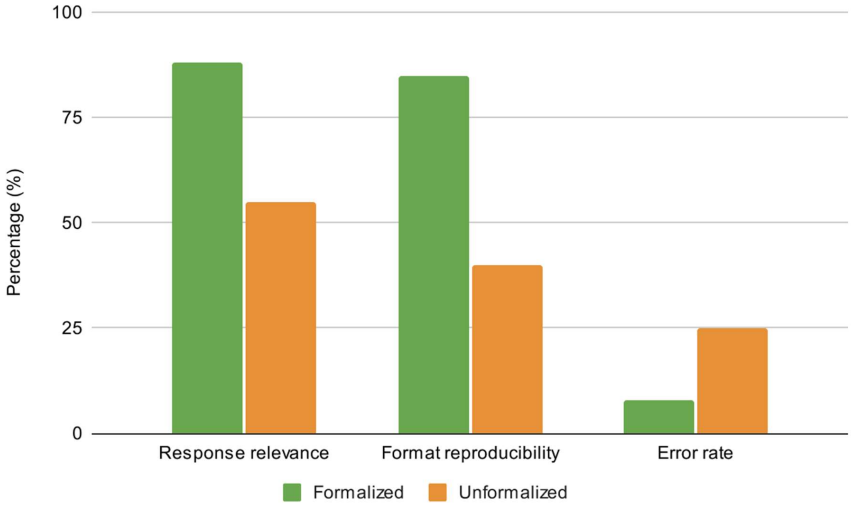

Fig. 2. Relative proportion of different types of errors in AI responses for unformalized and formalized prompts (proportions, 1 = 100%)

Our results became the basis for further defining the key principles of effective prompt formulation and choosing methods for their formalization. Particular attention was paid to the preparation of a localized glossary of terms for the behavior of an AI general in a wargame. In this glossary, each term has an unambiguous definition, examples of use, and assignment to a specific command style (aggressive, cautious, balanced), which reduces the risk of misinterpretation when generating actions.

Analysis also revealed that clear parameterization of the prompt (for example, indicating the type of unit, terrain, tactical situation) significantly increases the reproducibility of the results. If in non-formalized prompts a large difference was observed between several answers to the same question, then in formalized prompts this difference was reduced to a minimum, and the result of full compliance with the expected format reached over 85%.

It is important that the data obtained not only have theoretical significance but also directly influenced the planning of the next stages of work. Based on this analysis, it is planned to develop an application for automatic testing of bot actions according to scenarios generated based on formalized prompts. It is assumed that this application would be able to conduct several hundred game sessions under the same conditions for one scenario, collecting detailed statistics on the stability, predictability, and adequacy of AI behavior.

Our results are directly related to the formation of principles, methods, and templates for formalizing prompts, which became the basis for further verification and automated testing of models within the framework of this study.

5. 2. Results of formalizing the principles for constructing an effective prompt

Within the framework of our study, a typical structure of a formalized prompt was developed and tested (Table 3), which makes it possible to increase the accuracy, predictability, and stability of results from generating action cards of an AI opponent in tabletop hexagonal wargames.

The structure of the prompt included the following mandatory elements:

– a brief description of the situation – a contextual prerequisite for the action (taking into account key factors of the scenario: the location of forces, the phase of the game, available resources);

– a key object – a specification of the unit or group of units for which the action needs to be generated;

– a command style – a clearly defined style of tactical actions (aggressive, defensive, balanced), which set the behavioral framework;

– an expected response format – a predefined structured format (table, JSON, algorithm) that unifies the source data and facilitates their integration into the game system.

An additional element of each prompt was a localized glossary of terms specific to the selected game system. This allowed us to eliminate semantic fluctuations and ambiguity, reducing the number of irrelevant or overly general responses. According to the analysis, the use of a glossary reduced the number of linguistic errors by an average of ~30% while increasing the tactical relevance of the results.

Three types of template structures were developed and tested:

– action description, designed for the rapid generation of concise instructions, for example, “move one hex towards the nearest cover”;

– option table (d6), which is a structured list of six possible actions tied to the result of a six-sided die roll;

– action explanation, which contains a detailed interpretation of the tactical decision with a short analytical justification.

The highest formal compliance with the expected format was demonstrated by d6 tables (92% of responses clearly corresponded to the template), while “action description” and “action explanation” showed 88% and 85%, respectively. At the same time, d6 tables also provided the greatest predictability of the AI’s tactical behavior and were suitable for serial card generation.

To verify the effectiveness of the application of formalized prompts, a series of tests was conducted in five language models within the same game scenarios. Each model had over 50 test prompts built according to a typical structure. Ta-

ble 4 compares the success of responses based on two criteria: compliance with the format and tactical compliance.

Table 3

Example of the developed typical structures of a formalized prompt

Description	Structure
Basic d6 table template	CONTEXT: [Description of tactical situation, force disposition, game phase, available resources] OBJECT: [Unit type and its characteristics] STYLE: [Aggressive/Defensive/Balanced] FORMAT: d6 action table GLOSSARY: [Localized terms with unambiguous definitions] Generate d6 table: | d6 | Action | Expected effect |

JSON structure with full parameterization	CONTEXT: [Detailed description of the situation]
	OBJECT: [Unit specification]
	STYLE: [Tactical style]
	FORMAT: JSON
	GLOSSARY: [Terminological dictionary]
	Create JSON structure:
	{
	«id»: «»,
	«action_name»: «»,
	«instructions»: «»,
	«expected_outcome»: «»,
	«tactical_effect»: «»
	}

Algorithmic description of the action	CONTEXT: [Game situation] OBJECT: [Generation unit] STYLE: [Behavioral model] FORMAT: Algorithm GLOSSARY: [Specialized terms] Describe the sequence of actions as an algorithm with «IF-THEN» conditions.

Table 4

Comparing the success of responses from different AI models to formalized prompts

AI model	Compliance with the format	Tactical compliance
o3-mini	89%	83%
GPT-4.1	92%	87%
Mistral-7B	78%	71%
Claude 4.0 Sonnet	85%	81%
Gemini 2.5 Pro	88%	84%

Compared to unformalized prompts, formalized prompts produced approximately 35% more relevant and logically consistent responses. In particular, in the case of GPT-4.1, 92% of responses fully matched the expected pattern, even when the prompt formulation was intentionally varied.

Thus, the use of automatic prompt refinement reduced the time for their preparation by 25–40%. It also reduced the number of syntactic and semantic errors by 30% and increased the reproducibility of responses. These results confirmed the promising integration of automated tools at the stage of preparing instructions for AI.

One of the important observations was that a significant part of the responses remained within the expected format even when the prompt formulation changed or minor variations in the context. This indicates the robustness of the models' work, provided that clear structured instructions were used.

In some cases, the models correctly adapted the template to the specifics of the scenario, preserving key parameters – format, style, and logic of actions. Taking into account the results obtained, it is planned:

– to conduct large-scale automated testing with several hundred game sessions for one scenario to assess the stability of AI operation in reproducible conditions;

– to check the effectiveness of complex combinations of templates (for example, several d6 tables for one scenario with the choice of one of them depending on the tactical situation);

– to form a glossary of typical features of the behavior of an AI commander, which will be integrated into the prompt constructor for wargame designers.

Thus, a clearly structured, supported by a glossary, and automatically validated instruction to AI ensures stable quality and predictability of game content generation, which significantly facilitates the process of its preparation.

5. 3. Results of comparing the formalized and non-formalized text prompts to AI language models

Our study compared, experimentally tested, and systematized several methods for formalizing text prompts adapted for automated generation of action cards of AI opponents in hexagonal wargames. The considered solutions included template-based instructions (Template-Based Prompting), the use of a localized glossary of tactical terms, as well as structuring the prompt according to the "condition – action – format" scheme. The choice of these methods was based on a preliminary analysis, which revealed that they are the most suitable for reducing ambiguity, increasing the reproducibility of results, and ensuring compliance with the game context.

Each AI response to the corresponding prompt was checked according to three main criteria:

– compliance of the content with the expected style of behavior – an assessment given by experts, comparing the generated actions with the standards of aggressive, defensive, and balanced tactics;

– structured response – the presence of clearly defined elements, such as a description of the situation, action, and representation format;

– minimization of syntactic errors – the absence of grammatical, stylistic, or lexical shortcomings that may complicate the integration of the result into the game system.

For each method, a series of tests was conducted in five language models (o3-mini, GPT-4.1, Mistral-7B, Claude 4.0 Sonnet, Gemini 2.5 Pro). Each model processed at least 30 prompts with different scenarios and command styles. The assessment was carried out according to four metrics: structuredness, style compliance, relevance, and completeness. A five-point scale was used for the assessment, where 5 points – full compliance with the criterion, and 1 point – lack of compliance. Each answer was assessed by 3 experts independently; the final score was calculated as the arithmetic mean. The results are summarized in Table 5.

Our results confirmed that the combination of templatization and localized glossary gave the highest performance

in all metrics. Templated instructions ensured stability and predictability of the response structure, defined mandatory parameters (situation description, object, style, format), which significantly reduced the risk of deviation from the given logic. The localized glossary, in turn, minimized semantic discrepancies, increased the accuracy of terminology, and aligned actions with expected tactical models. Together, these elements created a synergistic effect: a clear structure of the prompt was combined with guaranteed correctness of terms, which allowed us to significantly improve both the quality and reproducibility of results.

Table 5

Comparing the success of model responses to formalized prompts

Methods of formalization	Structure	Style	Relevance	Completeness
Templating with key terms	4.2	4.1	4.3	4.0
Automatic grammar normalization	4.0	3.8	4.1	3.9
Linguistic restrictions (localized glossary)	4.5	4.3	4.4	4.2

An important part of our analysis was the comparison of formalized and non-formalized prompts. When working with unstructured instructions, experts, when evaluating AI responses, "intuitively" filled in the missing details, took into account the implicit context and logical connections that would be unavailable in an automated scenario. This led to an inflated subjective assessment but did not correspond to the real conditions of AI integration into the game process. Formalized prompts, on the contrary, limited the model to only the data that was explicitly provided in the prompt, and therefore demonstrated higher rates of accuracy, completeness, and relevance of responses.

As a result of the analysis, a sample of 100 AI responses was used: 50 for formalized prompts and 50 for non-formalized ones. The relevance of the responses was estimated as the ratio of the number of responses that the experts recognized as fully meeting the requirements to the total number of responses in the group. The reproducibility of the format was calculated as the proportion of responses that fully complied with the given template, and the error rate was calculated as the ratio of the number of responses with at least one violation of the format to the total sample size. Our results (Table 6) show that formalization made it possible to increase the relevance of the responses by an average of 30–35%, and the reproducibility of the format by almost two times. At the same time, it was possible to significantly reduce the number of errors, which is especially important when creating action cards in series, where even single failures can lead to an imbalance in the game system.

Table 6

Comparing the formalized and non-formalized prompts

Prompt type	Relevance of the answer	Format reproducibility	Error rate
Formalized	88–92%	85–90%	8–12%
Non-formalized	55–65%	40–50%	25–30%

A visual comparison of the formalized and non-formalized prompts in terms of response relevance, format reproducibility, and error rate is shown in Fig. 3 (based on Table 6).

Special attention was paid to the impact of prompt preparation automation. The use of auto-completion, contextual validation, and self-correction tools reduced the time for generating instructions for AI by 25–40%, reduced the number of syntactic and semantic errors by 30%, and helped maintain high reproducibility of responses in repeated generations. Compared to completely manual formulation, the use of such tools provides faster results without losing their quality.

Another aspect of our study was to check the stability of models when varying prompt formulations. Even with changes in the order of information representation or the use of synonyms of key terms, the results for templated and glossary-supported prompts remained within 85–92% of compliance with the given format. This confirms that a clear prompt construction methodology not only increases the accuracy of the response in a single case but also makes the work of AI stable in a series of repeated calls.

Thus, prompt formalization has shown that the most effective for generating AI action cards in wargames is an integrated formalization model that combines a template structure, a localized glossary, and mechanisms for automated prompt refinement. Such a model enables the achievement of high relevance and structured content, as well as significantly reducing the complexity of the process of preparing AI action cards in wargames.

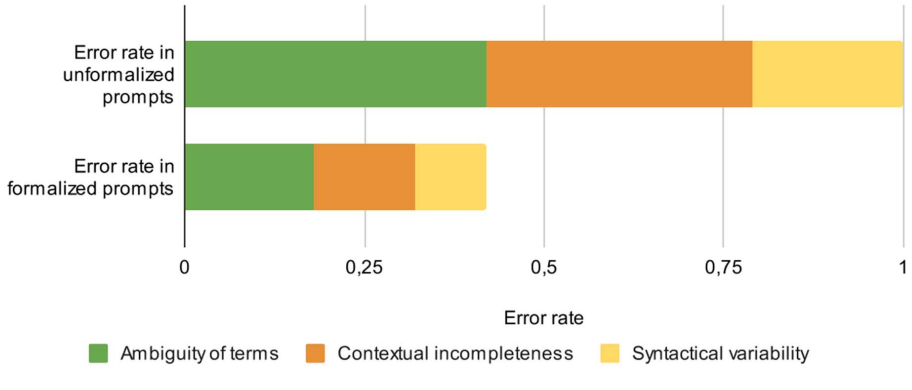

Fig. 3. Comparing the formalized and non-formalized prompts according to the specified indicators (based on Table 6)

5. 4. Results of development and testing of typical prompt structures

A set of typical prompt structures optimized for automated generation of AI opponent action cards in wargames was developed. The main task at this stage was to design such templates that would ensure the stability of the response format, compliance with the game logic, and ease of integration into game mechanics. As a result, three basic types were formed that structurally differ in the form of representation, instructional saturation, and level of analytical detail, which provides different approaches to modeling tactical decisions. The first type was called "Action Description" and is a concise, yet instructive fragment of text that conveys the essence of the operation, for example, moving the player in the direction of the nearest cover. The second format, defined as "d6-table", contains a structured list of six action options, logically tied to the results of a dice roll, which ensures random game development of the situation. The third type, "Explanation of the choice of action," is analytical in nature as it not only outlines a possible tactic but also provides a concise argument that deepens the understanding of the logic of the decision made.

During testing, it was found that the highest level of response controllability is achieved when using JSON formats and logical tables. JSON structures described the action parameters in detail (identifier, name, instructions, expected result), facilitating automatic processing and integration. However, they required strict syntax checking since even a minor error made it impossible to use the result. An example of such a structure is given in Table 7.

Table 7

Example of a JSON structure description fragment for a game action

Field	Description
id	Unique action identifier
action_name	Short name of tactical operation
instructions	Sequence of steps or logical conditions
expected_outcome	Result for current scenario

Despite the convenience of JSON, the results showed that the most stable and predictable for AI generation were "d6 tables". Their strength relates to simple but clearly regulated logic: each number from 1 to 6 corresponds to a separate action, which has its own tactical effect. Such a structure not only increases the predictability of AI responses but also minimizes cases of "mixing" of different options within one result. An example of a shortened d6 table is given in Table 8.

During testing that involved various language models, more than 240 structured output fragments were generated and tested (JSON structures – 60, d6 tables – 60, "Action description" – 60, "Explanation of action choice" – 60). Format stability was calculated as the proportion of fragments without any syntactic or logical errors out of the total volume of each type. Logic predictability was measured using expert assessment: the proportion of fragments that received at least 4 points for compliance with the expected result. The average integration time was normalized relative to manual prompt formulation (1.0×). It was found that "d6 tables" showed from 90 to 95% compliance with the specified format, while the proportion of logical deviations did not exceed 5–7%. JSON structures demonstrated about 85–90% format stability but required stricter syntax control since even a small error could complicate their use in automated scenarios. The "Explanation of action choice" format turned out to be the most variable: although it provided depth of context and allowed us to generate more "human" responses, the level of format stability fluctuated within 75–80%. To emphasize the difference in predictability and format stability, a generalized comparative Table 9 was compiled.

Table 8

Example of a fragment of a d6 table for an action in a tabletop wargame

d6	Action	Expected effect
1	Retreat to the nearest fortified position	Maintaining the viability of the unit
2	Use smoke screen to conceal maneuvers	Making it more difficult to direct enemy fire
3	Organize flank attack to probe enemy defensive positions	Identifying weak points in the defense
4	Carry out short-term intensive bombardment of enemy positions	Suppressing enemy firepower
5	Call in reinforcements from the rear	Increasing offensive potential
6	Carry out combined ground and artillery attack	Breaking through the enemy's defense line

Table 9

Comparing the stability and predictability of different types of structures

Structure type	Format stability	Predictability of logic	Average integration time
«Action description»	82–85%	80–85%	1,0×
«d6 table»	90–95%	90–94%	0,8×
«Action selection explanation»	75–80%	78–82%	1,2×
JSON structure	85–90%	88–90%	0,9×

The results confirmed that "d6 tables" are the optimal solution in the context of tabletop wargames due to the combination of simplicity, high predictability, and speed of integration into game mechanics. JSON structures take second place, providing excellent structuring, but requiring additional syntax checking. The "Action Description" and "Action Choice Explanation" formats are more suitable for scenarios where variability and a creative element are important, but they are inferior in stability and formal clarity. Thus, the development of typical prompt structures made it possible to build a standardized base that not only facilitates work with language models but also ensures stable quality and predictability of generated actions. Our results are consistent with the tasks defined in section 3 and confirm the achievement of the key research goals described in the conclusions. Further steps include expanding the base of typical structures, creating mixed formats (for example, combining "d6-tables" with JSON), and testing them on examples of more complex game scenarios.

After developing and testing the formalization methods, special attention was paid to practical verification of their effectiveness directly in the context of tabletop hexagonal wargames. The main object of testing was tabular structures based on "d6-tables", which at the previous stages showed the highest stability of the format (90–95%) and predictability of the logic of the AI opponent's actions. They were chosen as the basic format due to their ability to provide a clear connection between each action option and its tactical effect, which is critically important for solo-mode scenarios where a human player interacts with an automated opponent.

The testing was carried out by designing an experimental set of 60 action cards for an AI opponent in a wargame based on a historical scenario of the 18th century. To generate the content, pre-prepared prompts in the format of "d6 tables" and JSON structures were used, which were tested on several language models (GPT-4.1, Claude 4.0 Sonnet, Gemini 2.5 Pro, Mistral-7B, o3-mini). Each model generated action options based on the same conditions, which allowed for an objective comparison of the results. The evaluation was carried out by three independent game design experts according to the following criteria:

– compliance with game logic – how organically the action fits into the given scenario;

– structure and unambiguousness – clarity of the format and absence of contradictory instructions;

– relevance to tactical conditions – compliance of the action with the situation on the battlefield;

– level of interpretation errors – the number of cases when the model incorrectly read the structure or content of the prompt.

For ease of interpretation, the results were normalized on a five-point scale, where 5 points are the maximum correspondence, and 1 point is the absence of correspondence. The results are given in Table 10.

Table 10

Comparing the quality of cards generated by the formalized and non-formalized prompts

Prompt format	Compliance with logic	Structuredness	Relevance	Interpretation errors (↓)	Average score
Formal. (d6)	4.7	4.8	4.6	0.3	**4.7**
Formal. (JSON)	4.5	4.6	4.5	0.4	**4.6**
Non-formal.	3.6	3.4	3.5	1.1	**3.5**

In order to clearly represent the results of evaluating the qualitative characteristics of cards, Fig. 4 shows a comparison of prompt formats according to three main criteria.

The results show that the use of formalized prompts on average increases the quality of generated cards by 31–34% compared to non-formalized prompts. At the same time, the greatest increase was observed in the indicators of structuring (+41%) and a decrease in the number of interpretation errors (three times). It is important that even in cases where AI offered non-standard or creative options, the formalized prompt kept them within the given format, which made such answers suitable for direct use without additional editing.

It was separately noted that strict adherence to formatting and limiting linguistic variations in prompts minimizes ambiguities and significantly reduces the need for post-generational refinements. For example, when using non-formalized prompts, up to 40% of cards required correction, while in the case of "d6-tables" this indicator did not exceed 8%.

Thus, our validation has confirmed the practical applicability and high efficiency of the proposed method. Using standardized prompt structures makes it possible to design balanced and relevant game action cards for an AI opponent, which opens up the prospect of scaling the method to other board game genres and further automating the design process. The next step will be to integrate formalized prompts into the scenario builder interface, which will allow wargame designers to generate complete card sets directly in the digital environment.

6. Discussion of results based on investigating the formalization of text prompts to AI systems in the context of generating AI player actions in wargames

As shown in Table 2, the implementation of a clear structural organization and a localized glossary allowed us to reduce the number of interpretation errors by 58% (from 100% to 42%) compared to non-formalized instructions. The reduction in the number of ambiguous responses from large language models was made possible by eliminating ambiguity, contextual incompleteness, and grammatical variability inherent in non-formalized prompts. Due to the preliminary definition of the command style, game situation, and expected response format, the model received a more complete instruction, which reduced the need for "guessing" on the part of AI.

The systematization of linguistic metrics for assessing the quality of responses, given in Table 5, highlights that a localized glossary and structured templates contribute significantly to improving response performance. More consistent and accurate AI responses are provided by explicit context descriptions, a localized glossary of terms, and a structured prompt format. This minimizes ambiguity, increases semantic consistency, and simplifies the model's interpretation of complex game scenarios.

The use of standardized formats (tables, JSON, algorithm) enabled an increase in format reproducibility to 85–92% (Table 6), confirming the close relationship between strict formalization and the accuracy of the obtained results.

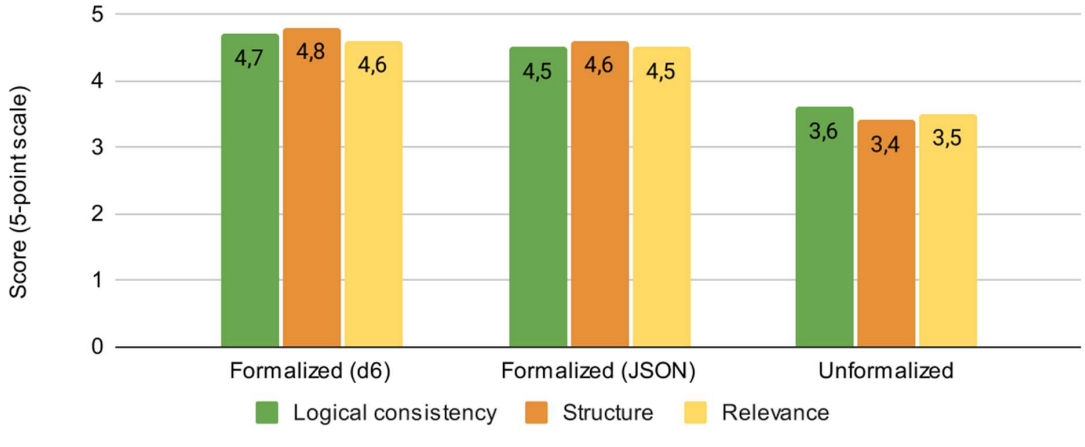

Fig. 4. Assessing the quality of AI responses for different prompt formats on a 5-point scale

The main factor of efficiency was templated structures, especially "d6-tables", which demonstrated the highest stability (90–95%, Table 9). This is explained by their logical simplicity: a fixed number of options and a single-valued correlation "number – action" minimizes the risk of generation errors. In turn, JSON structures provided high compliance with logic and structuredness (Table 10) but required stricter syntax checking, which indicates a compromise between formal accuracy and practical convenience at the stage of integration into the software environment.

The resulting indicators correspond to the trends identified in a number of works on prompt engineering, where structured methods (chain-of-thought, format-based prompting) increased the quality of answers by 30–40% [5–8]. Our results are consistent with those data but are specific in that they were first tested in the applied field of tabletop wargames where the contextual component is critically important. The achieved quality improvement (31–34%) surpasses the results of the reviewed studies (14%) [14]. The discrepancy is explained by the fact that the proposed solution achieves such indicators precisely in the applied context – the generation of game scenarios. Unlike general tasks of mathematical thinking or text generation, in the context of this study, contextual accuracy is critically important. This contrasts with previous works that focused mainly on the educational or programming context. The difference is the combination of templates, glossary, and contextual parameterization, which allowed us to achieve an average quality score of 4.6–4.7 versus 3.5 for non-formalized prompts (Table 10), while most known solutions are based only on templatization or instructiveness without taking into account industry specificity.

The developed formalized approach solves the problem of integrating formal structuring with subject-specific semantics. Unlike general prompt engineering methods that focus either on formal correctness (like SLOT) or on general improvement of "thinking" (like chain-of-thought), our methodology simultaneously provides:

– subject specificity; thanks to the localized glossary, semantic consistency specific to wargames is created;

– practical applicability; unlike abstract structures, d6-tables are integrated into game mechanics much easier;

– stable reproducibility (85–90%, Table 6), which is ensured by a combination of prompt templating and a glossary of terms.

This makes it possible to close the gap between highly technical solutions (often inaccessible to game designers) and intuitive methods (which give unstable results).

Our results have a number of limitations. First, the experiments were conducted under controlled conditions with fixed scenario parameters, which reduces the possibility of generalizing the results to more dynamic game situations. Second, the number of expert evaluators was limited to three individuals, which could affect the level of objectivity. Third, the testing covered a limited number of wargame scenarios while in real game practices the variability is much higher.

Among the shortcomings of the study, worth noting is the instability of the results in cases of complex or multi-component prompts, as well as the dependence on specific language models used in the experiments. In addition, the qualitative assessment of the results required manual participation of experts, which reduces the level of reproducibility.

To verify the adequacy of the formalized models, additional testing is planned in the format of group game sessions. One of the players will control the "AI opponent" using generated action cards, and these actions will be compared with the expected logic of the players. It is planned to record how much the AI behavior will correspond to the declared styles (aggressive, cautious, balanced), and whether it will cause a violation of the game balance or misunderstanding from real players.

In addition, it is planned to test a set of internal tools based on its own prompt generator, which is planned to be implemented in the form of a web interface. The tool will allow one to generate prompts to the AI under a semi-automatic mode based on input parameters (unit type, field situation, game style). In addition, one of the important functions will be the ability to specify the desired response format (table, JSON, plain text, etc.). The generator will use prompt templates and a glossary that will be compiled based on empirical data collected during this study.

The combination of theoretical analysis, automated generation, and empirical testing in a simulated game environment will provide an opportunity to obtain representative results that are subsequently used for interpretation and comparison with non-formalized methods.

Our results confirmed that the formalization of text prompts to AI significantly improves the quality of AI opponent action generation in wargames; however, in addition, aspects were identified that require further development.

First, there is a need to expand and detail prompt templates for various game situations and wargame genres. The existing templates have shown high efficiency in a rather narrow range of scenarios, but their suitability for large-scale multifactor simulations is still limited. Further experiments should involve the development of universal and at the same time flexible structures that can adapt to changes in the task without losing the clarity of the result.

Second, to increase the predictability of responses, it is advisable to expand the localized glossary by including terms from military affairs, the geopolitical context, and the psychology of the opponent's behavior. This will make it possible to reduce the dependence of the response quality on cultural and linguistic interpretations and model modifications.

Third, a promising area is the implementation of mixed formats of structured prompts, in particular combinations of d6 tables, JSON structures, and explanatory blocks. It is expected that this would make it possible to combine the advantages of compact formatting of tables and the flexibility of other formats, while increasing the scalability of the methodology. The development of a systematic approach paves the way for its scaling to other subject areas, and its implementation in visual prompt builders (the "prompt-as-interface" concept), which can simplify interaction with AI for non-technical users. In addition, pre-training models on structured responses seems promising, which can potentially eliminate some of the format errors.

The expected development difficulties relate to the methodological and technical planes. Formalization of complex multifactorial scenarios will require both combinations of templates and mechanisms for automatic adaptation to changes in the scenario in real time. From the experimental side, the greatest challenge will be stress testing under dynamic conditions where changing parameters affects the relevance of actions over several iterations. Since it is assumed that in further work those scenarios will be generated from a relatively short user description, an additional technical difficulty is the possibility that the model will not contain sufficiently reliable historical or factual information. This

could lead to the emergence of semantic inaccuracies in the content, especially in historically oriented wargames, where even small deviations from real events affect the gaming authenticity and educational value of the resulting scenarios and destroy the gaming experience of users.

Thus, the success of our research is attributed to a combination of structured templates, a glossary, and formal rules for prompt formation. It was this integration that enabled a reduction in errors by almost half and an improvement in content quality by a third. The proposed systematic approach stands out against the background of known solutions by its combination of templating, glossaries, and format control, which makes it relevant for complex gaming environments, in particular for tabletop wargames. At the same time, maintaining the effectiveness of such a solution in an expanded and interdisciplinary context requires further research and addressing the outlined methodological challenges. This creates a basis for further development toward automating the generation of game content with a high degree of structure using AI systems.

7. Conclusions

1. Our study has identified key linguistic difficulties that arise when users interact with artificial intelligence systems. The most common errors are due to the ambiguity of terms (42% of cases), the lack of full context (37%), and syntactic variability of formulations (21%). The introduction of structured prompts in combination with a localized glossary allowed us to reduce the number of interpretation errors by more than half (58%). This result is attributed to the fact that a clear definition of terms and their unambiguous interpretation significantly limits the scope for variability of the language model.

2. The key principles for constructing AI instructions were identified and experimentally confirmed. The highest efficiency was shown by templatization with key terms (4.2 points), use of a localized glossary (4.5 points), and automatic grammar normalization (4.0 points). The combination of these principles ensured an increase in the relevance of answers to 88–92% compared to 55–65% for non-formalized prompts. The efficiency is explained by the combination of structuring and semantic accuracy, which narrows the variability of language models, which directly affects the result.

3. The results of comparing the formalized and non-formalized text prompts to AI language models showed that formalization of prompts ensured the stability of results in more than 85% of cases of compliance with the requirements. The average quality scores were as follows: relevance – 4.4; completeness – 4.2; style compliance – 4.3 on a five-point scale. The result is attributed to fact that the methods simultaneously eliminate three main sources of errors: ambiguity, structural uncertainty, and contextual incompleteness.

4. The use of three types of templates has been proposed: "action description", "d6-table" and "action choice explanation" for generating action cards in hexagonal wargames. "d6-tables" demonstrated the highest stability – 90–95% format compliance. When testing 60 cards in wargames, formalized prompts showed an improvement of 31–34%, specifically, 4.7 points (d6) and 4.6 points (JSON) against 3.5 points on a five-point scale for non-formalized prompts. The number of interpretation errors decreased threefold (from 1.1 to 0.3–0.4). This is explained by the logical simplicity and rigid structure of the format.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable prompt.

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in Chapters 4–7.

References

1. Poola, I. (2023). Overcoming ChatGPTs inaccuracies with Pre-Trained AI Prompt Engineering Sequencing Process. International Journal of Technology and Emerging Sciences, 3 (3), 16–19. Available at: https://www.researchgate.net/profile/Indrasen-Poola/publication/374153552_Overcoming_ChatGPTs_inaccuracies_with_Pre-Trained_AI_Prompt_Engineering_Sequencing_Process/links/65109c34c05e6d1b1c2d6ae9/Overcoming-ChatGPTs-inaccuracies-with-Pre-Trained-AI-Prompt-Engineering-Sequencing-Process.pdf
2. Prompt engineering. OpenAI. Available at: https://platform.openai.com/docs/guides/text#prompt-engineering
3. Diab, M., Herrera, J., Chernow, B., Chernow, B., Mao, C. (2022). Stable Diffusion Prompt Book. OpenArt. Available at: https://cdn.openart.ai/assets/Stable%20Diffusion%20Prompt%20Book%20From%20OpenArt%2011-13.pdf
4. Shah, C. (2025). From Prompt Engineering to Prompt Science with Humans in the Loop. Communications of the ACM. https://doi.org/10.1145/3709599
5. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv. https://doi.org/10.48550/arXiv.2205.11916
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. et al. (2017). Attention Is All You Need. arXiv. https://doi.org/10.48550/arXiv.1706.03762

7. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 55 (9), 1–35. https://doi.org/10.1145/3560815
8. Wei, J., Zhou, D. (2022). Language Models Perform Reasoning via Chain of Thought. Google Research. Available at: https://research.google/blog/language-models-perform-reasoning-via-chain-of-thought/
9. Zaghir, J., Naguib, M., Bjelogrlic, M., Névéol, A., Tannier, X., Lovis, C. (2024). Prompt Engineering Paradigms for Medical Applications: Scoping Review. Journal of Medical Internet Research, 26, e60501. https://doi.org/10.2196/60501
10. Pawar, V., Gawande, M., Kollu, A., Bile, A. S. (2024). Exploring the Potential of Prompt Engineering: A Comprehensive Analysis of Interacting with Large Language Models. 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA), 1–9. https://doi.org/10.1109/iccubea61740.2024.10775016
11. Prompt engineering: overview and guide. Google Cloud. Available at: https://cloud.google.com/discover/what-is-prompt-engineering
12. Sui, Y., Zhou, M., Zhou, M., Han, S., Zhang, D. (2024). Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 645–654. https://doi.org/10.1145/3616855.3635752
13. Crabtree, M. (2024). What is Prompt Engineering? A Detailed Guide For 2025. DataCamp. Available at: https://www.datacamp.com/blog/what-is-prompt-engineering-the-future-of-ai-communication
14. Kryazhych, O., Vasenko, O., Isak, L., Babak, O., Grytsyshyn, V. (2024). Method of constructing requests to chat-bots on the base of artificial intelligence. International Scientific Technical Journal «Problems of Control and Informatics», 69 (2), 84–96. https://doi.org/10.34229/1028-0979-2024-2-7
15. Wang, D. Y.-B., Shen, Z., Mishra, S. S., Xu, Z., Teng, Y., Ding, H. (2025). SLOT: Structuring the Output of Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2505.04016
16. Shorten, C., Pierse, C., Smith, T. B., Cardenas, E., Sharma, A., Trengrove, J., van Luijt, B. (2024). StructuredRAG: JSON Response Formatting with Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2408.11061
17. Tang, X., Zong, Y., Phang, J., Zhao, Y., Zhou, W., Cohan, A., Gerstein, M. (2024). Struc-Bench: Are Large Language Models Good at Generating Complex Structured Tabular Data? Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 12–34. https://doi.org/10.18653/v1/2024.naacl-short.2
18. Sasaki, Y., Washizaki, H., Li, J., Yoshioka, N., Ubayashi, N., Fukazawa, Y. (2025). Landscape and Taxonomy of Prompt Engineering Patterns in Software Engineering. IT Professional, 27 (1), 41–49. https://doi.org/10.1109/mitp.2024.3525458
19. Cotroneo, P., Hutson, J. (2023). Generative AI tools in art education: Exploring prompt engineering and iterative processes for enhanced creativity. Metaverse, 4 (1), 14. https://doi.org/10.54517/m.v4i1.2164
20. Hau, K., Hassan, S., Zhou, S. (2025). LLMs in Mobile Apps: Practices, Challenges, and Opportunities. 2025 IEEE/ACM 12th International Conference on Mobile Software Engineering and Systems (MOBILESoft), 3–14. https://doi.org/10.1109/mobilesoft66462.2025.00008
21. Sharma, R. K., Gupta, V., Grossman, D. (2024). SPML: A DSL for Defending Language Models Against Prompt Attacks. arXiv. https://arxiv.org/abs/2402.11755
22. Mountantonakis, M., Tzitzikas, Y. (2025). Generating SPARQL Queries over CIDOC-CRM Using a Two-Stage Ontology Path Patterns Method in LLM Prompts. Journal on Computing and Cultural Heritage, 18 (1), 1–20. https://doi.org/10.1145/3708326
23. Wang, Z., Chakravarthy, A., Munechika, D., Chau, D. H. (2024). Wordflow: Social Prompt Engineering for Large Language Models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 42–50. https://doi.org/10.18653/v1/2024.acl-demos.5
24. Desmond, M., Brachman, M. (2024). Exploring Prompt Engineering Practices in the Enterprise. arXiv. https://doi.org/10.48550/arXiv.2403.08950
25. Garcia, M. C., Bondoc, B. C. (2024). Mastering the art of technical writing in it: Making complex things easy to understand in Atate campus. World Journal of Advanced Research and Reviews, 22 (1), 571–579. https://doi.org/10.30574/wjarr.2024.22.1.1020
26. Developing AI Literacy With People Who Have Low Or No Digital Skills (2024). Good Things Foundation. Available at: https://www.goodthingsfoundation.org/policy-and-research/research-and-evidence/research-2024/ai-literacy
27. Garg, A., Rajendran, R. (2024). The Impact of Structured Prompt-Driven Generative AI on Learning Data Analysis in Engineering Students. Proceedings of the 16th International Conference on Computer Supported Education, 270–277. https://doi.org/10.5220/0012693000003693