Large Language Models in metaphor identification

The case of presuicidal interaction

Gábor Simon, ELTE Eötvös Loránd University, simon.gabor@btk.elte.hu

Tímea Borbála Bajzát, ELTE Eötvös Loránd University, bajzat.timea@btk.elte.hu

Natabara Máté Gyöngyössy, ELTE Eötvös Loránd University, natabara@inf.elte.hu

Péter Gergő Molnár, ELTE Eötvös Loránd University, udwg3g@inf.elte.hu

Noémi Prótár, ELTE Eötvös Loránd University, protarnoemi@student.elte.hu

mailto:noemiprotar@gmail.com

Balázs Indig, ELTE Eötvös Loránd University, indigbalazs@inf.elte.hu

Abstract

This paper addresses the challenge of automating metaphor identification in Hungarian using generative AI technology. Our tool uses a dictionary-augmented chain-of-thought prompting method to perform metaphor identification and implements a widely used protocol for manual annotation. The study employs a curated corpus of online forum posts about suicidal ideation, with manually annotated linguistic metaphors. It describes the challenges of automating metaphor annotation, discusses alternative methodological approaches to address these issues, outlines the project's materials, and assesses the performance of the models (reporting on ~89% of general accuracy and ~79% of balanced accuracy in the case of the best-performing model, GPT-5), concluding with suggestions for further development.

Keywords

metaphor, annotation, automation, generative AI, dictionary

1. Introduction

The present paper has two aims. First, we contextualize the development of an AI annotator assistant within the field of automating metaphor identification. Second, we would like to present our updated findings from the recent testing of the LLM-based metaphor identifier, offering readers insights into the latest advancements and our efforts to enhance the tool's performance.

The main motivation for developing an automatic metaphor annotator lies in its potential application in the field of suicide prevention. According to Eurostat,[1] the suicide rate was the third highest in Hungary in 2021 (with 15.7 deaths per 100,000 inhabitants) among the EU states, and some parts of the country were among the regions most affected by suicide. This trend is particularly alarming considering that a more than 13% decrease was recorded in suicide-related deaths (compared with 2011). Unfortunately, this has not changed in recent years: in 2022, the suicide rate increased (to 16.7 deaths per 100,000 inhabitants); thus, the country retained its

[1] The data are available here: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20240909-1 (last accessed: 18/12/2025).

ominous third place in the EU.[2] These figures mean that nowadays a suicide attempt is made approximately every five hours in the country, making suicide the 8th most common cause of death.[3]

The research presented here is the direct continuation of the project launched to analyze linguistic metaphors in discussions about suicidal thoughts and, hence, to improve suicide prevention (for preliminary results see Simon et al. 2025a). The project aimed to identify common metaphors used by individuals contemplating suicide, which could help alert professionals to potential risks based on their communication about mental health.

The motivation behind targeting metaphorical expressions is twofold. On the one hand, a significant gap can be identified in the literature on studying presuicidal language use: while the prosodic, morphosyntactic, and lexical features of suicide talk have been investigated thoroughly in the last decade (see Homan et al. 2022), analyses focusing on metaphorical language use in presuicidal interaction are almost completely missing. On the other hand, indirect (i.e., non-literal) language is preferred over direct language in talking about suicidal thoughts and feelings, based on a qualitative study carried out in the U.K. (Owen et al. 2012). Even though it is widely recognized that metaphors play an important role in communicating suicidal thoughts and intention (for an early integration of the notion of metaphor into a constructionist approach to therapy see Andersen 1992), the exact patterns of suicide related metaphors are yet to be explored. A systematic corpus-based study of the metaphorical construal of suicide in the Spanish press was performed by Coll-Florit (2026), who found that the most frequent source domains of suicide are WAR, HIDDEN OBJECT, ILLNESS, JOURNEY, LIVING ORGANISM and CRIME. For Hungarian, however, only a few metaphors are mentioned in a qualitative analysis of presuicidal phone calls, such as FALL (OUT OF THE WORLD), SLEEPING, HAVING A REST, JOURNEY or CROSSING BOUNDARIES (B. Erdős 2006: 58–59, 63, 87, 103). Presuicidal metaphors constitute, thus, an important but relatively understudied field of discursive suicidology, where specific patterns are observed (some of them occurring across cultures), but the findings are rarely based on the comprehensive analysis of corpus data, especially in the case of the Hungarian language.

Consequently, exploring metaphors in online Hungarian presuicidal interactions bridges the gap found in previous research, sheds new light on a crucial aspect of presuicidal discourse, and paves the way for further applied research in the field.

One conclusion drawn from corpus compilation and the first rounds of data processing was that annotating metaphors in online suicide-related posts is exceptionally time-consuming and resource-intensive. Although about 70% of the posts in the research corpus contained metaphorical language (with every 10th word labeled on average, Simon et al. 2025a), the volume of the data was insufficient for big data analysis or LLM training, which could have exploited metaphorical language use as a potential feature of suicide talk. A potential solution to the data scarcity dilemma is to automate metaphor detection using AI tools, allowing for the collection of extensive data in a relatively short timeframe. Our paper presents such a solution that aims, however, to provide a broader baseline for the application of generative AI in metaphor analysis.

2. Challenges and advances in the automation of metaphor detection

[2] The data are available here: https://elitmed.hu/ilam/helyunk-a-vilagban/evente-1600-ember-vet-veget-az-eletenek-magyarorszagon-a-del-alfoldi-regio-egesz-europaban-sereghajto (last accessed: 18/12/2025).

[3] The data are available here: https://ongyilkossagmegelozes.hu/tenyek-es-statisztikak-az-ongyilkossagrol/ (last access: 18/12/2025).

One of the central issues analysts face during the exploration of meaning-making in corpora is the difficulty of identifying the target expressions. Danni Yu et al. (2024: 2) identify two main reasons for this challenging situation: the diverse scope of target phenomena (extending from units below the level of words to complex syntactic structures) and the lack of clear form-to-function correspondences. As an example of the latter challenge, Yu et al. (2024: 4) highlight that the English expression *sorry* is used not only for expressing apology but also for assuring the partner of our sympathy during negative events.

From the perspective of corpus annotation, this situation is unwelcome because manual work requires special training, as well as significant time and effort in research, which makes scaling up the analysis and extending it to the nuances of semantic phenomena extremely difficult. Moreover, human analysts may commit mistakes and inconsistencies during the annotation process. Thus, even gold standard corpora are not flawless (cf. the principle of imperfection in Sass 2022: 604–606). The main methodological challenge of annotating metaphors in the corpus can be formulated as follows: how can the reliability and effectiveness of metaphor detection be ensured while maintaining its diverse scope and extending its scale?

The logical solution to this problem would be to automate corpus annotation, as this would increase both the efficiency and the scale of corpus building. However, according to Levchenko et al. (2021: 16), "automatic/automated metaphor identification still remains a challenging problem", due to its limitation to the English language and the requirement of using a manually annotated corpus for testing. In other words, researchers cannot avoid the laborious work of corpus compilation, and available methods must be adapted to the target language.

Taking all these into consideration, the timing of the development of an automated metaphor annotation in Hungarian seems to be optimal: the protocol mostly used for manual metaphor identification (MIPVU, Steen et al. 2010) has been adopted and improved to cover Hungarian data as well (Simon et al. 2025b), and a manually annotated corpus of metaphors in suicide-related online discourse (Simon et al., 2025 a) is ready to use. However, deciding which approach to implement in the automation process requires further elaboration.

In their seminal work on tackling metaphor with computational tools, Veale et al. (2016: 45–46) distinguish between representation-centric and process-centric approaches. The former relies on how metaphors are typically organized in thought and language, while the latter explores how utterances are interpreted as metaphorical in discourse. Representational approaches attempt to observe and describe metaphors as schemata, either conceptual or linguistic (hence the alternative label: schematic approaches). These approaches provide a computational tool with significant generic knowledge about how metaphors are represented. This knowledge is obtained through the thorough analysis of metaphorical structures in corpora or by analyzing large datasets via computational (e.g., vector-based) techniques. Representational solutions to metaphor analysis are thus generic and data-driven in their nature. They can be based on detailed knowledge of the source and target domain vocabulary (Stefanowitsch 2006), constructions of metaphorical language use (Sullivan 2013), or other machine-extracted schematic patterns (e.g., mosaic n-grams, see Indig & Bajzát 2024).

Process-centric approaches (whether corrective or analogical, see Veale et al. 2016, 33) explore and implement the procedures by which human individuals understand the metaphorical nature of utterances. These solutions are based on the core notion of analogy, i.e., cross-domain correspondences between two semantic structures. Analogy is captured as some form of matching between data structures (predicates, graphs, explicated concepts, or other forms of knowledge representation, Veale et al. 2016, 39–45).

Representation-centric approaches show a clear overlap with corpus analysis (adopting a "first cognition then engineering" view). However, they are burdened by similar issues outlined in the introduction. Process-centric approaches, in turn, face challenges due to the complexity of modeling human metaphor comprehension in machine-based environments.

Applying generative AI tools to identify and label metaphorical expressions appears to combine these approaches with unprecedented efficiency. Large language models (LLMs) rely on extracting and learning the patterns of a language from a vast amount of data (representation-centrism), while they can be prompted to exploit a specific reasoning capacity in deciding whether an expression is metaphorically used in a context or not (process-centrism). According to previous research, these models achieve human-level accuracy in named entity recognition, discourse topic identification, sentiment analysis, and relation and information extraction (Yu et al. 2024, 6–7). Moreover, ChatGPT outperformed non-trained annotators in a wide range of tasks, including relevance, topic, and frame detection (Gilardi et al. 2023). In a recent study on a generative AI-based apology annotation experiment in English discourses, Yu et al. (2024) found that GPT-4 (implemented in the Bing chatbot) achieved a robust level of accuracy, which is slightly below the results of the human annotator (92.7% compared to 95.4% at the instance-level, respectively). Consequently, employing LLMs to annotate pragmatic and semantic phenomena "is a viable option", though "human oversight remains necessary" (Yu et al. 2024, 20). In other words, the aim of automation is not the total exclusion of the human analyst from the procedure but rather to provide a technical solution that is effective enough in detecting potential target expressions (knowing that even human annotation might be imperfect) and is capable of scaling up the entire process.

Identifying metaphors, however, presents additional and partly unique challenges for analysts (discussed in Fuoli et al. 2025; Liang et al. 2025). The first is the decision on the unit of metaphor detection: a narrow window of context seems to be typical in recent research, focusing on single pre-selected words or target words in sentences. Labelling multiword expressions or entire texts as metaphors is rather exceptional in current NLP approaches. This tendency seems problematic since metaphorical meaning is frequently not limited to individual words, as in (1). The verb *előtör* 'brake out' can have four meanings in Hungarian, and it is the subject nominal *gondolataim* 'my thoughts' that makes it clear that in the context of the discourse, the verb is used metaphorically ('<a feeling, an emotion> erupts from someone') and not in its literal sense ('bursts or rushes forth').

(1) *előtörtek* az öngyilkossági *gondolataim*
brake-out-pst-3pl the suicide-adj *thought*-pl-poss.3pl
'my suicidal thoughts emerged'

The second issue concerns the amount and the type of information provided for the LLM. The model may be prompted with simple examples of metaphorical and literal language (few-shot prompting); with examples accompanied by explanations about their required analysis (enhanced/fine-grained prompting); or with detailed task instructions, including the step-by-step procedure for metaphor identification, the main annotation categories, and their explanation through illustrative examples (instruction-rich and chain-of-thought prompting). The last approach adopted in previous studies on automation of metaphor detection is the so-called "retrieval-augmented generation" or RAG, where the model is provided with a knowledge resource (an

annotation guideline, for instance), and it requires careful preparation of not only test material but also the integration of the external data source into prompting.

A third decision point concerns how the LLM is used for metaphor detection: directly, i.e., with the description of the task in the prompts (with or without the adaptation of existing manual annotation protocols); or in a "Socratic" or indirect way (Liang et al. 2025, 310–311), when the analyst interacts with the model guiding it with questions and feedback on the results. Finally, the issue of the overlapping categories of indirect meaning-making needs to be addressed somehow: since different types of figurative language (metonymy or personification) may be identified falsely as metaphors by the model (Fuoli et al. 2025, 13–14), handling alternative understandings (and hence, classification) of the same expression is a key issue in both evaluation and fine-tuning.

Facing these challenges, we decided to develop an LLM-based tool that detects linguistic metaphors directly at the word level using the Hungarian-specific version of the MetaID protocol (Simon et al., 2025b), which is based on MIPVU (Steen et al., 2010), a widely used protocol for manual metaphor annotation. We implemented a chain-of thought (CoT) prompting approach in combination with a dictionary-augmented generation (DAG) method (see the next section for details). As an improvement of our tool after the first test results, in a modified version of the prompt, we instructed the model to capture further categories of indirect language use (metonymy, generalization, and specification, respectively).

3. Material and infrastructure[4]

This section summarizes the annotation protocol adapted for automation, presents the features of the test material, describes the project's overall infrastructure (including the integration of dictionary data), and demonstrates both the original and fine-tuned pipelines of our tool.

3.1 The annotation protocol and its implementation in an AI environment

We aimed to develop an LLM-based NLP pipeline for identifying linguistic metaphors in Hungarian texts. The original annotation procedure (Simon et al., 2025b), a hybrid protocol inspired by MIPVU for identifying metaphors, was adapted for the direct detection of metaphors. It defines the basic and contextual meanings of expressions using dictionary entries and allocates the metaphor label when a cross-domain mapping is observed. The method can be outlined briefly as follows. In the procedure of manual annotation, there are two basic criteria for labelling an expression as a potential metaphor: (i) the basic and the contextual meaning need to be distinct based on the dictionary entry, and (ii) a potential cross-domain mapping needs to be identified as the explanation of using the target expression in the given context. Whether the metaphorical contextual meaning is documented in the dictionary as one meaning of the expression or not is out of the scope of the original protocol, since this question concerns the conventionality of the linguistic metaphor. However, if the basic meaning is itself metaphorical (i.e., there is no prior, more physical and more concrete meaning available in the dictionary entry), the expression is excluded from metaphor identification.

To capture metaphorical meaning above and below the level of the word, the basic unit of analysis in the MetaID protocol was shifted from lexical units to morphemes (in accordance with the rich morphological repertoire of the Hungarian language), and the analysis extends to multiword expressions as well (including their semantic relations and prefabricatedness measured as collocations). Due to its complexity, this protocol cannot be directly applied to machine environments. As a temporary solution, our pipeline evaluated the metaphorical usage of

[4] Data and linguistic resources generated in the project are available here: https://github.com/metaphoraid.

individual words in sentences. However, the broader scope of the manual protocol leaves room for future development of the automatic identification tool. Nevertheless, the strict boundaries between metaphors and further categories of indirect language use introduced by the MetaID method were adopted in prompting.

Since the manual annotation protocol (as well as the original MIPVU) relies on meaning descriptions provided in the dictionary, an additional challenge of adapting the manual protocol was to integrate this type of information into the pipeline. We adopted dictionary-augmented generation (DAG) to AI-based annotation (see Hämäläinen 2024). For each target word, the LLM is given pre-processed meaning descriptions (extracted from a reference dictionary by a Python script, see 3.3 for further details). The dictionary entries serve as vantage points for evaluating the potentiality of metaphorization: the meaning listed first in the dictionary was considered the basic meaning in prompting, while finding an appropriate contextual meaning was one of the main tasks of the model.[5] By combining structured lexical knowledge with the reasoning capacity of an LLM on the one hand, and an extended protocol relying on dictionary descriptions on the other hand, our tool applies a "CoT-DAG" experimental design. While existing solutions exploit dictionary augmentation for formal NLP tasks (e.g., lemmatization, Hämäläinen 2024) or in machine translation (Peng et al. 2020), to our knowledge, this is the first attempt to increase the effectiveness of automatically identifying semantic phenomena by using an additional lexicographical resource.

3.2 Test material

For testing the effectiveness of our tool, we used a manually annotated metaphor corpus compiled to investigate patterns of metaphorical language use in presuicidal interactions in Hungarian during the previous stage of the project (Simon et al. 2025a). The database consists of online posts sampled from three archived thematic forums of the Búra self-support community for people with mental disorders and their relatives: Öngyilkosság 'Suicide', Szuicid gondolatok 'Suicidal thoughts', and Krízis, öngyilkossági szándék 'Crisis, suicidal intention'.[6] The posts, which were published between 2010 and 2021, were purposefully selected with special attention to the use of the first-person perspective, discussion of individual experiences related to suicide, and reporting on suicidal ideation. The research corpus totals 22,767 words. During the processes of individual annotation and pairwise curation, all linguistic metaphors were identified and labeled in the corpus, ranging from inflections to multi-word expressions and including the semantic relations between the components of the expressions. The processing of the corpus extended to the annotation of the source domains of the metaphors, adopting the MSDIP protocol (Reijnierse & Burgers 2023) and using the category set[7] of the UCREL Semantic Analysis System.

---

[5] If the target expression is used in a contextual meaning that is not given in the dictionary, our tool will exclude it from the evaluation of metaphorical usage. This weakness of the method clearly suggests that the dictionary-based approach has its own limitations. But the precedents for the lack of contextual meaning in the dictionary are rather infrequent; thus, it does not cause a serious issue in solving the task of automatic metaphor identification. In the future, a possible direction of improving the tool can be to prompt the LLM to substitute the missing meaning definition with a generated sense description relying on the context of the target expression.

[6] The website is available here: https://bura.hu/ (last access: 18/12/2025). The posts address sensitive and critical issues in the lives of the users. However, they are publicly accessible and posted under pseudonyms, ensuring that no personally identifiable information is disclosed. Given these conditions, the Research Support Committee at the Faculty of Humanities at ELTE determined that research ethics approval was not required. During the compilation of the corpus, only the following metadata was collected: the URL of the post, the date of sampling, the date of posting, the username, and the title of the forum.

[7] The USAS category system can be found here: https://ucrel.lancs.ac.uk/usas/usas_guide.pdf (last access: 18/12/2025).

Although this previously established research corpus is not a gold standard Hungarian metaphor corpus, it provides a standardized dataset for examining how metaphors are used in the context leading up to a suicide attempt. To apply it as test material in automating metaphor identification, additional preprocessing steps were performed. First, approximately 30% of the corpus was randomly selected (6192 words in the first experimental round, and 6787 in the second). Then it went through normalization (including the corrections of spelling errors, resolving abbreviations, removing non-standard punctuations and emoticons, and fixing the character encoding inconsistencies). In the first run of the experiments, the entire random sample was used as the input material. In the second run, however, 2,110 words were omitted due to modifications in preprocessing. Thus, 4677 words were tested as potential metaphors, 696 of them were labelled as metaphorical in the corpus, and 3981 were non-metaphorical according to the manual analysis.

3.3 The infrastructure of the project

In addition to the test material, the infrastructural background of the research includes three main components: the dictionary, the LLM, and the digital environment. These components are integrated into a system. Regarding the first component, we chose the Concise Hungarian Explanatory Dictionary)[8] because it is a relatively recent and partly corpus-based lexicographical database of the Hungarian language (frequency information about the usage of the lemmas is based on the data from an earlier version of the Hungarian National Corpus). However, the published version of the dictionary had to be adjusted to align with the aims of the present research. First, all meanings of the lemmas were separated and listed by PoS categories. Then, the "~" symbol was replaced with the actual lemma (with necessary formal modifications). To grasp meaning differences as precisely as possible, broader sense descriptions were split into submeanings (demarcated by the "|" symbol, a semicolon, alphabetical ordering, or documented between <> symbols in the printed version of the dictionary). As a last step, all the examples provided in the dictionary were formally separated from the meanings to obtain the senses of the words introduced with an example in the entries.

The following language models were used for automation: in the first stage of the experimentation, GPT-4o, 4.1-nano, 4.1-mini, 4.1 (OpenAI, 2024), and 5[9] (used only with zero-shot prompting); in the second stage: GPT-4.1, 5, 5.1, 5-chat, 5.1-chat. We conducted experiments with zero-shot and three-shot chain-of-thought prompting. For the latter, we designed alternative experimental settings, using randomly selected examples from the corpus, providing the LLM with positive (metaphorical) and negative (non-metaphorical) data. In the first stage, we also tested the effect of PoS-alignment of the examples on the accuracy of metaphor identification.

In its first version, our genAI-based metaphor identifier consisted of a preprocessing phase (including segmentation, tokenization, lemmatization, and PoS-tagging) applying the huSpaCy NLP tool (Orosz et al. 2023), a prompt creation phase (integrating the relevant meaning descriptions from the dictionary), and an LLM call resulting in a binary decision about the metaphorical use of the word and reasoning. Figure 1 below demonstrates the first pipeline of the research tool. (Testing and development focused on the framed part of the diagram.)

[8] The query interface of the dictionary is available online here: https://eksz.nytud.hu/ (last access: 18/12/2025).
[9] https://openai.com/gpt-5/

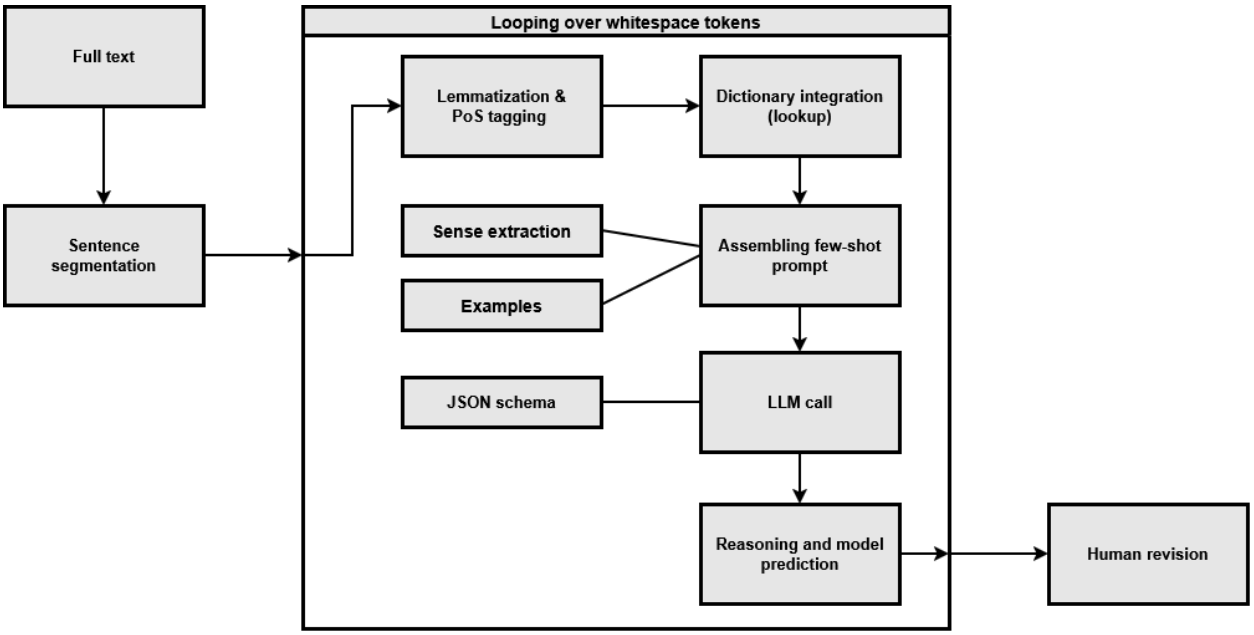

Figure 1. The original pipeline of the research tool for automating metaphor identification

After reviewing the results of the initial experiments, additional refinements have been made to the pipeline to enhance the model's performance and improve the tool's effectiveness. Instead of huSpaCy, we integrated the e-magyar digital language processing system (Indig et al. 2019)[10] into the preprocessing phase. But Figure 2 also highlights three additional modifications that warrant closer examination. First, the preprocessing phase now includes a step for named entity recognition, which allows for the exclusion of proper names and name elements from consideration as potential metaphors. Second, before assembling the prompt, the pipeline filters words that are typically not used as metaphors, such as conjunctions, articles, and negation words. This filtering process reduces the number of target words and consequently, the number of calls to the language model, leading to lower overall costs for using the tool.

[10] The system is available here: http://emtsv.elte-dh.hu:5000/ (last access: 18/12/2025).

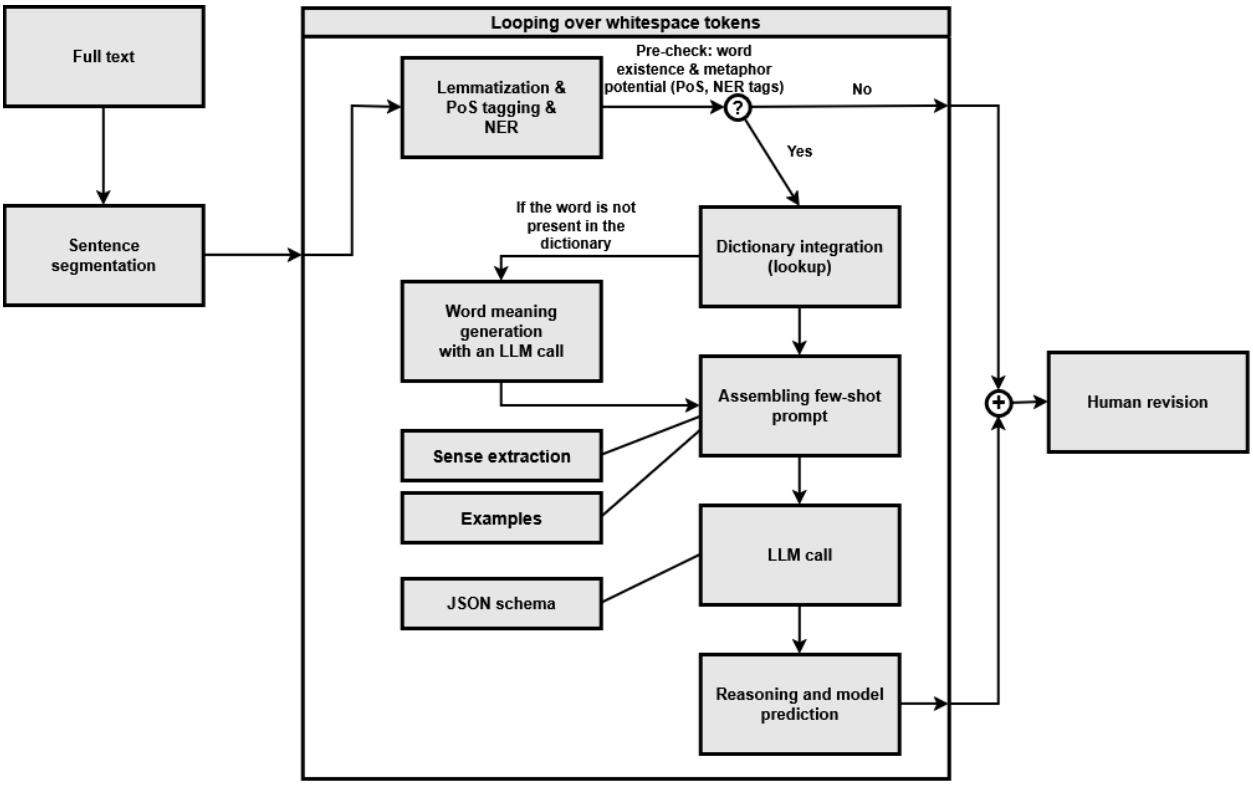

Figure 2. The modified pipeline of the research tool for automating metaphor identification

The third intervention in the pipeline offers a solution for cases where no dictionary entry exists for a specific lemma. It could be attributed to the following causes: (i) the word was incorrectly lemmatized; (ii) the word is not a correct Hungarian word (it is either a word from another language, just a random string of characters, or a Hungarian word with typos); (iii) the word is a correct Hungarian word, but it is not present in the dictionary. To resolve the last problem, an additional LLM call (using GPT-5-chat) was added to the pipeline for meaning creation, in which an additional LLM call generates a meaning description (including an example sentence) for the lemma under analysis. Consequently, the outcome of meaning generation serves as the vantage point in the subsequent analysis.

3.4. Prompting

After thoroughly analyzing the preliminary results (i.e., the performance observed in the first experiments, see Section 4.1 below), we refined our prompts multiple times. In our final system prompt (written in English), we used the following structure:

- Persona description: the persona to which the model must adhere to.
- Task description: the brief description of the task that must be carried out by the model, introducing the category of metaphorical expressions. This section also contains two subsections: one where the input resources are explained, and another that details the steps of identification for the metaphors. These steps are the following:
    - First, the model must identify the contextual meaning of the analyzed word.

    - Then the model must identify the dictionary entry that the contextual meaning matches best. We also accounted for contextual meanings that were not included in the dictionary – in this case, the model returns null for the dictionary entry's index.
    - Then the model must compare the contextual (identified previously) and the basic meaning (coming from the prepared dictionary entry) of the word.[11]
    - If the word's contextual meaning does not match the basic meaning, the model checks whether the word is metaphorical or has any other indirect meanings.
- Definition of a metaphor: In this section, we define the concept of metaphor and demonstrate how the same word can be metaphorical and non-metaphorical based on its context. We also provide the model with dictionary entries and examples of correct reasoning.
- Other indirect meaning types: Three additional categories of indirect language use are introduced (metonymy, generalization, specification) with the same type of examples as metaphors. This step was used to prevent the model from falsely tagging (i.e., from overgeneralizing the category of metaphor).
- Answer format: This section defines the expected JSON output format returned by the model, which has the following attributes:
    - reasoning for the selected contextual meaning (string)
    - index of the dictionary sense corresponding to the contextual meaning of the examined word (integer/null)
    - chain-of-thought reasoning for determining whether the word is used metaphorically (string)
    - index of the other indirect meaning type, or null if the word does not exhibit indirect meaning (integer/null; where 1 = metonymy, 2 = generalization, and 3 = specification)
    - result of the metaphor detection (boolean)

In our user prompt (written in Hungarian), we employed a zero- and a three-shot approach (with or without PoS-alignment in the first round, and only without PoS-alignment in the second round). The model was provided with the target sentence, the word under analysis (both in its original and lemmatized form), its basic and alternative dictionary meanings, and, in the three-shot configuration, illustrative examples.

In a separate LLM call for meaning creation, we prompted the GPT-5-chat model in the user prompt to do the following:

- Check if the word is a meaningful Hungarian word: The model must only create meanings for words it deems meaningful.
- Define the meaning of the word: In this step, the model must define one or more meaning(s) for the given word.

[11] Recall that the basic meaning was automatically borrowed from the dictionary entry provided to the model. In other words, finding a more basic meaning (or refining the dictionary's meaning descriptions) was not the task of the model. Determining the basic meaning is challenging, even in manual annotation, as factors like concreteness, specificity, and human orientation play a significant role in the decision (Steen et al. 2010, 35). To reduce computational costs and inconsistencies in automation, we decided to rely on dictionary entries in the identification of basic meanings instead of choosing one from the meaning description (which would be a separate subtask for the model), or defining one by itself. Asking the LLM to identify the basic meaning of the word can be a potential future improvement of prompting.

- Check the meanings: In this step, the model must check every meaning: if a meaning is found that describes the same meaning as another one, it merges them, and if it finds missing meanings, it completes the list.
- Arrange the meanings: In this step, the model must arrange the meanings, putting the most concrete and most often used meaning in the first place.[12]

Return a structured output, in which all the meanings can be found, with a definition and an example. Our system prompt for this use-case was a brief persona description in English.

In both stages of the experimentation, the models were accessed via their Application Programming Interfaces (APIs) in batch execution of prompts (for the GPT-4 model family), or in case-by-case LLM calls (for the GPT-5 model family). We conducted two experimental runs: one to assess the applicability of generative AI in metaphor detection and another to evaluate the increase in model performance after optimizing the prompt and refining the pipeline.

LLMs have tunable parameters that affect reasoning and output length. For models without a "-chat" marking, the reasoning step enhances problem-solving but is not user-visible. The internal reasoning length can be set to "none", "low", "medium", "high", or "xhigh". Output length, controlled by "Verbosity", has values of "low", "medium", or "high".

When generating text, LLMs output probabilities for vocabulary elements. Temperature affects randomness in sampling: 1.0 means training-like sampling, while 0.0 selects the highest probability element. Reasoning models are more sensitive to temperature variations, whereas chat models can handle adjustments more freely. As a result, reasoning models are restricted to operating at a temperature value of 1.

In the first experiment, the Reasoning component of the GPT-5 model was set to a "low" level, as it is recommended to be used with multi-step decision-making and text interpretation tasks, while the Verbosity component was set to a "medium" level, which refers to the default answer length. In the second run, the components were adjusted to a "medium" level for the reasoning models (GPT-5-nano, mini, GPT-5, GPT-5.1). The Temperature parameter was set to 1.0 by GPT-5 and 5.1, and to 0.0 by the chat model versions.

Note, that the main tasks of the tool were (i) to analyze the grammatical features (i.e., lemma, PoS category and morphological structure) of the target words, (ii) to figure it out what is the contextual meaning of the target word (relying on dictionary sense descriptions), (iii) to decide whether the target word is used metaphorically in the given context (providing also a reasoning for the decision), and (iv) to evaluate whether other indirect usage of the target word can be recognized beyond metaphor. Thus, the tool exploits the reasoning capacity of generative AI models and not the classification capacity of the LLM.

3.5. Evaluation methodology

To evaluate our experiments, we resort to statistical measures of accuracy and balanced accuracy. Simple accuracy is calculated as the percentage of the correctly classified words out of all words processed. As the number of non-metaphors was significantly larger in our annotated dataset than the number of metaphors, the simple accuracy measure is dominated by non-metaphors. Balanced accuracy calculates the percentage of correctly identified words versus all words for both classes

[12] Note that concreteness and frequency are not the only factors of identifying a basic meaning of the word: specificity and human-orientation are also important in the manual protocol (see Steen et al. 2010, 35). However, for a straightforward solution to generate dictionary entries, we prioritize concreteness and commonness over the other two, since they seem easier to evaluate for the LLM. In the future, the meaning-generating prompt can be refined and tested with the additional aspects, too.

(metaphor and non-metaphor), then takes the average of these, ensuring the same weight for the two classes when calculating the accuracy score. We differentiate between prompting techniques based on the example task-solution pairs we provide to the model. We denote the number of examples with the few-shot notation used by most generative AI literature. Here, zero-shot refers to no examples given, providing the detailed task description only, while 3-shot refers to providing the detailed task description and three example task-solution pairs in addition. PoS-alignment of prompts refers to choosing examples from the same PoS as the task to be solved.

4. Results and discussion

4.1 The first stage of the experiments

Table 1 below summarizes the results of our experiments applying zero-shot and three-shot prompting with and without adjusting the examples to the part-of-speech category of the target word. In the evaluation phase, we considered all possible PoS categories[13] occurring in the corpus (third column). Then, the evaluation procedure was restricted to considering only content words (last column). Moreover, accuracy was calculated both as an overall and as a balanced score.

Model	Prompt	all PoS	VERB+NOUN+ADJ+ADV
		balanced accuracy % / accuracy %	
4.1-nano	Zero-shot	63.77 / 45.29	60.56 /41.57
4.1-mini		72.68 / 85.99	71.84 /81.67
4.1		**75.67** / 87.69	**74.97** /83.52
4o		73.59 / 87.37	72.98 / 83.28
5 (R="low" V="med")		69.72 /**89.88**	69.60 /**86.38**
4.1-nano	3-shot non-PoS-aligned	59.49 /36.35	56.57 /33.75
4.1-mini		73.37 /83.58	71.99 /78.58
4.1		74.96 /87.92	74.37 /83.74
4o		69.87 /88.18	69.69 /84.32
4.1-nano	3-shot PoS-aligned	59.35 /36.20	56.67 /34.02
4.1-mini		72.04 /83.11	70.85 /77.98
4.1		74.91 /87.73	74.37 /83.64
4o		70.76 /**88.28**	70.39 /**84.37**

Table 1. Results of the first stage of testing (best scores are highlighted in bold)

We first observed that accuracy increased with the model size in zero-shot testing: GPT-4.1 yielded better results than GPT-4o (although the latter outperformed the smaller versions of 4.1). The highest accuracy scores were obtained with the GPT-5 model. Surprisingly, however, 4o performed slightly better than the other models in 3-shot prompting (being roughly on a par with the results

[13] To exemplify two specific PoS-categories in linguistic metaphors, an adjectival metaphor is for example: *rossz gondolatok* 'bad thoughts' (where *rossz* 'bad' refers to have unpleasant or harmful effect on an organism in its basic meaning); while an example for adverbial metaphor is *próbálnék fent maradni* 'I would try to stay up' (where the adverb *fent* 'up' refers basically to the upper domain of a vertical scale, in the context, however, it has the meaning of emotional and/or mental stability).

of automated metaphor detection in Spanish: the highest general accuracy score in our first round is slightly better than the performance of the GPT-4o model in identifying the metaphorical usage of Spanish verbs; 86.30% compared to our result, 88.28%, respectively, see Puravian, Renau & Riquelme 2024, 7). This suggests that when the task is clearly described and examples are given, the model size is not the only decisive factor. Nonetheless, the 4.1 nano model is extremely underpowered, meaning we must consider a lower limit regarding the size of the model. Additionally, smaller models have reduced capacity in Hungarian, which limits their ability to represent and process complex linguistic expressions. Since Hungarian is a morphologically rich and relatively low-resource language, these limitations restrict the models' effectiveness in solving the task.

When aiming for general accuracy in prompt settings, we found that providing examples improved results with both the 4.1 and 4o models. Therefore, using preprocessed sentences with explanations of metaphorization from the test corpus seems beneficial, and this observation is consistent with the conclusions of Fuoli et al. (2025), who claim that fine-tuned and CoT-prompting yielded the best results. Again, the 4.1 nano and 4.1 mini models were less effective in 3-shot prompting, which highlights that below an optimal model size, examples may hinder the exploitation of the argumentative capacity of the LLM. (Adjusting the examples to the PoS category of the target word resulted only in a modest rise in accuracy, and only with GPT-4o. The larger model proved to be less effective in PoS adjustment.) This observation aligns with the findings of Liang et al. (2025, 315), which indicate that even a single example can enhance the model's performance; however, increasing the number of shots does not lead to improved F1 scores. One potential explanation for these observations is that the examples may confuse the LLM, leading it to overgeneralize the category of metaphor in the actual discourse. Limiting the evaluation to content words resulted in lower accuracy scores. Thus, word classes do not appear to be crucial factors in the language model's performance.

Though the overall results are promising, the picture is less bright when considering balanced accuracy. The highest score, just above 75%, was achieved using zero-shot prompting with GPT-4.1, confirming that the number of shots does not necessarily positively affect the performance of the models. This suggests that the models have difficulties in identifying potential metaphors, despite being very effective at locating non-metaphorical expressions within the corpus. Additionally, few-shot prompting with examples may decrease the LLM's ability to determine if an expression is used metaphorically in discourse. The less promising balanced accuracy scores led us to improve our tool's pipeline and run a second round of testing.

4.2 The second stage of the experiments

After reviewing the preliminary results, modifications were made to the pipeline, which are summarized in Section 3.3. Additionally, the project's infrastructure was expanded to include the GPT-5 model family. Table 2 demonstrates the results of the second round of experimentation, focusing only on the tokens that remained in the sample after preprocessing and filtering.

Model	Prompt	Filtered tokens	
		balanced accuracy %	accuracy %
4.1-nano	Zero-shot	65.67	75.14
4.1-mini		73.86	82.59
4.1		68.61	86.92

5-chat		78.08	86.15
5.1-chat		74.81	88.45
5-nano		69.89	85.91
5-mini		76.02	86.61
5		75.13	88.88
5.1		74.23	88.11
4.1-nano	3-shot (non-PoS-alignment)	66.85	76.42
4.1-mini		74.56	81.55
4.1		73.87	87.70
5-chat		**78.72**	84.85
5.1-chat		74.69	88.34
5-nano		68.75	85.48
5-mini		76.16	86.64
5		75.01	**88.90**
5.1		74.22	88.41

Table 2. Results of the second stage of testing (best scores achieved with different prompting techniques are underlined; overall best scores of the tool are highlighted in bold)

As the results indicate, we successfully enhanced our tool's effectiveness: the highest balanced accuracy score (using GPT-5-chat with 3-shot prompting) improved from approximately 76% to 79%, meaning a nearly 80% performance. In terms of overall accuracy, GPT-5 outperformed the other models slightly, although its results were somewhat more modest in the second round. The accuracy slightly increased from 88.88% (in a zero-shot scenario) to 88.90% (in a three-shot scenario). However, the new pipeline produced a general accuracy score exceeding 85% in 13 out of 18 tests, which is approximately 72% of all trials, demonstrating consistently high performance.

The improvements of our metaphor-detecting tool extend to three aspects of the task. First, filtering out obviously non-metaphorical words after linguistic preprocessing made it more challenging to complete the task at a high level, as evidently false elements are excluded from performance measurement. In other words, the tool had to maintain or increase its performance while working on a restricted sample of the test material. Second, GPT-5-chat proved to be the best-performing model with a notable increase in balanced accuracy even for the more difficult task. Third, the recall (i.e., the rate of true positive predictions) of the metaphorical class rose from 0.55 to 0.70. This signifies that fewer metaphorical words have been "lost", making the tool more suitable as an assistant for human annotation.

Since the revised versions of the prompt predict further types of indirect meaning-making (based on the non-match of the basic and the contextual meaning of the word, as well as on a brief introduction of other indirect meaning categories with examples), we made observations about handling non-metaphoric but indirect meanings by the models, and about the capacity of the LLMs in distinguishing metaphors from other phenomena of indirect meanings. Table 3 summarizes the results of the best-performing model.

Type of indirect meaning	GENERALIZATION	METONYMIC	SPECIFICATION

Metaphor recognized as metaphor	0	0	0
Non-metaphor recognized as non-metaphor	149	63	18
Non-metaphor recognized as metaphor	0	0	0
Metaphor recognized as non-metaphor	12	14	3
Total	161	77	21

Table 3. Detection of other indirect meaning categories by GPT-5-chat (in 3-shot scenario)

Before discussing performance, it is necessary to comment on the "zero lines" in the table. The zeros demonstrate that the expressions recognized as metaphors by the model were not classified under any other indirect meaning. These seemingly "empty" cells serve as evidence that the metaphor decision was effective in the pipeline.

In total, generalization was the most frequent prediction by GPT-5-chat: metonymies only extend to less than half of generalizations, and specifications barely exceed one-eighth of the generalizations (13%). Concerning the second aspect of the task, the model proved to be highly accurate: the lack of any positive predictions shows that it does not categorize other indirect meanings as metaphors; in other words, if the contextual meaning is classified as "other indirect", the model always rejects a positive decision about metaphors. (This does not, unfortunately, indicate that the indirect meaning is predicted correctly – see the typical failures in section 4.4 below.) The low number of false negative results also demonstrates that the risk of confusing metaphors with other indirect meanings is relatively low. Only 7% of the analyzed words were falsely predicted as generalization instead of metaphor, and this proportion is slightly higher in the case of specification (14%). However, the percentage of false negatives rises to 18% when considering metonymies. This suggests that distinguishing metaphors from metonymies poses the greatest challenge for the LLM in this respect. This finding is consistent with the previous literature: as Fuoli et al. (2025: 13–14) report, the "blurred relationship" between metaphors and metonymies caused many discrepancies in automatic metaphor detection. Recognizing that this is a key theoretical issue in contemporary metaphor analysis within cognitive linguistics, the higher percentage of metaphors mistakenly classified as metonymies presents a valuable opportunity for further enhancement of the tool.

4.3 Model behavior and trends

Figure 3 shows a cost tradeoff analysis of different models, comparing their performance in zero-shot and three-shot prompting with the cost of evaluating the target words. (Empty circles correspond to zero-shot tests, while dark ones designate three-shot experiments. Data represents only the results of the second round.)

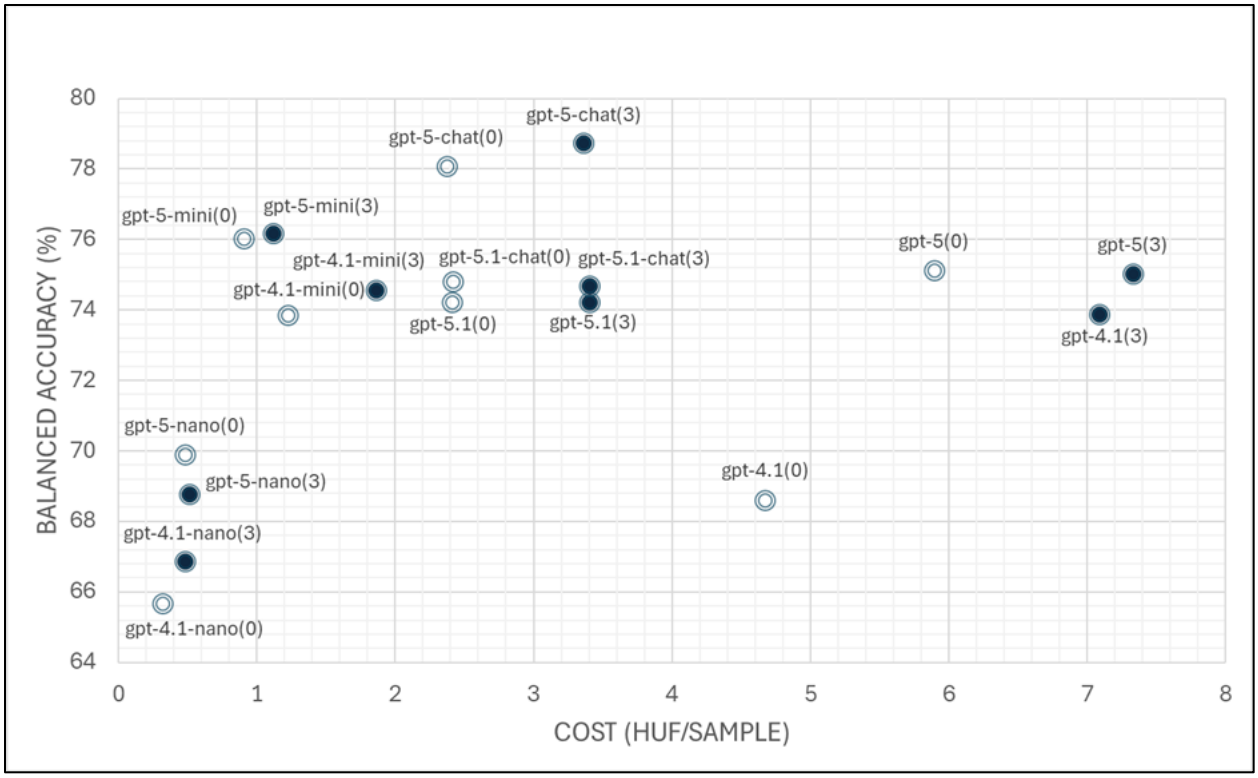

Figure 3. Model performance compared to the cost of use

As the data indicate, GPT-5-nano, mini, and chat outperform other generations. Given the model's size, nano and mini reasoning models are much more effective than their older counterparts. However, with the newer, larger models (5 and 5.1), reasoning becomes counterproductive in terms of costs. Therefore, from the perspective of usage, chat models appear to be more profitable. It is also clear from the diagram that three-shot prompting is more beneficial than zero-shot prompting in terms of balanced accuracy, with only a few exceptions observed in the case of the gpt-5-nano (3-shot is worse) and gpt-5, gpt-5.1 reasoning models (they match the zero-shot performance). Using more examples generally leads to higher accuracy; however, this improvement also incurs approximately 20–30% additional costs.

It is also worth noting that a remarkable reduction in reasoning length can be observed with the 5.1 model: on roughly the same level of accuracy, the cost of running the model was almost halved in the case of GPT-5.1 (compared to GPT-5) with both prompting scenarios. This means that GPT-5.1 is more token-efficient if the tool exploits the reasoning capacity of LLMs in automatic metaphor detection.

4.4 Typical failures

The thorough examination of the models' false predictions would require a new study; thus, here we can only mention some of the recurring failures the models made during testing. Table 4 displays the total number of correct and incorrect decisions made by the best-performing models in the two rounds of experimentation.

	Model prediction

		GPT 4.1 (1st round)		GPT-5-chat-latest (2nd round)	
		Metaphorical	Non-Metaphorical	Metaphorical	Non-Metaphorical
True	Metaphorical	424	344	463	197
	Non-Metaphorical	432	4944	506	3475

Table 4. The predictions of the models against the corpus annotation in the first and the second testing rounds

Before discussing the typical false predictions made by LLMs, we provide a brief overview of the effectiveness of our tool. Keeping the decrease of the amount of target words evaluated in mind (due to the improvements of the pipeline), a remarkable reduction in the proportion of false negative results can be observed (from 44.8% to 29.8%); meanwhile, however, the proportion of false positive hits has been slightly increased (from 8.0% to 12.7%). These numbers clearly indicate that the performance improvement, reflected by a higher balanced accuracy score in the second round, was primarily due to a decrease in the proportion of false negative data. This observation is promising for the annotation process because it is easier for human annotators to eliminate incorrectly identified metaphors than to identify relevant data that has not been recognized by the system at all.

In general, it can be observed that the LLM is more likely to analyze an expression as a metaphor than not; however, it is also common to miss possible metaphors in the input text. Example (2) demonstrates the first type of error, namely, overgeneralization.

(2) Már ott tartunk, hogy ha még egy kísérletem lesz,
already there keep-PRS-1PL that if still one attempt-POSS-1SG be-FUT-3SG

akkor krízis *helyzet* van […].
then crisis *situation* be-PRS-3SG
'We're at the point where if I have one more attempt, it's a crisis *situation* […].'

The tool interpreted the word *helyzet* 'situation' as metaphorical based on the idea that it refers to a dynamic state of events and circumstances rather than a spatial or geographical location (identified as basic meaning). While this argument is valid, the spatial meaning of the word is still active in its contextual use, since a suicide attempt implies physical space. In other words, the expression describes the circumstances more broadly (hence, it can be considered rather a generalization), but it is not the instantiation of the TIME IS SPACE metaphor.[14]

The common source of false negative cases is the model's ignorance or misinterpretation of the basic meaning of the word. Example 3 demonstrates this type of failure.

(3) Munka *után* assszem kisétálok a sínekhez.
work *after* think-PRS-1SG out-walk-PRS-1SG the track-PL-ALL
'*After* work I think I walk to the tracks'

[14] Although *helyzet* 'situation' can be interpreted as the instantiation of SITUATIONS ARE LOCATIONS general metaphor (proposed by one reviewer of our paper), the context of the noun makes it clear that it refers back to the noun *kísérlet* 'attempt', i.e., a concrete physical act. This anaphoric use of the term can maintain a physical (or more basic) interpretation of the expression, keeping the spatial meaning still active, but also making its extension to a less concrete meaning possible. This is why we find the use of the noun to be better categorized as a generalization here.

According to the model's argumentation, the contextual meaning of the postposition *után* 'after' refers to a position in a temporal sequence. While this interpretation is appropriate, the tool mistakenly considers it to have the basic meaning ("This meaning corresponds exactly to the basic meaning in the dictionary"), overlooking the potential metaphorical usage of the word.

Following the qualitative analysis of the false predictions made by GPT-5 in both the zero-shot and three-shot scenarios during the second round (carried out by two authors of the present paper individually), it can be concluded that both types of issues are present in the model's behavior. However, these problems have become less salient in the list of errors. As we observed, true generalizations are sometimes categorized as metaphors (increasing the number of false positives), since the model falsely overinterprets any difference between the contextual and the basic meaning as metaphors. It seems typical that any departure from the physical and concrete basic meaning is understood as an analogical extension in the reasoning of the model, despite the contextual meaning being interpretable with a more general (but not necessarily metaphorical) scope. Moreover, ignoring the basic meaning (as a counterpart of the overgeneralization of analogical extensions) also occurs in the list. Still, it is limited only to specific lemmas (e.g., the adjective *rossz* 'bad', which has a physiological basic meaning in the dictionary but typically an emotional contextual meaning in the corpus).

Beyond the problems of generalization and of skipping the basic meaning, a third rather frequent cause of false negative predictions can be called "the fallacy of matching meanings". By these instances, the model misinterprets the indirect contextual meaning of a word as its conventional usage. It relies on the argument that the correctly identified contextual meaning "can be found" in the dictionary, suggesting that this interpretation reflects a habitual use of the word without any potential metaphorical implications. This fallacy can be considered the opposite of the issue of generalizations mistakenly labeled as metaphors.

The last failure we encountered frequently is related to the category of the register. Since the test material consists of online posts about suicidal ideation, and the task of the model is to find potential metaphors in them, it overgeneralizes the meaning of words in cases when they are used in their basic sense.

(4) Valahogy muszáj *vágnom*, hogy életben maradjak.
somehow must *cut*-INF-1SG that life-INE remain-SUBJ-1SG
'Somehow I have to *cut* to stay alive.'

In (4), the word *vágnom* 'to cut' (in first-person-singular conjugation) is mistakenly identified as metaphorical by the model, arguing in the following way: "In this sentence, however, *vágnom* does not refer to physical cutting, but to the beginning of an action, a change of direction (used in the sense of 'cut through the forest'). This meaning is based on analogy: in physical cutting, someone breaks through an obstacle, opens a path, while here the speaker 'cuts a path' for himself in his own life situation, i.e., he tries to move forward through action. There is therefore an analogy between the two meanings, which is why the usage is metaphorical." Even though this interpretation is acceptable regarding the non-literal use of the verb *vág* 'cut' in other contexts, in actual discourse, it refers to the physical (self-harming) act of cutting ourselves. We can conclude that the LLM cannot take register-specificity, i.e., the actual meaning of the verb referring to self-harming and the broader context of mental crisis, into account, within which a specific

combination of literal and figurative language use makes it difficult to make the same prediction at every occurrence of a target word.

The issues of generalization, ignoring dictionary definitions, the fallacy of matching meanings, and the problem of registers are not the only limitations of our tool. However, they effectively illustrate the challenges we encounter when evaluating the performance of the models. The first three issues can be addressed through prompt optimization and further fine-tuning in the future. However, the incorrect predictions related to the register appear to require manual revision. This process would involve human expertise in the annotation to achieve the highest level of precision.

4.5 Limitations

Beyond the relatively acceptable test results, the tool we developed for automatic metaphor identification evidently has some limitations. The most important one among them is the reliance on the dictionary, which, on one hand, contributes to making the decision of the model more precise, but, on the other hand, restricts the flexibility of the tool. Since one cannot apply any comprehensive corpus-based dictionary of the Hungarian language, specific usages of the words, or even specific words from different registers of the language, may be missing from the dictionary entries integrated in the procedure. While a human annotator can easily overcome this issue based on their native knowledge of the target language, in the case of an automated system, missing data causes the exclusion of expressions from the analysis.

Another limitation resides in the insufficient accuracy of metaphor detection: although the tool almost reaches a 90% percent general accuracy (and 80% in balanced accuracy), both the overgeneralization of the category of metaphors and the ignorance towards specific metaphors and borderline cases are issues that need to be solved in the future. Since these failures partly stem from the influence of the context or the inaccurate handling of meanings separated from the dictionary entries, a careful refinement of prompting may increase the level of accuracy in metaphor identification. Furthermore, providing additional examples of metaphors for the LLM (such as 5-shot or 10-shot prompting) may positively influence the tool's effectiveness.

Finally, the demarcation of metaphors from metonymies seems to be the hardest challenge in the semantic analysis of indirect language use. Either additional examples of metonymic expressions or the elaboration of metonymy identification within the system prompt could help the model in making a subtle distinction between metaphors and metonymies. However, an intense interplay of these two categories can be observed in several discourse fields; therefore, a rigid separation of them seems to be neither feasible nor required.

5. Conclusions

The present paper is the first report on a pipeline for automatically identifying metaphors in Hungarian. This pipeline relies on generative AI models and an NLP infrastructure augmented with meaning descriptions extracted from the dictionary. Considering that (i) we implemented the methodological framework only in zero-shot and few-shot prompting, (ii) we applied the recently released GPT models, and (iii) even human annotators can disagree on the potential metaphorical usage of words in context, i.e., one cannot expect 100% accuracy in the case of manual annotation either, the performance of our solution looks promising. Another advantage of the genAI-based

tool is its speed: our rough estimates suggest that a professional human annotator would need to annotate around 30 words per minute to match the pace of the automated tool[15].

Our most important findings are as follows. We presented novel best practices for metaphor annotation using structured outputs that enforce annotation method-specific chain-of-thought reasoning. We benchmarked state-of-the-art OpenAI model families, with GPT-5-chat-latest being the best solution that also benefits from few-shot prompting. The future development of the tool needs to focus on the iterative refinement of prompts (using 5-shot and 10-shot prompting with positive (metaphorical), negative (non-metaphorical), and ambiguous sentences) and the preprocessing workflow (providing the model only with the meanings relevant for the actual PoS category of the word in case of polysemy or homonymy), as well as on experimentation with large-scale open-source models and fine-tuning. Additionally, an interactive experimental annotation user interface with a real-time dictionary and LLM integration is under development, too.

Concerning the long-term application of the tool, we would like to test and apply automated metaphor detection in the exploration of online presuicidal discourse. An effective genAI-based annotation platform paves the way for compiling a large-scale research corpus and observing the lexicogrammatical patterns of metaphorical expressions related to suicidal ideation. Such a catalogue of presuicidal metaphors, evaluated by experts working in the field of suicide prevention, will then increase further the effectiveness of automatic metaphor extraction on the one hand, and render the assessment of suicide risk possible. Nevertheless, the CoT-DAG pipeline for metaphor identification can be applied to other research areas that involve metaphorical language use, including health communication, politics, scientific, or literary discourse.

Acknowledgements
The research was funded by the National Laboratory for Social Innovation at ELTE University. The paper was also supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00260/25/1).

References

Andersen, Tom 1992. Reflections on Reflecting with Families. In S. McNamee and K. J. Gergen (eds.) Therapy as Social Construction. London, Thousand Oaks, New Delhi: SAGE Publications. 54–68.

B. Erdős, Márta 2006. A nyelvben élő kapcsolat. Egy öngyilkosság-megelőző sürgősségi telefonszolgálat beszélgetéseinek vizsgálata. [Relationship Living in Language. An investigation into conversations on a suicide prevention hotline] Budapest: Typotex.

Coll-Florit, Marta 2026. Suicide Metaphors in the Press: Breaking the Taboo. Metaphor and Symbol 41. 175–191. doi: 10.1080/10926488.2025.2580642

Fuoli, Matteo, Weihang Huang, Jeannette Littlemore, Sarah Turner and Ellen Wilding. 2025. Metaphor identification using large language models: A comparison of RAG, prompt engineering, and fine-tuning. arXiv:2509.24866

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv:2303.15056

Hämäläinen, Mika. 2024. DAG: Dictionary-Augmented Generation for Disambiguation of Sentences in Endangered Uralic Languages Using ChatGPT. In M. Hämäläinen, F. Pirinen, M. Macias and M. C. Avila (eds.) Proceedings of the 9th International Workshop on

[15] The actual cost of using the system is 0.45 cents per whitespace split word, which roughly equals 4% of the minute projection of the minimum wage for an assistant professor at ELTE Eötvös Loránd University, Budapest.

Computational Linguistics for Uralic Languages. Helsinki: Association for Computational Linguistics. 36–40. url: https://aclanthology.org/2024.iwclul-1.4/

Homan, Stephanie, Marion Gabi, Nina Klee, Sandro Bachmann, Anni-Marie Moser, Martina Duri, Sofia Michel, Anna-Marie Bertram, Anke Maatz, Guido Seiler, Elisabeth Stark and Birgit Kleim. 2022. Linguistic features of suicidal thoughts and behaviors: A systematic review. Clinical Psychology Review 95. 102161. doi: https://doi.org/10.1016/j.cpr.2022.102161

Indig, Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Vadász Noémi and Makrai Márton. 2019. One format rule them all – The emtsv pipeline for Hungarian. In A. Friedrich, D. Zeyrek and J. Hoek (eds.) Proceedings of the 13th Linguistic Annotation Workshop. Florence: Association for Computational Linguistics. 155–165. url: https://aclanthology.org/W19-40/

Indig, Balázs and Bajzát Tímea Borbála. 2024. Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian. In In M. Hämäläinen, F. Pirinen, M. Macias and M. C. Avila (eds.) Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages. Helsinki: Association for Computational Linguistics. 96–103. url: https://aclanthology.org/2024.iwclul-1.12.pdf

Levchenko, Olena, Oleh Tyschenko and Marianna Diali. 2021. Automated Identification of Metaphors in Annotated Corpus. (Based on Substance Terms). In N. Sharonova, V. Lytvyn, O. Cheredichenko, Y. Kupriianov, O. Kanischeva, T. Hammon, N. Grabar, V. Vysotska, A. Kowalska-Styczen and I. Jonek-Kowalska (eds.) Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference. Kharkiv: CEUR-WS.org. 16–31. url: https://ceur-ws.org/Vol-2870/paper3.pdf

Liang, Jiahui, Aletta G. Dorst, Jelena Prokic and Stephan Raaijmakers. 2025. Using GPT-4 for Conventional Metaphor Detection in English News Texts. Computational Linguistics in the Netherlands Journal 14. 307–341.

Peng, Wei, Chongxuan Huang, Tianhao Li, Yun Chen and Qun Liu. 2020. Dictionary-based Data Augmentation for Cross-Domain Neural Machine Translation. arXiv:2004.02577

OpenAI (2024). GPT-4 Technical Report. arXiv:2303.08774

Orosz, György, Szabó Gergő, Berkecz Péter, Szántó Zsolt and Farkas Richárd. 2023. Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. arXiv:2308.1263

Owen, Gareth, Judith Belam, Helen Lambert, Jenny Donovan, Frances Rapport and Christable Owens. 2012. Suicide communication events: Lay interpretation of the communication of suicidal ideation and intention. Social Science & Medicine 75. 419–428. doi: https://doi.org/10.1016/j.socscimed.2012.02.058

Puraivan, Eduardo, Irene Renau and Nicolás Riquelme. 2024. Metaphor Identification and Interpretation in Corpora with ChatGPT. SN Computer Science 5. 976. doi: https://doi.org/10.1007/s42979-024-03331-0

Pusztai, Ferenc. (ed.). 2003. Magyar értelmező kéziszótár. [Concise Hungarian Explanatory Dictionary]. Budapest: Akadémiai.

Reijnierse, W. Gudrun and Christian Burgers. 2023. MSDIP: A Method for Coding Source Domains in Metaphor Analysis. Metaphor and Symbol 38. 295-310. doi: 10.1080/10926488.2023.2170753

Sass, Bálint. 2022. Principles of corpus querying: A discussion note. Acta Linguistica Academica 69. 599–614. doi: 10.1556/2062.2022.00581

Simon, G., T. B. Bajzát, K. Árvay, J. Ballagó, K. Hauber, Zs. Havasi, Á. Kuna, E. K. Molnár, N. Prótár, E. Szlávich and Zs. Kaló. 2025a. Metaphors as markers for suicidal intent: Patterns in metaphorical language used in online forum posts about suicidal thoughts. In K. Fogarasi and D. Mányi (eds.) Impact of Sociocultural Factors on Health Communication. Berlin, Bruxelles, Chennai, Lausanne, New York, Oxford: Peter Lang. 221–238.

Simon, G., T. Bajzát, J. Ballagó, Zs. Havasi, E. K. Molnár and E. Szlávich 2025b. When MIPVU goes to no man's land: a new language resource for hybrid, morpheme-based metaphor identification in Hungarian. Language Resources & Evaluation 59. 77–108. doi: https://doi.org/10.1007/s10579-023-09705-9

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr and Trijntje Pasma. 2010. A Method for Linguistic Metaphor Identification. From MIP to MIPVU. Amsterdam, Philadelphia: John Benjamins.

Stefanowitsch, Anatol. 2006. Corpus-based approaches to metaphor and metonymy. In A. Stefanowitsch and St. Th. Gries (eds.) Corpus-based Approaches to Metaphor and Metonymy. Berlin, New York: Mouton de Gruyter. 1–16.

Sullivan, Karen. 2013. Frames and Constructions in Metaphoric Language. Amsterdam, Philadelphia: John Benjamins.

Veale, Tony, S Ekaterina Shutova and Beata Beigman Klebanov. 2016. Metaphor. A Computational Perspective. s. l.: Morgan & Claypool Publishers.

Yu, Danni, Luyang Li, Hang Su and Matteo Fuoli. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis. The case of apology. International Journal of Corpus Linguistics 29. 534 – 561. doi: https://doi.org/10.1075/ijcl.23087.yu