Appendix A

Work schedule throughout semesters

Winter semester

Semester week number	Description
1	Studying the problem
2	Finding literature
3	Finding more sources
4	EU AI Act analysis
5	Jailbreak analysis
6	AI analysis
7	Content filter analysis
8	Risks of AI solutions analysis
9	Experimenting with Jailbreaking
10	More experimenting with Jailbreaking
11	Evaluating of the experiments
12	Document revision and final changes

Plan evaluation

The first part of my Bachelor's thesis (BP1) was done in the winter semester and in the following examination period. In the first weeks I was finding sources for my thesis, studying them, and finally started writing. Unfortunately, I set my expectations a little higher. The winter semester was tough and I needed to meet deadlines for other courses, so I was writing less than I intended to. Sometimes I did my analyzes of the mentioned problems in another order because I was more interested in them. My supervisor and I agreed to have consultations every week. I tried to be present every week, but sometimes I did not have time due to the busy semester. In each consultation, I have discussed progress with my supervisor, and he gave me valuable feedback on changes or additions that I could implement. Finally, I finished the first part of the thesis during the examination period, where I had time because my examinations did not come until January. Some of the mentioned topics in the plan are set as TBD, and other topics which were not in this semester plan as well where I will be working on them in the summer semester as part of BP2 and the final thesis.

Summer semester

Semester week number	**Description**
1	Revision and proofreading of the document
2	US Legislation Analysis
3	China Legislation Analysis
4	Solution proposal
5	Experimenting with Jailbreaking
6	Experimenting with Jailbreaking
7	Experimenting with Jailbreaking
8	Evaluating of the experiments
9	Guidelines for users
10	Guidelines for users
11	Conclusion and Resumé
12	Document revision and final changes

Plan Evaluation

The second part of my Bachelor's thesis (BP2) was completed during the summer semester. I started by reviewing and proofreading the document that I had created in the winter. I was able to fix some unclear wording, formatting, etc. Then I continued where I left off, which was the legislation. Then I wrote a solution proposal for the practical part of the thesis which were the guidelines. After that, I came up with several experiments that needed more time for execution. Following the experiments, I evaluated them and started working on the guidelines. Lastly, I wrote the conclusion for my thesis and translated the most important parts to Slovak for the Resumé as a requirement of writing the thesis in English. Finally, I reviewed the entire document, made final changes, and submitted it. I was able to follow the plan with some slowdowns that were caused by my other assignments, and I was about a week behind the schedule that I needed to catch up.

Appendix B

Guidelines for users

Contents

B.1 Overview

These guidelines were created to establish ethical handling of artificial intelligence, primarily for non-expert users of the general public and developers. The guidelines explain the basics of prompt engineering for novice users, and later focus on recommendations for the ethical and fair use of AI systems. The guidelines also cover examples for identifying AI-generated content and the beneficial uses of AI.

These guidelines are grounded in the EU Ethics Guidelines for Trustworthy AI [1], which were created by the independent High-Level Expert Group (HLEG) on AI appointed by the European Commission.

B.2 Introduction to prompt engineering

This section focuses mostly on non-experts or novice users in this field. As explained in the thesis, prompt engineering involves designing and optimizing text instructions called prompts, which are mainly used to communicate with chatbots that use LLMs in their background. To generate a desired output the prompts need to have certain quality and clarity, and they need to be specific as the LLMs are prone to generating vague, irrelevant, incorrect and unnecessarily long responses that often do not include the desired answers.

Many people think that prompt engineering is closely related to computer science and programming. However, a lack of expertise in these areas should not discourage people from using LLMs, as such knowledge is not essential.

Now, let us explain two prompt techniques that could help achieve better results from LLM prompting.

[1] AI HLEG (2019), Ethics Guidelines for Trustworthy AI, URL: `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`

Few-shot prompting

Few-shot prompting provides AI models with some task examples to improve accuracy [2]. It can be used as a technique to enable in-context learning, where we provide demonstrations in the prompt to steer the model to better performance. The demonstrations serve as a conditioning for the subsequent examples in which we would like the model to generate a response [3].

Example of a few-shot prompting:

```
Input:
    Task:    "What is the rating of follwing movie from 1 (worst)
             to 10 (best) ?"
    Example 1:
    Review: "Absolutely amazing! A must-watch." Rating: 10/10
    Example 2:
    Review: "It was okay, not great but not terrible." Rating: 5/10
    Example 3:
    Review: "Terrible plot and weak acting." Rating: 3/10
    Now you try:
    Review: "The visuals were stunning but the story dragged."
    Rating:
Output:
    7/10
```

We can observe that the model evaluated the rating based on the examples presented. This method is sufficient for simpler tasks; however, it is not suitable for complex reasoning tasks. On such tasks, we can use another method called

[2] https://www.ibm.com/think/topics/few-shot-prompting

[3] https://www.promptingguide.ai/techniques/fewshot

Chain-of-thought prompting.

Chain-of-thought prompting

Chain-of-thought (CoT) prompting enables complex reasoning capabilities through intermediate reasoning steps. It can be combined with few-shot prompting to get better results on more complex tasks that require reasoning before responding. [4] Example of CoT prompting:

```
Input:
    Task:        "Solve the following age-related word problems.
                 Think step-by-step before answering."
    Example 1:
    Problem:     "Alice is 12 years old. Her brother is 3 years
                 older. How old is her brother?"
    Reasoning:   Alice is 12. Her brother is 3 years older,
                 so 12 + 3 = 15 years old.
    Answer:      15
    Example 2:
    Problem:     "Tom is 18 years old. His sister is half his
                 age. How old is his sister?"
    Reasoning:   Tom is 18. Half of 18 is 9, so his sister
                 is 9 years old.
    Answer:      9
```

[4] https://www.promptingguide.ai/techniques/cot

```
    Example 3:
    Problem:    "A mother is 30 years older than her son.
                The son is 10 years old. How old is the mother?"
    Reasoning:  The son is 10. The mother is 30 years older,
                so 10 + 30 = 40 years old.
    Answer:     40
    Example 4:
    Problem:    "Emily is twice as old as her sister. Her sister
                is 7 years old. How old is Emily?"
    Reasoning:  Emily's sister is 7. Emily is twice as old,
                so 7 × 2 = 14 years old.
    Answer:     14
    Now you try:
    Problem:    "Emily's sister has a brother that is 4 years
                younger than her. Jake is thrice as old as her
                brother. How old is Jake?"
    Answer:
Output:
    Reasoning:  Emily's sister is 7 years old. Her brother
                is 4 years younger than her, so 7 - 4 = 3
                years old. Jake is thrice as old as her
                brother, so 3 × 3 = 9 years old.
    Answer:     9
```

The model demonstrates reasoning behind its answer, leading to improved responses and transparency, as users can understand the model's "thinking" process behind its answers.

B.3 Interacting with AI generated content

In current state of AI, we interact with the content generated by AI on a daily basis without our realization. It is present nearly everywhere on the Internet, including social networks, personalized ads, search engine results, recommendation systems, automated customer support, and even news summaries or product descriptions. That is why everyone should know the basics on identifying content generated by an AI.

Identifying AI-generated content online

Identifying AI-generated online content is tricky due to the existence of advanced AI models. However, some tell-tale signs of AI-generated content can still be spotted. The most tricky part is to recognize text generated by an AI without the help of a detector (computer program). These detectors are widely used; however, their detection accuracies vary and their results are questionable. Our suggestion is to use common sense and look for generic language or repetitive structure in the text. If the text is about something scientific or presents some facts in general, cross-check the information with additional sources because texts generated by an AI may also contain confidently inaccurate facts. AI-generated images have common tell-tale signs, such as odd hands, unnatural textures, or distorted backgrounds. For audio, look for unnatural speech intonation ("robotic" voice).

Despite the presence of these signs, sometimes they can be insufficient to determine if the content is AI-generated. That is why everyone on the Internet should be more careful and not believe everything seen there, use common sense, and always check the presented facts.

Beneficial use of AI

In the thesis, we discussed various risks associated with AI. However, there are also many beneficial uses for artificial intelligence.

AI for creativity

AI tools, mainly generative AI is a great tool to boost human creativity. For example, writers can use AI to brainstorm ideas or get multiple enhancement suggestions for their stories. Artists can use generative AI to visualize ideas or concepts for inspiration before making a commitment to an actual project. This creates a great opportunity for everyone to be a potential artist or content creator.

AI for explanation and learning

Generative AI tools (mainly LLMs) are also a great resource for students to help them study. They can help with research, summarization of long academic papers, drafting document outline, creating quizes for tests, or even generating code. In this area, the AI tools are very useful; however, students must be careful with using such tools because of the inaccuracies of the LLMs as previously mentioned.

B.4 Ethical and fair use of AI systems

This and the next section are grounded in the EU Ethics Guidelines for Trustworthy AI [1] and also draw on these guidelines throughout. Prerequisites for Trustworthy AI are that the AI should be lawful, ethical, and robust. The High-Level Expert Group (HLEG) identified requirements for Trustworthy AI as following:

- Human agency and oversight
- Technical robustness and safety

- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

Human agency and oversight

AI systems should maintain human autonomy, rights, and oversight. They must support informed decision making and fundamental rights, avoiding manipulation of its users. Developers should assess risks to user rights early in the development process to ensure compliance with regulations such as the EU AI Act, as mentioned in the thesis. Oversight of AI systems should ensure control through mechanisms such as human-in-the-loop (HITL) or others. HITL refers to the capability for human intervention in every decision cycle of the system. These control mechanisms ensure that an AI does not undermine human autonomy.

Technical robustness and safety

Technical robustness and safety are essential for a trustworthy AI. AI systems must function reliably, prevent harm, and adapt to changing conditions. They should resist attacks such as data poisoning, where adversaries inject malicious data into the model's training dataset; and adversarial inputs, which, when targeting LLMs, are referred to as jailbreaking, as mentioned in the thesis. AI system should also include fallback mechanisms and minimize unintended outcomes. For example, in the case of LLM, when an adversary succeeds in jailbreaking the model, an independent system should check the output of the model for harmful content, and if it detects such content, the harmful response is deleted or the user cannot continue with the conversation. Finally, high-risk applications, as explained in the thesis on the EU AI Act, must be thoroughly tested to ensure their reliability and accuracy.

Privacy and data governance

Privacy and data governance are essential to protect user rights in AI systems. Developers of AI systems must ensure data protection throughout its lifecycle, from data collection to output generation. AI systems throughout the interaction with the user could collect sensitive data about them; because of this, these systems must ensure that they avoid unlawful profiling or discrimination against them. This type of data must be accessible only by an authorized personel.

Diversity, non-discrimination and fairness

AI models can be trained on datasets that include biases to some groups, as mentioned in the thesis. These biases can lead to discrimination of said groups which does not align with requirements for trustworthy and ethical AI. Diversity and fairness require AI systems to avoid these biases and treat all users equally. AI systems should be designed with inclusivity in mind to help reduce the discrimination that can occur.

Societal and environmental well-being

Environmental well-being is not a particular part of ethics in regard to the users of AI systems; however, it is ethical in a sense in regards to the environment to use such AI systems, which are considerate of the environment in which we live. For example, large AI models should consider using small amounts of electricity for their training.

In addition to environmental well-being, the social impact of AI systems, for example, on democracy and mental health, should be monitored to prevent and minimize harm.

Accountability

Accountability requires responsibility for AI systems and their outcomes. AI systems should be auditable by internal or external auditors to assess risks and allow evaluation without the need for intellectual property (IP) information about such systems. Organizations or individuals (whistleblowers) that found a negative or unlawful impact of AI systems and report this to authorities must be protected to ensure that companies comply with the safety rules and that large corporations do not eliminate such people.

There are always some trade-offs regarding AI systems. These trade-offs must be transparently evaluated to prevent harmful behavior of these systems. The person who makes the decision on allowing/halting further development of such systems must be accountable for their actions.

B.5 Improvements to transparency of AI systems

A primary characteristic of a transparent AI system is communication. AI systems should never misrepresent themselves as humans to their users. It is essential for these systems to clearly communicate that it is not a human being because users have the right to know that they are interacting with an AI system.

Transparent AI systems should make the decision-making process from data gathering to algorithms more traceable to increase its transparency.

Decisions of AI systems should be understandable to users, balancing accuracy and clarity. Systems with significant human impact, for example, systems that assess medical conditions and base treatment on the evaluation, must provide an explanation for their decision-making and must be deployed carefully to prevent causing harm to its users.

To improve the transparency of AI models, we suggest the use of reasoning models. These models use their output to fact-check themselves. For example, the AI agent is tasked with assessing the health condition of a cancer patient. Before the agent provides the final response, it generates a step-by-step explanation of how it interpreted the patient's symptoms, medical history, and relevant medical approaches. This process allows professional in the field to verify the logic behind the assessment and identify potential errors.

B.6 Conclusion

These guidelines aim to promote the ethical, safe, and responsible use of AI. We started with a brief introduction to prompt engineering with examples of some methods to increase the quality of the responses from AI models. Next, we presented tips on how to spot content generated by an AI. We also explained clear benefits of AI systems to balance the guidelines. As for our main section of these guidelines on ethical and fair use of AI systems, grounded in the EU Ethics Guidelines for Trustworthy AI [1], we explained the requirements to achieve trustworthy and ethical AI. In conclusion, these guidelines encourage AI system developers not only to develop functional, but also ethical, transparent, and trustworthy AI systems.

Appendix C

Use of AI in the thesis

- DeepL (2025), `https://www.deepl.com/en/translator`, translation of multiple parts of the text
- Grammarly (2024), `https://app.grammarly.com`, grammatical correction of multiple parts of the text
- Writefull (2025), `https://x.writefull.com/`, grammatical correction and rephrasing of multiple parts of the text

Appendix D

Survey questions

The following pages show our survey questions. For simplicity, the survey was conducted in the Slovak language. The results of the survey are presented in Section 5.1.

Etické a bezpečnostné aspekty prompt engineeringu

Volám sa Marek a som študentom bakalárskeho štúdia na **Fakulte informatiky a informačných technológií Slovenskej technickej univerzity (FIIT STU)**. Tento dotazník je súčasťou mojej bakalárskej práce zameranej na skúmanie **etických a bezpečnostných rizík** spojených s tzv. **prompt engineeringom.** Jedná sa o navrhovanie a optimalizáciu textových pokynov (tzv. "promptov"), ktoré slúžia prevažne na komunikáciu s četbotmi (chatbot), ktoré na pozadí používajú modely umelej inteligencie (LLMs - Large Language Models), ako je napríklad OpenAI ChatGPT, Google Gemini a Microsoft Copilot. Môže však ísť aj o modely, ktorých výstupom nie je text, ale napríklad obrázok, video alebo zvuk. Keďže táto technológia je ešte veľmi nová, tak prirodzene má určité riziká spojené s jej používaním hlavne v oblasti súkromia, bezpečnosti a etiky.

Cieľom tohto dotazníka je zistiť, či ľudia rozumejú týmto technológiám, aké s nimi majú skúsenosti a či si uvedomujú ich bezpečnostné a etické riziká.
Vyplnenie dotazníka trvá **5-8 minút.** Snažte sa, prosím, odpovedať úprimne.

Vopred vám ďakujem za vaše odpovede.

Marek Čederle
xcederlem@stuba.sk

* Označuje povinnú otázku

1. Do ktorej vekovej skupiny patríte? *

Označte iba jednu elipsu.

○ Menej ako 18 rokov
○ 18-24
○ 25-34
○ 35-44
○ 45-54
○ 55 a vyššie

2. Aké je vaše pohlavie? *

Označte iba jednu elipsu.

○ Muž

○ Žena

○ Iné / Nechcem uviesť

3. Do akej kategórie vzhľadom na technologické znalosti by ste sa zaradili? *

Označte iba jednu elipsu.

○ Pracujem v IT (2 roky a viac)

○ Pracujem v IT (menej ako 2 roky)

○ Študent vysokej školy s odborom informatika (alebo podobným technickým odborom)

○ Študent strednej školy s odborom informatika (alebo podobným technickým odborom)

○ Nadšenec do IT

○ Skúsený používateľ internetu

○ Bežný používateľ internetu

○ Iné: ____________________

4. Stretli ste sa s pojmom **umelá inteligencia (AI - Artificial intelligence)?** *

Označte iba jednu elipsu.

○ Áno

○ Nie

5. Vedeli ste, že umelá inteligencia (**AI**) sa využíva v každodenných aplikáciách (napr. chatboty, generovanie obrázkov, preklad textu, atď.)? *

Označte iba jednu elipsu.

○ Áno

○ Nie

6. O ktorých z nasledujúcich nástrojov využívajúcich AI **ste počuli**? (Výber viacerých možností) *

Začiarknite všetky vyhovujúce možnosti.

□ ChatGPT
□ Microsoft Copilot
□ Google Gemini
□ DALL·E
□ MidJourney
□ Stable Diffusion
□ Soundraw
□ DeepL
□ Žiadne z uvedených
□ Iné: ______________________

7. Ktoré z nasledujúcich nástrojov využívajúcich AI **ste už použili**? (Výber viacerých možností) *

Začiarknite všetky vyhovujúce možnosti.

□ ChatGPT
□ Microsoft Copilot
□ Google Gemini
□ DALL·E
□ MidJourney
□ Stable Diffusion
□ Soundraw
□ DeepL
□ Žiadne z uvedených
□ Iné: ______________________

8. O ktorých z nasledujúcich rizík AI **ste už počuli**? (Výber viacerých možností) *

Začiarknite všetky vyhovujúce možnosti.

□ Šírenie dezinformácií
□ Deepfake (falošný obrazový alebo zvukový obsah generovaný pomocou AI)
□ Krádež identity
□ Generovanie škodlivého kódu (malware)
□ Únik/Exfiltrácia osobných údajov
□ Sociálne inžinierstvo (phishing)
□ Generovanie škodlivého obsahu (napr. návody na výrobu zbraní, násilný alebo sexuálny obsah, atď.)
□ Žiadne z uvedených
□ Iné: ______________________

9. S ktorými z nasledujúcich rizík AI **ste sa už osobne stretli alebo ste nimi boli zasiahnutí**? (Výber viacerých možností) *

Začiarknite všetky vyhovujúce možnosti.

□ Šírenie dezinformácií
□ Deepfake (falošný obrazový alebo zvukový obsah generovaný pomocou AI)
□ Krádež identity
□ Generovanie škodlivého kódu (malware)
□ Únik/Exfiltrácia osobných údajov
□ Sociálne inžinierstvo (phishing)
□ Generovanie škodlivého obsahu (napr. návody na výrobu zbraní, násilný alebo sexuálny obsah, atď.)
□ Žiadne z uvedených
□ Iné: ______________________

Hodnotenie úrovne vnímanej hrozby

Pre nasledujúce otázky vyberte číslo na stupnici od 0 do 10, kde **0 znamená vôbec nie závažné** a **10 znamená mimoriadne závažné**

10. Aká vážna je podľa vás hrozba šírenia dezinformácií generovaných pomocou umelej inteligencie (AI)? *

Označte iba jednu elipsu.

	0	1	2	3	4	5	6	7	8	9	10	
vôbe	○	○	○	○	○	○	○	○	○	○	○	mimoriadne závažné

11. Ako vážne je podľa vás riziko použitia AI na krádež identity? *

Označte iba jednu elipsu.

	0	1	2	3	4	5	6	7	8	9	10	
vôbe	○	○	○	○	○	○	○	○	○	○	○	mimoriadne závažné

12. Za aké závažné považujete zneužitie AI pri vytváraní škodlivého alebo nezákonného obsahu (napr. deepfake, návody na vytváranie zbraní)? *

Označte iba jednu elipsu.

	0	1	2	3	4	5	6	7	8	9	10	
vôbe	○	○	○	○	○	○	○	○	○	○	○	mimoriadne závažné

13. Aké obavy máte z možného zneužitia AI na kybernetické útoky, ako sú tvorba škodlivého kódu (malware) a sociálne inžinierstvo (phishing)? *

Označte iba jednu elipsu.

	0	1	2	3	4	5	6	7	8	9	10	
žiad	○	○	○	○	○	○	○	○	○	○	○	veľké

14. Myslíte si, že ľudia plne chápu, ako môže byť umelá inteligencia (AI) zneužitá? *

Označte iba jednu elipsu.

○ Áno

○ Nie

○ Neviem

15. Aké je podľa vás najväčšie riziko umelej inteligencie (AI) a jej potenciálne zneužitie?

____

Záver

Ďakujem vám veľmi pekne za vyplnenie dotazníka.
Verím, že ste odpovedali čo najviac úprimne.
Odpovede budú použité výlučne na akademické účely v rámci mojej bakalárskej práce.

PS: Nezabudnite potvrdiť vyplnenie dotazníka tlačidlom **odoslať**.