Framing Prompts as User Stories: Effects on the Output Quality of Generative AI

Tomislav Car
University of Rijeka, Faculty of Tourism and Hospitality Management, Croatia
Ivan Šimac
University of Applied Sciences of Rijeka, Croatia
Sabrina Šuman
University of Applied Sciences of Rijeka, Croatia

Abstract

This study investigates how framing prompts as user stories affects the output quality of generative AI systems. It examines whether structuring prompts in the form commonly used in software development and human-centred design ("As a user, I would like to in order...") improves the relevance, clarity, and contextual accuracy of generated responses. A controlled experiment was conducted with tourism-related scenarios in which a large language model (LLM) was presented with prompts in both traditional and user story formats. The outputs were evaluated using a hybrid method that combines automatic metrics (BERTScore) and human expert evaluation based on the IE-Information Extraction framework (precision, recall, F1, and error rate) to capture qualitative aspects of response quality. The results show that prompts formatted as user stories consistently yield higher-quality responses, especially in matching the user's intent and providing accurate, relevant content. These findings highlight the role of prompt structure in shaping LLM performance and suggest that user-centred prompt design can improve generative AI applications in domain-specific contexts. The study contributes to the prompt engineering literature and offers practical implications for improving human–AI interaction by formulating inputs more deliberately and structurally.

Keywords: prompt engineering, user story, generative AI, output quality, tourism scenarios

JEL classification: L86, O31, O32, O33, Z32

Paper type: Research article
Received: 15 June 2025
Accepted: 10 August 2025

DOI: 10.54820/entrenova-2025-0057

Citation: Car, T., Šimac, I., & Šuman, S. (2025). Framing Prompts as User Stories: Effects on the Output Quality of Generative AI. ENTRENOVA - ENTerprise REsearch InNOVAtion, 11(1), https://doi.org/10.54820/entrenova-2025-0057.

Introduction

Artificial intelligence (AI) is a field of computer science that focuses on the development of technologies that enable machines and computer systems to simulate human cognitive functions such as learning, understanding, problem solving, decision making and creativity, and to act as autonomous agents capable of perceiving their environment and make decisions aimed at achieving the most favourable outcome, even in situations involving uncertainty (Russell & Norvig, 2022; Paloalto networks, 2025; Stryker, 2024).

In recent years, AI has become mainstream, a foundational technology for all industries, enabling everything from voice assistants and chatbots to complex data analytics platforms. At its core, AI enables machines to simulate human cognitive abilities — to learn from data, understand language, and solve problems. AI enables machines to solve ambiguity problems like humans do by taking into account broader context and reasoning from experience and past knowledge (Šuman, 2021). This transformative capability has driven the market for AI tools to an estimated value of $244bn by 2025, with forecasts pointing to over $800bn by 2030.

The shift in 2025 is the rise of a new generation of generative models capable of deep research, advanced reasoning, and iterative planning, a sub-sector of generative AI that is expected to grow exponentially between 2025 and 2030 (Thormundsson, 2025).

One of the most fascinating topics in AI is Natural Language Processing (NLP), which enables machines to communicate like humans and understand natural language.

Since 2022, there has been tremendous progress in NLP, driven by the development of large language models (LLMs), massive training datasets, deep learning technologies, and transformer architectures. The power of machine understanding and natural language generation is increasing exponentially, especially at the levels that have long posed the most significant challenges in NLP – the semantic level, the discourse level, and the pragmatic level (Liddy, 2001; Feldman, 1999; Khurana et al., 2023):

- The semantic level focuses on understanding the meaning of words and sentences, including the relationships between concepts and the disambiguation of meanings.
- The discourse level is concerned with how sentences relate to each other in a coherent text, taking into account aspects such as the resolution of references and thematic continuity across multiple sentences.
- The pragmatic level interprets language based on the context and the speaker's intention. Factors such as implicit meanings, social cues, and user goals are taken into account to derive the intended meaning beyond the literal content.

This progress in NLP is also evident in the performance of modern virtual assistants powered by LLMs, such as ChatGPT, Gemini, Copilot, Perplexity, Claude, DeepSeek, Grok, and others. Continuous monitoring of performance is necessary because these tools can deliver attractive, plausible-sounding results that are, at the same time, false or nonsensical (Rigin et al., 2025). These tools require input in the form of a so-called prompt, in which you must describe and define what is to be executed. To achieve an output that best suits the user's needs, the prompt must be designed to produce a high-quality, more accurate, customised, and comprehensive result.

In this sense, this article aims to analyse the impact of the prompt structure on the output performance of the generative AI tools selected for this study (ChatGPT Plus Model 4o, Gemini 2.5 Pro, Copilot Pro, and Perplexity Pro). Reference sentences were

created for the study to represent the expected output of the AI tools. For each reference sentence, both the basic prompt format and the user story format were created.

Performance was measured using a combination of the traditional performance evaluation approach, based on parameters from the Information Extraction (IE) domain, and BERTScore with the Roberta large language model. The following research questions were defined:

- RQ1- Is there a difference in the performance of the generated output content when comparing a basic-standard structured prompt with a prompt formulated as a user story as used in agile information systems development?
- RQ2- Is there a difference in the performance of the generated output content between different generative AI tools?

In addition to these research questions, the paper explores the complexity of evaluating the quality of generative AI outputs and selecting the appropriate method—or combination of methods—to objectively assess this quality. The Methodology section describes the processes of the two evaluation techniques used, as well as all the necessary steps performed before the actual evaluation. In the Background and Related Work section, prompt engineering is explained, and the user story model from the field of agile information systems development is presented. In the Results and Discussion section, all results are presented in tables, followed by a discussion.

Background and related work

Prompt engineering has become a crucial technique for improving the reliability, performance, and generalisability of large language models (LLMs) for a wide range of NLP tasks. It refers to the practice of designing, structuring, and refining prompts to guide LLMs to the desired results. Initially explored as an experimental interaction method, prompt engineering has since evolved into a systematic approach for matching model behaviour to task requirements and has proven particularly useful in domains such as medical decision-making, recommendation systems, and programming support (Wang et al., 2024; Kusano et al., 2024).

Prompt engineering involves manipulating various elements of the prompt, including the instruction (or directive), examples (in zero-shot, one-shot, or few-shot configurations), context (task-relevant data), persona (e.g., "act as a professional lawyer"), and output formatting constraints (Schulhoff et al., 2025). Researchers have proposed and evaluated a wide range of prompting methods, including standard instruction-based prompts, Chain-of-Thought (CoT), Self-Consistency, and Tree-of-Thought (ToT) prompting. Each of these approaches involves trade-offs among interpretability, effectiveness, and computational cost. The emergence of these prompting paradigms has driven the development of taxonomies and systematic surveys — such as that of Liu et al. (2023) — that catalogue many unique prompting strategies and emphasise the urgent need for consistent terminology and evaluation protocols. He et al. (2024) also explored how prompt formatting (from plain text to JSON or a bullet list) can lead to differences in LLM output and performance. Wang et al. (2024) investigated how prompting strategies affect the results. They showed that the same model can produce significantly different results depending on how the input question is formulated and which prompting strategy is used. Kusano et al. (2024) investigated prompt selection for recommendation systems and concluded that prompt performance is highly dependent on the task, with no single format consistently outperforming others across datasets.

Despite the extensive focus on prompt design, comparatively few papers have addressed systematic frameworks for prompt evaluation.

The Prompt Report Schulhoff et al., 2025, argues for the standardisation of prompt evaluations through benchmark tasks, performance metrics, and reproducibility studies across models. This is particularly important given the rapid evolution of LLMs, where new architectures and training paradigms may respond differently to previously effective prompts. This underpins the argument that prompt design affects not only raw accuracy but also the interpretability and trustworthiness of model outputs.

This type of variability emphasises another key problem: the lack of universal or optimal prompts. This emphasises the argument that the development of prompts is not just a superficial formatting issue but a key methodological problem when using LLMs in high-stakes contexts.

Researchers have shown that prompts are not just inputs, but crucial interfaces through which LLM capabilities are accessed and controlled. The existing literature suggests that prompt development is not a trivial task, but a complex, empirical process that requires systematic experimentation. Although considerable progress has been made in documenting prompting strategies and evaluating their effects, further research is needed to standardise evaluation protocols, develop automatic prompting selection methods, and understand prompting behaviour in multilingual and multimodal contexts.

In this study, the authors propose a strategy for structuring prompts as user stories from agile software development.

User stories, which are widely used in agile software development, serve as a semi-structured method for formulating functional requirements in natural language. They are usually structured along the lines of "As a [user type], I want [goal/task] so that [reason or benefit]" and facilitate communication between developers and stakeholders by ensuring that users' intentions are clearly formulated and contextualised (Cohn, 2004). This format strikes a balance between expressiveness and constraint by providing a controlled syntax while retaining the benefits of natural language. Formulating NLP tasks in user story format is beneficial as it aligns tasks with user-centred requirements and improves understanding of users' needs (Raharjana et al., 2021). We could consider a user story format as a type of Controlled Natural Language (CNL) - a subset of natural language designed to reduce ambiguity and complexity, thereby increasing precision and enabling automatic interpretation by machines or non-expert humans. Kuhn 2014) defines CNLs as "engineered subsets of natural languages that are restricted in vocabulary and grammar in order to reduce ambiguity and complexity". Initially developed for technical documentation, knowledge representation, and formal specification, CNLs have recently found application in software development, particularly in requirements engineering, where clarity is paramount.

The convergence between CNLs and user stories is particularly evident in Natural Language Processing (NLP) applications. User stories can be seen as a lightweight, domain-specific controlled natural language. Mainly because of their clarity and easy-to-remember format, the user story format has also been observed in the prompt engineering field, as in this research.

Methodology

The main goal of this research is to show how prompts formatted as user stories affect the quality of generative AI output. To this end, several steps were taken:

1) Selection of generative AI tools.

For reasons of research rigour, we only used paid Pro versions: Gemini 2.5 Pro, ChatGPT Plus-Model 4o, Microsoft Copilot Pro, Perplexity Pro.

2) Creation of a set of valid sentences for evaluation.
   In this phase, we examined many possible sentences that satisfied the conditions for a unique reference sentence to apply the information extraction technique and calculate Precision, Recall, F1, and Error Rate. Each reference set (REF1,…REF4) was created by three experts who consulted official sources such as websites, tourism agency offers, and the knowledge of other experts. For each sentence, reference elements that were of interest to us were identified. In the end, four sentences about tourism and leisure in Primorsko-Goranska County in summer 2025 were formed for the references.
3) Creation of the prompt versions that should produce an output corresponding to the reference sentence: four basic prompts and four user story prompts for each of the four AI tools. The outputs of the AI tools are called candidates. In total, we obtained 32 candidates (8 for each AI tool, 4 for basic prompts, and 4 for user story prompts). Basic prompts were labelled BP1,...BP4, while user story-formatted prompts were labelled USP1,…USP4.
   The main semantics of the question (prompt) were kept in both formats to show how the AI tool reacts to the prompt format.
4) To allow a better and more objective comparison between the reference set and the output of the AI tools, we have created an additional prompt (AP1,…AP4) that follows the CNL topic mentioned above and specifies the structure of the output to make it more transparent and more comparable.
5) For each AI tool, the BERTScore is calculated for each candidate sentence (4 introductory and four user story prompts) using the pre-trained model of Roberta Large Language, namely P-Precision, R-Recall, and F1-measure. In addition, four measures from the IE domain (Precision, Recall, F1, and Error rate, which are explained in more detail in this section) were calculated by the experts. All results are presented in Tables 1-5.
6) Step 6: Automatic calculation of the BERTScore metrics P, R, and F1 with Python and the corresponding libraries, as well as the IE metrics (D, C, M, N, S, I, P, R, F1, ERR), which were explained in the Methodology section and calculated by the experts – Table 3. Although experts carried out the IE calculations, the formulae are rigorous (there are reference sentences and clear parameters), so it is a valid quantitative method.
7) The Discussion of the results obtained can be found in the Results and Discussion sections. The focus here is on the performance of each AI tool (across all sentences), for each sentence by each tool, and on the differences in results when formulating a simple prompt versus a user story prompt.

Combining two different metrics for

Evaluating the performance of AI systems — especially those that process or generate natural language — requires reliable, interpretable, and task-appropriate metrics. Both metrics used in this study are quantitative and based on specific parameters calculated automatically by an algorithm (BERTScore using Python and its libraries) and manually by experts, and they follow strict formulas (well-known in the field of information extraction) and reference entities for each sentence. Therefore, this is a valid research method to answer research questions.

In the field of information extraction (IE), Makhoul et al. (1999) proposed a basic framework for evaluating system performance based on three key metrics: Precision, Recall, and the F-measure. These metrics quantify the system's ability to accurately

and completely extract relevant information from the text. Precision is the ratio of correctly identified elements to all elements extracted by the system. At the same time, recall is the ratio of correctly identified elements to all correct elements that should have been extracted. The F-measure (or F1-score) combines these two values into a single harmonic mean to balance the trade-offs between completeness and correctness.

Although this scoring paradigm was initially developed for structured information extraction tasks — such as named entity recognition or slot filling in pre-annotated corpora - it is also highly relevant for modern generative AI models. These tools can generate complex outputs, such as summaries, explanations, reasoning chains, and even structured specifications or privacy requirements.

For tasks where factual correctness and completeness are crucial, traditional metrics remain a foundation. For example, if a model has the task of extracting the "who"," what", and "why" components from the user stories, its outputs can be evaluated directly using the IE proposed metrics. However, generative models introduce qualitatively new challenges beyond traditional IE tasks. Their outputs are often open, creative, or context-dependent, which complicates the direct application of rigid evaluation metrics. For such outputs, researchers have proposed broader assessment frameworks that incorporate the following:

- Human judgement (e.g., ratings on a Likert scale of relevance, coherence, or usefulness)
- Automated metrics (e.g., BLEU, ROUGE, BERTScore)
- Task-specific benchmarks (e.g., exact accuracy in argumentation or logical consistency in the generation of arguments).

Despite these innovations, most assessment criteria are still based on the core principles of precision (correctness) and recall (completeness). For example, when summarising or answering questions, ROUGE scores are essentially recall-oriented and measure the overlap of n-grams with reference texts, whereas BLEU has a precision bias.

Consequently, the F-measure remains a critical metric even in modern contexts, particularly when generative systems are used in information-sensitive domains such as law, healthcare, or safety-critical software. In such domains, hallucinated content (low precision) or omitted facts (low recall) can pose significant risks.
Furthermore, understanding the trade-offs between these metrics enables more targeted model refinement and better risk management. For example, a generative model tuned for high recall may be used for exploratory search or brainstorming tasks, while a model tuned for high precision may be better suited for formal document creation or compliance reporting.

In this research, we combined traditional metrics from the IE domain with the BERTSscore, which captures the context and deeper semantics of the observed text. The IE metrics and then BERTScore are explained below.

The following labels, taken from Makhoul et al., 1999, are used to calculate IE metrics such as Precision, Recall, F1 measure, and Error Rate:

- N_i = total number of identified elements in the i-th reference sentence
- M_i = Total number of elements identified in the i-th candidate sentence (i.e., how many elements the AI tool identified in total in the output);
- C_i = number of correct elements in the i-th candidate sentence – the identified elements in the candidate sentence that match both the content and the role of the reference elements;

- S_i = number of substitutions in the i-th candidate sentence – elements in the candidate sentence that match the reference elements in terms of content but not in terms of role. This parameter is not used in the AI tool's evaluation, as this type of error is recorded under insertions (I). Si will therefore be 0 for all records;
- D_i = Number of deletions – elements of the i-th reference sentence that are missing in the candidate sentence (no match with a candidate element);
- I_i = number of insertions – elements identified in the candidate sentence that do not match any reference element (false positives).

The following values are calculated for each sentence:

P_i, R_i, $F1_i$, and ERR_i.

For a given sentence i, all elements are considered at once:

$$N_i=C_i+S_i+D_i \quad (1)$$
$$M_i = C_i + S_i + I_i \quad (2)$$

Based on the above, the precision measure for an i-th candidate sentence is defined as follows:

$$P_i = \frac{C_i}{M_i} = \frac{C_i}{C_i + S_i + I_i} \quad (3)$$

Moreover, the recall for the i-th candidate sentence is:

$$R_i = \frac{C_i}{N_i} = \frac{C_i}{C_i + S_i + D_i} \quad (4)$$

Precision refers to the percentage of correct candidate elements, while recall refers to the percentage of correct candidate elements in relation to the total number of referential elements.

The F1 score, which is used as a single measure of evaluation performance, is defined as the weighted harmonic mean of precision and recall:

$$F1_i = \frac{2 * P_i * R_i}{P_i + R_i} \quad (5)$$

A metric that focuses on all three types of errors is the error rate, which is defined as follows for a given candidate sentence:

$$ERRi = \frac{S_i+D_i+I_i}{C_i+S_i+D_i+I_i} \quad (6)$$

BERTScore (*Bidirectional Encoder Representations from Transformers*) is an advanced evaluation metric developed to measure the semantic similarity between texts. It overcomes the limitations of traditional lexical overlap metrics, such as BLEU and ROUGE, which do not capture deeper semantics and context. By using contextual embeddings derived from pre-trained transformer models such as BERT, it captures subtle semantic relationships rather than relying solely on exact lexical matches. The metric first embeds reference and candidate texts into high-dimensional

vector spaces and then computes token-level similarities using pairwise cosine similarity, allowing evaluation of semantic closeness independent of differences in word choice. BERTScore has three key components: Precision, which evaluates how accurately candidate tokens reflect the reference; Recall, which measures how thoroughly candidate tokens cover the reference; and an F1 score, which provides a balanced view of Precision and Recall. Its benefits include better correlation with human judgment and the flexibility to adapt seamlessly to different NLP tasks, such as text summarisation, machine translation, and text production quality assessment. Furthermore, BERTScore's approach to semantic evaluation benefits greatly from BERT's deep bidirectional context encoding, a critical aspect highlighted by Devlin et al. (2019, the original authors of BERT. Despite higher computational costs than simpler metrics, BERTScore positions itself as a valuable tool for robust NLP evaluation due to its superior performance at capturing semantic details (Bansal, 2025; Van Otten, 2024; Devlin et al., 2019).

It should be noted that the BERTScore metrics Precision, Recall, and F1 are calculated in a different way than the traditional IE metrics we mentioned in (3), (4), and (5). BERTScore measures the semantic similarity between the predicted/candidate and reference texts, whereas IE aims to measure the exact match between the predicted/candidate and reference elements. Precision in BERTScore measures how well each predicted element from the candidate text semantically matches the reference elements. In contrast, the traditional Precision metric in IE measures the percentage of predicted elements in the candidate text that are correct. In addition, Recall in BERTScore measures how well each reference element is captured by the predicted element in the candidate text, while Recall in the IE metric measures the percentage of reference elements that were successfully predicted in the candidate text.

Results and discussion

Following the seven steps described in the Methodology section, the results of steps 2, 3, and 4 are shown in Table 1:

Table 1

Defining basic prompts, user story prompts, additional prompts, and reference text for each sentence

	Text 1
BP_1	Please give me a list of three to five open-air cinemas operating in July 2025 in the Primorsko-Goranska county in Croatia (excluding islands), each sentence specifying the cinema's name and venue.
USP_1	As a big film fan, I would like a list of three to five open-air cinemas operating in July 2025 in Primorsko-Goranska County in Croatia (excluding islands) so that I can watch films in the open air on holiday in July.
AP_1	Please formulate output in the form of a sentence: Here is the list of open-air cinemas in Primorsko-Goranska county in July of 2025: Name_1 -Venue/City, Name_2-Venue/City...where Name_1 is the name of the first open-air Cinema/festival.
REF_1	In July 2025, there are the following open-air cinemas on the mainland of Primorsko-Goranska County: the Open Air Theatre in Opatija, the Summer Art-kino in Rijeka, the Cinehill festival in Delnice, and Fužine.
	Text 2
BP_2	Please give me a list of all hotels in Rijeka, Croatia, for the summer of 2025.

USP$_2$	As a tourist who stays exclusively in objects categorized as hotels, give me all open hotels in summer 2025 in Rijeka, Croatia, so that I can choose the best hotels that fit my needs.
AP$_2$	Please formulate output in the form of the sentence: Here is the list of hotels in Rijeka operating in July and August of 2025: HotelName_1, HotelName2...where HotelName1 is the name of the first Hotel.
REF$_2$	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka, Costabella Beach Resort & Spa, Grand Hotel Bonavia, Hotel Neboder, Hotel Jadran, Hotel Continental, Aparthotel Villa V, Old Town Inn, Tre Re Inn.
	Text 3
BP$_3$	List me all five-star hotels in Opatija, Croatia.
USP$_3$	As a traveller looking for first-class accommodation (exclusively 5-star hotels) in Opatija (Croatia), I would like to see a complete list of five-star hotels in the city so that I can choose the luxury accommodation that best suits my plans.
AP$_3$	Please formulate the output in the form of a sentence: Here is the list of all five-star hotels in Opatija, Croatia: HotelName_1-number of stars, HotelName2...where HotelName1 is the name of the first hotel.
REF$_3$	Here is the list of all five-star hotels in Opatija, Croatia: Amadria Park Hotel Milenij-5 stars, Hotel Ambasador-5 stars, Hotel Bevanda-5 stars, Keight Hotel Opatija Curio Collection by Hilton-5 stars, Amadria Park Hotel Sveti Jakov – 5 stars, and Boutique & Design Hotel Navis – 5 stars.
	Text 4
BP$_4$	List all the tourist attractions/sights that are located in the city of Opatija (Croatia).
USP$_4$	As a tourist or user of a travel application, I would like to receive a list of all tourist attractions and sights located exclusively in the city of Opatija (Croatia) so that I can plan my visit and sightseeing without being distracted by attractions in the surrounding areas.
AP$_4$	Please formulate the output in the form of a sentence: Here is the list of all attractions or sights in Opatija, Croatia: 1. Name 1, 2. Name2â€¦.... where Name 1 is the name or title of the first attraction or sight of the entire list. Please integrate a numbered list into one paragraph.
REF$_4$	Here is the list of all attractions or sights in the city of Opatija, Croatia: (1) St. James's Church, (2) Church of the Annunciation of the Blessed Virgin Mary; (3) Villa Angiolina & Croatian Museum of Tourism; (4) Swiss House; (5) Juraj Šporer Art Pavilion; (6) Hotel Kvarner; (7) Hotel Imperial; (8) Angiolina Park; (9) St. James' Park, (10) Lungomare Promenade; (11) Maiden with the Seagull Statue; (12) Croatian Walk of Fame; (13) Opatija Small Harbour ("Portić"); (14) Carmen Sylva Forest Promenade; (15) American Gardens; (16) Open Air Theatre.

Note: BP – basic prompt; AP – additional prompt; USP – user story prompt; REF - reference
Source: Authors' work

Now that all reference sentences, basic prompts, and user story prompts, as well as additional prompt text, are known, the results of the AI tools for each BP and each USP are given in Table 2 (methodology step 5).

Table 2
Results of four AI tools for each BP and USP

	ChatGPT
BP1	Here is the list of open-air cinemas in Primorsko-Goranska county in July of 2025: CINEHILL 2025 – Fužine, Ljetno Art-kino – Rijeka, Tobogan Festival – Rijeka, Open-Air Theatre – Opatija, Cinema under the stars -Hotel Admiral.
BP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka, Costabella Beach Resort & Spa, Grand Hotel Bonavia, Hotel Continental, Hotel Neboder, Teatro Suite & Rooms, Old Town Inn, Hotel Jadran, and Hotel Marina.

BP3	Here is the list of all five-star hotels in Opatija, Croatia: Ikador Luxury Boutique Hotel & Spa – 5 stars, Design Hotel Navis – 5 stars, Hotel Bevanda – 5 stars, Amadria Park Hotel Milenij – 5 stars, Hotel Ambasador – Liburnia – 5 stars, Amadria Park Hotel Sveti Jakov – 5 stars, Keight Hotel Opatija (Curio Collection by Hilton) – 5 stars, Hotel Imperial – Liburnia – 5 stars, Hotel Kristal – Liburnia – 5 stars.
BP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Lungomare Coastal Promenade, 2. Villa Angiolina, 3. Park Angiolina, 4. Maiden with the Seagull Statue, 5. Saint Jacob's Church, 6. Saint Jacob's Park, 7. Church of the Annunciation, 8. Croatian Walk of Fame, 9. Port of Opatija, 10. Open Air Summer Theater, 11. Šetalište Carmen Sylva, 12. Juraj Šporer Art Pavilion, 13. Swiss House, 14. Hotel Kvarner, 15. Casino Rosalia, 16. Slatina Beach, 17. Lido Beach, 18. Tomaševac Beach, 19. Volosko Neighborhood, 20. Habsburg-era Villas, 21. Park Margarita, 22. Opatija Circuit.
USP1	Here is the list of open-air cinemas in Primorsko-Goranska County in July 2025: Cinehill Film Festival – Fužine, Summer Art-kino – Rijeka, Open Air Theatre – Opatija.
USP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka Costabella Beach Resort & Spa, Grand Hotel Bonavia, Hotel Continental, Hotel Neboder, Hotel Jadran, Aparthotel Villa, Teatro Suite & Rooms, Old Town Inn, Tre re Inn
USP3	Here is the list of all five-star hotels in Opatija (Croatia): Boutique & Design Hotel Navis – 5 stars, Hotel Ambasador – 5 stars, Keight Hotel Opatija, Curio Collection by Hilton – 5 stars, Amadria Park Hotel Milenij – 5 stars, Hotel Bevanda – 5 stars, Amadria Park Hotel Sveti Jakov – 5 stars.
USP4	Here is the list of all attractions and landmarks in Opatija: (1) Lungomare Coastal Promenade; (2) Villa Angiolina and Angiolina Park; (3) Church of St. James and St. James's Park; (4) Church of the Annunciation; (5) Maiden with the Seagull statue; (6) Juraj Šporer Art Pavilion; (7) Croatian Walk of Fame; (8) Portić (Opatija's small harbour); (9) Hotel Kvarner; (10) Hotel Imperial; (11) Swiss House and American Gardens; (12) Carmen Sylva Forest Path; (13) Slatina Beach and Lido Beach; (14) Open-Air Summer Theater.
	Gemini 2.5 Pro
BP1	Here is the list of open-air cinemas in Primorsko-Goranska county in July of 2025: Ljetna pozornica - Opatija, Summer Art-kino - Rijeka, Kino Sloboda - Lovran, and Open-Air Cinema - Baška. *
BP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Grand Hotel Bonavia - Rijeka, Hotel Continental, Molo Longo - Integrated hotel (apartments & rooms), Hilton Rijeka Costabella Beach Resort & Spa, Hotel Neboder.
BP3	Here is the list of all five-star hotels in Opatija, Croatia: Keight Hotel Opatija, Curio Collection by Hilton-5 star, Hotel Milenij-5 star, Amadria Park Hotel Sveti Jakov-5 star, Hotel Ambasador-5 star, Hilton Rijeka Costabella Beach Resort & Spa-5 star.
BP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Lungomare (Franz Joseph I Promenade), 2. Maiden with the Seagull, 3. Villa Angiolina, 4. Angiolina Park, 5. St. James's Church, 6. St. James's Park, 7. Juraj Šporer Art Pavilion, 8. Hotel Kvarner, 9. Croatian Walk of Fame, 10. Slatina Beach, 11. Tomaševac Beach, 12. American Gardens, 13. Church of the Annunciation of Mary, 14. Open Air Summer Theater, 15. Opatija's Small Harbor (Portić), 16. Volosko, 17. Učka Nature Park.
USP1	Here is the list of open-air cinemas in Primorsko-Goranska county in July of 2025: Ljetno Art-kino - Rijeka, Mala Ljetna pozornica - Opatija, Ljetno kino Kraljevica - Kraljevica. *
USP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Grand Hotel Bonavia - Rijeka, Hotel Continental, Botel Marina Rijeka, Hotel Neboder, Hotel Jadran.
USP3	Here is the list of all five-star hotels in Opatija, Croatia: Keight Hotel Opatija, Curio Collection by Hilton - 5-star hotel, Hotel Milenij - 5-star hotel, Amadria Park Hotel Sveti Jakov - 5-star hotel, Hotel Ambasador - 5-star hotel, Hotel Bevanda - 5-star hotel.
USP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Lungo Mare (Franz Joseph I Promenade), 2. Maiden with the Seagull (Djevojka s galebom), 3. Helios

	and Selena Fountains, 4. Croatian Walk of Fame, 5. Villa Angiolina, 6. Croatian Museum of Tourism, 7. Hotel Kvarner, 8. Hotel Imperial, 9. Juraj Šporer Art Pavilion, 10. Museum of Chocolate (located in Hotel Continental), 11. Church of St. James (Crkva Sv. Jakova), 12. Church of the Annunciation of the Blessed Virgin Mary (Crkva Navještenja Blažene Djevice Marije), 13. Madonna del Mare Statue, 14. Angiolina Park, 15. St. James's Park (Park Svetog Jakova), 16. American Gardens, 17. Margarita Park, 18. Slatina Beach, 19. Lido Beach, 20. Opatija Harbor/Port. *
	Copilot Pro
BP1	Here is the list of open-air cinemas in Primorsko-Goranska county in July of 2025: CINEHILL 2025, Fužine, Summer on Gradina – Trsat Castle, Rijeka, Ljetna pozornica Opatija – Open Air Theatre, Opatija. *
BP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka Costabella Beach Resort & Spa, Grand Hotel Bonavia, Hotel Jadran, Teatro Suite & Rooms, Hotel Continental, Old Town Inn, Botel Marina, Hostel Dharma, Hostel Morcic-Ri, Youth Hostel Rijeka, Hostel 1W, Apartments Del Molo, Apartments Marija Sucic, Apartments tRIo, Sir John's Rooms, Engine Rooms, Casa San Vito.
BP3	Here is the list of all five-star hotels in Opatija, Croatia: Ikador Luxury Boutique Hotel & Spa – 5 stars, Amadria Park Hotel Sveti Jakov – 5 stars, Boutique & Design Hotel Navis – 5 stars, Amadria Park Hotel Milenij – 5 stars, Hotel Ambasador – Liburnia – 5 stars, Keight Hotel Opatija, Curio Collection by Hilton – 5 stars.
BP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Lungomare Coastal Promenade, 2. Maiden with the Seagull, 3. Croatian Walk of Fame, 4. Opatija's Small Harbour (Portić), 5. Villa Angiolina & Angiolina Park, 6. Villa Amalia, 7. Juraj Šporer Art Pavilion, 8. Open Air Theatre, 9. Gervais Centre, 10. Church of St. James (Crkva Sv. Jakova), 11. Church of Our Lady of the Annunciation (Crkva Navještenja), 12. St. James's Park, 13. American Gardens, 14. Carmen Sylva Forest Promenade, 15. Učka Nature Park, 16. Slatina Beach, 17. Lido Beach, 18. Opatija Market Hall (Mrkat), 19. Croatian Museum of Tourism, 20. Galerija Antikvarijat, 21. Volosko, 22. Ičići & Ika, and 23. Veprinac. *
USP1	Here is the list of open-air cinemas in Primorsko-Goranska County in July of 2025: Summer Cinema at Trsat Castle Rijeka, Art-kino Summer Rijeka, Summer Cinema Gorski Kotar – Delnice or Fužine, Kastav Cultural Summer – Fortica/Kastav.
USP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka Costabella Beach Resort & Spa, Teatro Suite & Rooms, Grand Hotel Bonavia, Hotel Continental, Hotel Jadran.
USP3	Here is the list of all five-star hotels in Opatija, Croatia: Ikador Luxury Boutique Hotel & Spa – 5 stars, Amadria Park Hotel Sveti Jakov – 5 stars, Boutique & Design Hotel Navis – 5 stars, Hotel Milenij – 5 stars, Hotel Ambasador - Liburnia – 5 stars, Keight Hotel Opatija, Curio Collection by Hilton – 5 stars, Bevanda Hotel & Restaurant – 5 stars, Hotel Kvarner - Liburnia – 5 stars.
USP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Lungomare Promenade, 2. Slatina Beach, 3. Opatija's Small Harbour (Portić), 4. Villa Angiolina & Angiolina Park, 5. Croatian Museum of Tourism, 6. Maiden with the Seagull Statue, 7. St. James's Church (Crkva Sv. Jakova), 8. Church of Our Lady of the Annunciation, 9. St. James's Park, 10. American Gardens, 11. Carmen Sylva Forest Promenade, 12. Juraj Šporer Art Pavilion, 13. Open Air Theatre, 14. Gervais Centre, 15. Croatian Walk of Fame, 16. Hotel Kvarner Terrace, 17. Opatija Market Hall (Mrkat).*
	Perplexitiy Pro
BP1	Here is the list of open-air cinemas in Primorsko-Goranska county in July of 2025: Summer Stage Opatija – Angiolina Park/Opatija, Art-kino Croatia Summer Cinema – Trg Riječke rezolucije/Rijeka, Ljetno kino Delnice – King Tomislav Square/Delnice, Open Air Fužine – Fužine Town Center/Fužine. *
BP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka Costabella Beach Resort & Spa, Hotel Jadran, Botel Marina Rijeka, Hotel Neboder, Hotel Continental, Tre Re Inn, HI Hostel Rijeka, Riva Rooms.

BP3	Here is the list of all five-star hotels in Opatija, Croatia:Hotel Ambasador – 5 stars, Amadria Park Hotel Milenij – 5 stars, Keight Hotel Opatija, Curio Collection by Hilton – 5 stars, Design Hotel & Restaurant Navis – 5 stars, Amadria Park Hotel Sveti Jakov – 5 stars, Hotel Bevanda – 5 stars, Ikador Luxury Boutique Hotel & Spa – 5 stars.
BP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Lungomare Coastal Promenade, 2. Maiden with the Seagull, 3. Villa Angiolina & Angiolina Park, 4. St. James' Church & Park, 5. Croatian Walk of Fame, 6. Park Margarita, 7. Juraj Šporer Art Pavilion, 8. Hotel Kvarner, 9. Swiss House, 10. Fountain of Helios and Selena, 11. Portić Harbor, 12. Wall of Fame (Graffiti Mural), 13. Slatina Beach, 14. Lipovica Beach, 15. Tomasevac Beach, 16. Volosko, 17. Church of the Annunciation, 18. Opatijska balinjerada, 19. Miroslav Krleža Monument, 20. Opatija Adventure I POINT I.
USP1	Here is the list of open-aire cinemas in Primorsko-Goranska county in July of 2025: Ljetno Art-kino – Capuchin Monastery Courtyard/Rijeka, Open Air Cinema – Balustrada Beach/Crikvenica, Open Air Theatre – Angiolina Park/Opatija. *
USP2	Here is the list of hotels in Rijeka operating in July and August of 2025: Hilton Rijeka Costabella Beach Resort & Spa, Botel Marina Rijeka, Tre Re Inn, Hotel Jadran, Hotel Continental, Hotel Neboder.
USP3	Here is the list of all five-star hotels in Opatija, Croatia: Keight Hotel Opatija, Curio Collection by Hilton – 5 stars, Amadria Park Hotel Sveti Jakov – 5 stars, Amadria Park Hotel Milenij – 5 stars, Hotel Ambasador – 5 stars, Hotel Bevanda – 5 stars, Boutique Design Hotel, Restoran Navis – 5 stars.
USP4	Here is the list of all attractions or sights in Opatija, Croatia: 1. Maiden with the Seagull, 2. Lungomare - Promenada Franje Josipa I., 3. Villa Angiolina u Opatiji Museum, 4. Park Margarita, 5. Hrvatska ulica slavnih (Croatian Walk of Fame), 6. Church of St. James, 7. Slatina Fountain, 8. Opatija's Small Harbour, 9. Opatijska balinjerada, 10. Švicarska kuća, 11. Miroslav Krleža Monument, 12. Opatija Adventure I POINT I.*

Note: * The results were all translated into English first (e.g., Ljetno → Summer) or the Croatian version was removed if there were English and Croatian terms together (e.g., Church (Crkva)).
Source: Authors' work

Step 6: automatic calculation of the BERTScore metrics P, R, and F1 with Python and the corresponding libraries, as well as the IE metrics (D, C, M, N, S, I, P, R, F1, ERR) explained in the methodology section, which the experts calculated – Table 3.

Table 3
Both metric results from each AI tool for each BP and USP

		BERTScore			IE measures								
	Results	P	R	F1	D	I	C	M	P	R	F1	ERR	N
ChatGPT	BP1	0,88	0,90	0,89	1	2	2	5	0,40	0,67	0,50	0,60	3
	BP2	0,91	0,92	0,92	2	2	6	8	0,75	0,75	0,75	0,40	8
	BP3	0,93	0,96	0,94	0	3	6	9	0,67	1,00	0,80	0,33	6
	BP4	0,90	0,90	0,90	1	7	15	22	0,68	0,94	0,79	0,35	16
	USP1	0,92	0,95	0,93	0	0	3	3	1,00	1,00	1,00	0,00	3
	USP2	0,92	0,96	0,94	0	1	8	9	0,89	1,00	0,94	0,11	8
	USP3	0,96	0,96	0,96	0	0	6	6	1,00	1,00	1,00	0,00	6
	USP4	0,94	0,94	0,94	0	1	16	16	1,00	1,00	1,00	0,06	16
Gemini	BP1	0,93	0,93	0,93	1	2	2	4	0,50	0,67	0,57	0,60	3
	BP2	0,93	0,92	0,92	4	1	4	5	0,80	0,50	0,62	0,56	8
	BP3	0,94	0,93	0,94	2	1	4	5	0,80	0,67	0,73	0,43	6
	BP4	0,92	0,92	0,92	4	5	12	17	0,71	0,75	0,73	0,43	16
	USP1	0,92	0,92	0,92	1	2	2	3	0,67	0,67	0,67	0,60	3
	USP2	0,95	0,91	0,93	4	1	4	5	0,80	0,50	0,62	0,56	8
	USP3	0,95	0,93	0,94	1	0	5	5	1,00	0,83	0,91	0,17	6

		BERTScore			IE measures							
Copilot	USP4	0,91	0,91	0,91	1	5	15	20	0,75	0,94	0,83	0,29	16
	BP1	0,93	0,93	0,93	1	1	2	3	0,67	0,67	0,67	0,50	3
	BP2	0,88	0,93	0,91	3	12	5	17	0,29	0,63	0,40	0,75	8
	BP3	0,94	0,95	0,95	1	1	5	6	0,83	0,83	0,83	0,29	6
	BP4	0,90	0,91	0,90	3	8	14	22	0,64	0,88	0,74	0,44	16
	USP1	0,90	0,93	0,91	1	2	2	4	0,50	0,67	0,57	0,60	3
	USP2	0,96	0,93	0,94	4	1	4	5	0,80	0,50	0,62	0,56	8
	USP3	0,93	0,95	0,94	0	2	6	8	0,75	1,00	0,86	0,25	6
	USP4	0,93	0,92	0,92	2	3	13	16	0,81	0,81	0,81	0,28	16
Perplexity	BP1	0,91	0,94	0,92	1	2	2	4	0,50	0,67	0,57	0,60	3
	BP2	0,95	0,94	0,95	3	3	5	8	0,63	0,63	0,63	0,55	8
	BP3	0,95	0,96	0,96	0	1	6	7	0,86	1,00	0,92	0,14	6
	BP4	0,89	0,91	0,90		8	14	22	0,64	0,88	0,74	0,36	16
	USP1	0,92	0,93	0,92	1	1	2	3	0,67	0,67	0,67	0,50	3
	USP2	0,97	0,94	0,96	3	1	5	6	0,83	0,63	0,71	0,44	8
	USP3	0,97	0,96	0,96	0	0	6	6	1,00	1,00	1,00	0,00	6
	USP4	0,90	0,88	0,89	8	5	8	13	0,62	0,50	0,55	0,62	16

Source: Authors' work

In the following Table 4, the best and worst values for each parameter are shown - green is the best and red is the worst.

Table 4
The results of BERTScore and IE metric - highlighted best and worst values by the AI tools (green and red).

		BERTScore			IE			
	Results	P	R	F1	P	R	F1	ERR
chatGPT	BP1	0,883	0,901	0,892	0,400	0,667	0,500	0,600
	BP2	0,914	0,925	0,919	0,750	0,750	0,750	0,400
	BP3	0,933	0,956	0,944	0,667	1,000	0,800	0,333
	BP4	0,900	0,900	0,900	0,682	0,938	0,789	0,348
	USP1	0,915	0,951	0,933	1,000	1,000	1,000	0,000
	USP2	0,915	0,957	0,935	0,889	1,000	0,941	0,111
	USP3	0,962	0,963	0,962	1,000	1,000	1,000	0,000
	USP4	0,945	0,940	0,942	1,000	1,000	1,000	0,059
Gemini	BP1	**0,932**	0,930	0,931	0,500	0,667	0,571	0,600
	BP2	0,930	0,916	0,923	0,800	0,500	0,615	0,556
	BP3	0,941	0,933	0,937	0,800	0,667	0,727	0,429
	BP4	0,920	0,917	0,918	0,706	0,750	0,727	0,429
	USP1	0,919	0,919	0,919	0,667	0,667	0,667	0,600
	USP2	0,953	0,905	0,928	0,800	0,500	0,615	0,556
	USP3	0,950	0,929	0,939	1,000	0,833	0,909	0,167
	USP4	0,910	0,909	0,909	0,750	0,938	0,833	0,286
Copilot	BP1	0,930	0,926	0,928	0,667	0,667	0,667	0,500
	BP2	0,881	0,935	0,907	0,294	0,625	0,400	0,750
	BP3	0,942	0,950	0,946	0,833	0,833	0,833	0,286
	BP4	0,900	0,910	0,905	0,636	0,875	0,737	0,440
	USP1	0,899	0,927	0,913	0,500	0,667	0,571	0,600
	USP2	0,962	0,925	0,943	0,800	0,500	0,615	0,556
	USP3	0,927	0,947	0,937	0,750	1,000	0,857	0,250

		BERTScore			IE			
Perplexity	USP4	0,926	0,923	0,924	0,813	0,813	0,813	0,278
	BP1	0,908	0,937	0,922	0,500	0,667	0,571	0,600
	BP2	0,954	0,940	0,947	0,625	0,625	0,625	0,545
	BP3	0,952	0,960	0,956	0,857	1,000	0,923	0,143
	BP4	0,889	0,908	0,898	0,636	0,875	0,737	0,364
	USP1	0,915	0,930	0,923	0,667	0,667	0,667	0,500
	USP2	0,974	0,941	0,957	0,833	0,625	0,714	0,444
	USP3	0,966	0,959	0,962	1,000	1,000	1,000	0,000
	USP4	0,899	0,878	0,888	0,615	0,500	0,552	0,619

Source: Authors' work

Step 7: Discussion on the results obtained to answer the research questions – the performance of AI tools for BPs vs. USPs - Tables 3 and 4.

In Table 5, there are overall mean values for each metric overall for BPs and USPs, the mean values for BPs and USPs separately (the bold fonts represent the better results – the lower the better, while all other metrics are the higher the better).

Table 5
Overall means, BP means and USP means by AI tools

		BERTScore			IE			
		P	R	F1	P	R	F1	ERR
chatGPT	Mean	0,921	0,936	0,929	0,798	0,919	0,848	0,231
	Mean BP	0,907	0,920	0,914	0,625	0,839	0,710	0,420
	Mean USP	**0,934**	**0,953**	**0,943**	**0,972**	**1,000**	**0,985**	**0,042**
gemini	Mean	0,932	0,920	0,926	0,753	0,690	0,708	0,453
	Mean BP	0,931	**0,924**	**0,927**	0,701	0,646	0,660	0,503
	Mean USP	**0,933**	0,915	0,924	**0,804**	**0,734**	**0,756**	**0,402**
copilot	Mean	0,921	0,930	0,925	0,662	0,747	0,687	0,457
	Mean BP	0,913	0,930	0,921	0,608	**0,750**	0,659	0,494
	Mean USP	**0,929**	**0,931**	**0,929**	**0,716**	0,745	**0,714**	**0,421**
perplexity	Mean	0,932	0,932	0,932	0,717	0,745	0,724	0,402
	Mean BP	0,926	0,936	0,931	0,655	**0,792**	0,714	0,413
	Mean USP	**0,938**	**0,927**	**0,933**	**0,779**	0,698	**0,733**	**0,391**

Source: Authors' work

Looking at the results, **research question 1** can be answered by noting that almost all parameters perform better in USPs across *all AI tools*.

The answer to **research question 2** is that some AI tools responded better to the USPs, in particular ChatGPT, which performed best on almost all parameters related to the USPs and also in general. Perplexity had the best results on BPs, while Copilot performed the worst on USPs and Gemini on BPs.

Conclusion

The topic of this study is the rapid development of generative AI tools, taking a closer look at the quality of their outputs and the impact of prompt formatting on their performance. Prompt engineering shows that prompts are not just inputs but also

interfaces through which humans access and control LLM capabilities. The relevant literature emphasises that prompt development is a complex, empirical process that requires systematic experimentation, and that, despite the progress made in documenting prompting strategies and their effects, further research is needed.

This study aimed to investigate how the prompt format affects the output quality of generative AI tools. The idea of structuring prompts as user stories originates in agile software development and relies on its simple structure to improve communication between stakeholders, provide clarity, minimise ambiguity, and maximise specificity.

As described in the methodology, seven steps or phases were carried out to objectively evaluate the quality of 4 generative AI tools: ChatGPT 4o, Gemini Pro 2.5, Copilot Pro, and Perplexity Pro.

Two research questions were posed:

- RQ1- Is there a difference in the performance of the generated output content when comparing a basic-standard structured prompt with a prompt formulated as a user story as used in agile information systems development?
- RQ2- Is there a difference in the performance of the generated output content between different generative AI tools?

To answer the research questions, the authors proposed a method combining two quantitative assessment metrics: BERTScore, which captures deeper semantics, and traditional IE metrics, which remain a reasonable basis for evaluating factual accuracy and completeness.

Answer to RQ1: All observed AI tools showed better results in almost all observed metric parameters in the case of user story prompts (Tables 3-5). The most significant difference between basic and user story prompts was observed with the ChatGPT 4o model.

Answer to RQ2: There are differences among the tools, as shown in Table 4. ChatGPT performed best overall, including best on USPs, while Perplexity performed best on BPs. Copilot scored worst on USPs and in general, while Gemini scored worst on BPs.

As AI tools continue to advance, there is a need to develop strategies to maximise their power and utility continually. This research demonstrates the complexity of objectively assessing AI performance and selecting appropriate evaluation metrics. The limitations of this research include the small number of observed sentences and the inclusion of only four AI tools, although others exist. Future research plans include experimenting with various AI tools and prompting strategies to evaluate outputs.

References

1. Bansal, R. (2025, April). BERTScore: A contextual metric for LLM evaluation. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2025/04/bertscore-a-contextual-metric-for-llm-evaluation/
2. Cohn, M. (2004). *User stories applied: For agile software development*. Addison Wesley Longman Publishing.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
4. Feldman, S. (1999). NLP meets the Jabberwocky: Natural language processing in information retrieval. *Information Today, 16*(7), 38–42.
5. He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does prompt formatting have any impact on LLM performance? *arXiv*. http://arxiv.org/abs/2411.10541

6. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications, 82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4
7. Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics, 40*(1), 121–170. https://doi.org/10.1162/COLI_a_00172
8. Kusano, G., Akimoto, K., & Takeoka, K. (2024). Are longer prompts always better? Prompt selection in large language models for recommendation systems. *arXiv*. http://arxiv.org/abs/2412.14454
9. Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of library and information science* (2nd ed.). Marcel Dekker.
10. 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), Article 195. https://doi.org/10.1145/3560815
11. Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. (1999). Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop* (pp. 249–252). Morgan Kaufmann.
12. Palo Alto Networks. (2025). *What is AI*. https://www.paloaltonetworks.com/cyberpedia/artificial-intelligence-ai
13. Raharjana, I. K., Siahaan, D., & Fatichah, C. (2021). User stories and natural language processing: A systematic literature review. *IEEE Access, 9*, 53811–53826. https://doi.org/10.1109/ACCESS.2021.3070606
14. Rigin, S., Kottekkadan, N. N., Toney, T., & KK, M. N. (2025). Generative AI tools (ChatGPT) in tourism research: An experimental conversation. *Tourism and Hospitality Management, 31*(2), 251–263. https://doi.org/10.20867/thm.31.2.13
15. Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education.
16. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). The prompt report: A systematic survey of prompt engineering techniques. *arXiv*. http://arxiv.org/abs/2406.06608
17. Stryker, C. (2024, August 9). What is AI. *IBM Think Blog*. https://www.ibm.com/think/topics/artificial-intelligence
18. Šuman, S. (2021). Pregled metoda obrade prirodnih jezika i strojnog prevođenja. *Zbornik Veleučilišta u Rijeci, 9*(1), 371–384. https://doi.org/10.31784/zvr.9.1.23
19. Thormundsson, B. (2025, July 25). Artificial intelligence (AI) worldwide – Statistics & facts. *Statista*. https://www.statista.com/topics/3104/artificial-intelligence-ai-worldwide/
20. Van Otten, N. (2024, August 20). BERTScore – A powerful NLP evaluation metric explained & how-to tutorial in Python. *Spotintelligence*. https://spotintelligence.com/2024/08/20/bertscore/
21. Wang, L., Chen, X., Deng, X. W., Wen, H., You, M. K., Liu, W. Z., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digital Medicine, 7*(1), 147. https://doi.org/10.1038/s41746-024-01029-4

About the authors

Tomislav Car, PhD, is an associate professor at the Faculty of Tourism and Hospitality Management at the University of Rijeka, where he received his doctorate in 2017. His research interests include mobile technologies, mobile applications, social media, the Internet of Things, and artificial intelligence in tourism, as well as e-business and information systems. He has participated in several national and international research projects. The author can be contacted at: **tcar@fthm.hr**

Ivan Šimac, Master in Computer Science, is a lecturer at the Department of Information and Communication Technologies at the University of Applied Sciences of Rijeka. He is currently completing a doctoral programme at the Faculty of Computer Science and Digital Technologies at the University of Rijeka. His main areas of interest are the application of artificial intelligence, especially computer vision, programming, and the development of information systems. The author can be contacted at: **isimac@veleri.hr**

Sabrina Šuman, PhD in computer science, is a college professor at the Department of Information and Communication Technologies at the University of Applied Sciences of Rijeka. Her interests lie in artificial intelligence, business analytics, decision support, and programming. Author of four books in the research areas of programming, decision support, and business analysis. Published numerous articles in international journals and conferences. Involved in various professional and scientific projects. The author can be contacted at: **ssuman@veleri.hr**