UDC 81’42:004.9
DOI https://doi.org/10.32782/folium/2025.7.9

COMPARATIVE ANALYSIS OF THE EFFECTIVENESS OF LARGE LANGUAGE MODELS FOR METAPHOR IDENTIFICATION: ZERO-SHOT AND FINE-TUNING METHODS

Yakiv Bystrov
Doctor of Philology, Professor,
Vasyl Stefanyk Carpathian National University
ORCID ID 0000-0002-6549-8474
Scopus Author ID 56268636100
yakiv.bystrov@pnu.edu.ua

Nestor Bolshakov
Master’s Student at the Faculty of Foreign Languages
Vasyl Stefanyk Carpathian National University
ORCID ID 0009-0004-0917-4514
BolshakovNestor@gmail.com

Key words: *large language models (LLM), metaphor, metaphor identification, fine-tuning, zero-shot, computational linguistics, natural language processing (NLP).*

The article addresses the challenge of automatic metaphor identification, one of the most complex tasks in natural language processing (NLP). Based on the principles of cognitive linguistics, which define metaphor as a fundamental mechanism of thinking (Lakoff & Johnson, 1980), the role of metaphor as a powerful framing tool in political and media discourse is explored. Although the ability to analyse metaphorical patterns at scale is crucial for identifying manipulative technologies, the process of recognising them is complicated by contextual dependence, creativity, and the need for encyclopaedic knowledge. One of the main issues addressed in this article is the assessment of the potential of modern large language models (LLMs) for solving the task of automatic metaphor identification. The paper compares two key approaches: using the so-called ‘innate’ knowledge of models without additional tuning (the ‘zero-shot’ approach) and their specialised adaptation through fine-tuning. The effectiveness of the latest models (as of July 2025) from leading developers was investigated: OpenAI (GPT-4o), Google (Gemini 2.5 Pro, Gemini 2.5 Flash), and Anthropic (Claude Sonnet 4). Special attention was paid to the methodology of the experiment. The analysis was based on the NAACL 2020 Shared Task on Metaphor Detection corpus, and standard binary classification metrics were used to evaluate the performance of the models: precision, recall, and the F1-score. The article describes the fine-tuning procedure and identifies practical limitations associated with varying levels of tool availability in leading artificial intelligence ecosystems. The results of the study showed that the baseline models demonstrate low and unbalanced performance, while the fine-tuning procedure significantly improves their output (F1-Score increases by 24-29%). A comparative analysis of the retrained models revealed that GPT-4o achieves a better balance between recall and precision (F1-Score 64.20%), while Gemini 2.5 Flash retains a slight advantage in precision. The article makes an important contribution to the study of the capabilities of LLMs for analysing figurative language, demonstrating that fine-tuning is an extremely important method for adapting them to complex linguistic tasks.

ПОРІВНЯЛЬНИЙ АНАЛІЗ ЕФЕКТИВНОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ІДЕНТИФІКАЦІЇ МЕТАФОР: МЕТОДИ НУЛЬОВОГО ЗАПИТУ ТА ДОНАЛАШТУВАННЯ

Яків Бистров
доктор філологічних наук, професор,
Карпатський національний університет імені Василя Стефаника

Нестор Большаков
студент магістратури факультету іноземних мов,
Карпатський національний університет імені Василя Стефаника

Ключові слова: *великі мовні моделі, метафора, ідентифікація метафор, доналаштування, нульовий запит, комп›ютерна лінгвістика, обробка природної мови.*

У статті окреслено проблему автоматичної ідентифікації метафор – одного з найскладніших завдань у сфері обробки природної мови. Базуючись на положеннях когнітивної лінгвістики, що визначає метафору як фундаментальний механізм мислення (Lakoff & Johnson, 1980), досліджено її роль як інструменту фреймінгу в політичному та медійному дискурсах. Підкреслено, що здатність до масштабованого аналізу метафоричних патернів є важливою для виявлення маніпулятивних технологій, однак процес розпізнавання метафор ускладнюється контекстуальною залежністю, креативністю та необхідністю енциклопедичних знань. Дослідження має на меті здійснення порівняльної оцінки потенціалу великих мовних моделей (LLM) для вирішення завдання автоматичної ідентифікації метафор. У роботі порівнюються два ключові підходи: тестування базових можливостей моделей за допомогою методу нульового запиту (zero-shot) та їх спеціалізована адаптація з використанням методу доналаштування (fine-tuning). Досліджено ефективність флагманських(станом на липень 2025 року) моделей від провідних розробників: OpenAI (GPT-4o), Google (Gemini 2.5 Pro, Gemini 2.5 Flash) та Anthropic (Claude Sonnet 4). Окрему увагу приділено методології обробки експериментальних даних на основі корпусу NAACL 2020 Shared Task on Metaphor Detection. Для оцінки продуктивності моделей використано стандартні метрики бінарної класифікації: точність (Precision), повнота (Recall) та узагальнена оцінка F1-Score. У статті описано процедуру доналаштування та виявлено практичні обмеження, пов’язані з різним рівнем доступності інструментів у провідних екосистемах штучного інтелекту. Результати дослідження продемонстрували, що фундаментальні моделі показують низьку та незбалансовану ефективність, тоді як процедура доналаштування значною мірою покращує їхню продуктивність (F1-Score зростає на 24-29%). Порівняльний аналіз доналаштованих моделей штучного інтелекту виявив, що GPT-4o досягає кращого балансу між точністю та повнотою (F1-Score 64,20%), тоді як Gemini 2.5 Flash зберігає невелику перевагу в точності. Проведений порівняльний аналіз робить важливий внесок у дослідження можливостей LLM для аналізу метафор, демонструючи, що доналаштування є важливим методом для їхньої адаптації до складних лінгвістичних завдань.

Introduction. A new stage in the development of cognitive linguistics, initiated by Lakoff and Johnson’s influential book *Metaphors We Live By*, revealed that metaphor is not merely a peripheral stylistic device but a core mechanism shaping human thought and cognition According to the theory of conceptual metaphor, people systematically draw upon knowledge from concrete, bodily experienced domains (such as journeys, wars, or construction) in order to make sense of and structure abstract concepts, including love, evil, or time. Metaphors are ‘fundamental mechanisms of human cognition’ that shape our understanding by projecting everyday, familiar experiences onto abstract ideas

(Lakoff & Johnson, 1980). This function makes metaphor a powerful device not only for everyday communication but also for exerting targeted influence in socially significant discourses.

The metaphor takes on particular significance in the political and media domains, where it serves as a key tool for 'framing', which is the process of constructing social reality through language. As demonstrated by research in the field of critical analysis of metaphors, particularly in the works of J. Charteris-Black, the strategic use of metaphorical models allows communicators to simplify complex political phenomena, give them emotional colouring and promote certain ideological positions (Charteris-Black, 2004). Metaphorical frames such as 'taxes as a burden,' 'immigration as a flow,' or 'the economy as a sick organism' do not simply describe reality, but actively shape its perception, influencing public opinion and political decisions (Musolff, 2006). In this context, the ability to systematically and scalably analyse metaphorical patterns in large text corpora becomes a critical task for identifying manipulative techniques, analysing propaganda, and understanding the dynamics of public sentiment.

However, automating this process is one of the classic 'hard' tasks for natural language processing (NLP). First of all, the challenge comes from the deep contextual dependence of metaphors: a word can be used literally in one sentence but metaphorically in another, which requires the system to be able to subtly distinguish between shades of meaning. Such a task is known as Word Sense Disambiguation. Additionally, effective recognition of metaphors often requires encyclopedic knowledge that extends beyond linguistic patterns. For example, to interpret the English expression *'He shot down my arguments'* as a metaphor, the system must know that no real weapons were used and, as a result, activate the conceptual metaphor ARGUMENT IS WAR. This knowledge gap, which goes beyond linguistic patterns, makes automatic metaphor recognition one of the most difficult problems for NLP (Shutova, 2010). Moreover, language is a dynamic system that constantly generates new metaphors that cannot be predicted in advance, as well as included in any dictionary or knowledge base.

With the advent of large language models (LLMs) based on the Transformer architecture (Vaswani et al., 2017), new opportunities for solving this problem have emerged. Notably, in Ukrainian linguistics, the advantages and limitations of using large language models have been described in the context of automating the process of genre classification of literary texts (Pasichnyk, Yaromych, 2025). Having been trained on vast amounts of text, these models are capable of capturing complex semantic and contextual patterns, potentially enabling them to identify metaphorical expressions. As a result, two fundamentally different approaches have emerged: using the model's basic knowledge through carefully constructed instructions without providing examples of how to perform a similar task (the zero-shot method) and specialised adaptation of the model to a specific task by fine-tuning it on labelled examples.

This research paper focuses on identifying metaphors at the level of individual lexical units, which is in line with modern computational linguistics methodology, specifically the MIPVU procedure (Steen et al., 2010). Such a token-level approach allows for the creation of the most objective and quantifiable criteria for model evaluation.

For a long time, one of the key problems in the study of metaphor was the lack of uniform, consistent criteria for its identification in real discourse. Researchers often relied on their own intuition, which led to inconsistencies in analysis and made it impossible to reproduce results. To address this issue, the Pragglejaz research group proposed the *Metaphor Identification Procedure* (MIP) in 2007, the first clear and systematic method for identifying metaphorically used words in a text. MIP defines a reproducible, step-by-step process: the analyst identifies the contextual meaning of each lexical unit and then compares it with its most basic, concrete meaning, which is often determined by a dictionary. If the contextual meaning contrasts with the basic meaning but can be understood in comparison with it, the word is marked as metaphorical (Pragglejaz Group, 2007). This key criterion formalised the intuitive definition of 'understanding one thing through another' into a practical rule, enabling different researchers to reliably and consistently identify metaphors, achieving a significantly higher level of consistency.

Based on MIP, J. Steen and his colleagues developed an extended protocol called MIPVU *(Metaphor Identification Procedure VU University)* with more detailed instructions for borderline cases. Using MIPVU, a large corpus of contemporary English text was systematically annotated, resulting in the creation of the *VU Amsterdam Metaphor Corpus* (Steen et al., 2010). This corpus, consisting of 117 text fragments (~190,000 lexical units) from four different genres (academic, news, colloquial, literary), has become the industry's gold standard and benchmark. Each word in these texts was labelled as metaphorical or literal with a high level of agreement between annotators. Virtually all subsequent computer studies on metaphor identification have used the VUA corpus to train and evaluate their models, making it a fundamental empirical basis for the entire field.

The first wave of metaphor identification automation coincided with the emergence of deep

transformer models such as BERT. Researchers have begun to use large pre-trained encoder models (BERT, RoBERTa, etc.) to identify metaphors at the individual word level. An important milestone was the *Shared Task on Metaphor Detection* competition at the NAACL conference in 2020, which used the VUA corpus as the main data for evaluation.

In this competition, nearly all teams that achieved the best results utilised fine-tuned transformer encoder models. In general, the most successful models were based on BERT-class architectures adapted to the task at hand through controlled fine-tuning of specially labelled data. This approach has, in fact, transformed general language models into highly specialised metaphor classifiers. The results of the competition confirmed the effectiveness of this approach: the winner achieved an *F1-Score of approximately 77%* on the VUA test set, which significantly exceeded the baseline models (Leong, Beigman Klebanov & Shutova, 2020). Thus, the fine-tuned transformer encoder has become the benchmark against which all newer approaches are compared.

The latest achievements presented a fundamentally different approach to understanding language, and on this basis, large *generative* language models such as GPT-3 and its successors were developed, differing from previous BERT-class models in both architecture and capabilities. The key differences can be summarised as follows:

Architecture: BERT is an *encoder-only* transformer. It reads the entire sentence at once (bidirectionally) to create a rich contextual representation of each word, making it ideal for text *comprehension* and *classification* tasks. GPT-class models, on the other hand, are *decoder-only* transformers. They read text from left to right and are trained to solve a single task: *predicting the next word.* This makes them ideal models for text *generation* and *continuation*.

The principle of fine-tuning, training, and adaptation: BERT is tuned on a *discriminative* task (e.g., 'fill in the blank in a sentence'), which makes it a powerful 'expert analyst' with a deep understanding of context. However, to perform any new task (e.g., identifying metaphors), it requires fine-tuning on thousands of examples to add a new classifier layer to its 'brain.' GPT, on the other hand, learns from a *generative* task. This training makes it a kind of 'universal writer' that can perform tasks after receiving instructions in a prompt *without any additional training (zero-shot)*. This ability to learn in context, first demonstrated in the work on GPT-3 (Brown et al., 2020), made it possible to perform tasks that previously required significant effort to prepare data and fine-tune models.

Thus, the evolution from MIP and manual annotation, through the first wave of automation with pre-trained BERT classifiers, to modern massive generative models illustrates the progress in metaphor identification research. Each link builds on the previous one: The 'gold standard' annotations from the MIP era are still used to evaluate fundamental and fine-tuned models. The study aims to explore the capabilities and limitations of the latest available technologies (as of July 2025), determining the extent to which modern generative LLMs can match or surpass carefully fine-tuned models of the previous generation in such a complex task.

The purpose of this article is to conduct a comprehensive comparative analysis of the effectiveness of leading large language models for identifying metaphors using two methods: testing the model's capabilities using zero-shot and performance evaluations after specialised fine-tuning. To achieve the set goal, the following research tasks were formulated:

- What is the initial effectiveness of the GPT-4o, Gemini 2.5 (Pro and Flash) and Claude Sonnet 4 models in identifying metaphors without special training?
- How significantly does the fine-tuning procedure with a relatively small sample size (1,500 examples) improve the performance of the GPT-4o and Gemini 2.5 Flash models?
- Which of the pre-tuned models demonstrates the best balance between precision and recall, and how does pre-tuning affect their trade-off between these two metrics?
- What practical limitations exist in the ecosystems of leading AI developers regarding the availability of customisation tools for individual researchers?

Materials and methods. *Data corpus: NAACL 2020 Metaphor Detection Corpus.* The experimental part of the work was conducted based on the *NAACL 2020 Shared Task on Metaphor Detection* competition dataset, which is based on the *Amsterdam Metaphor Corpus (VUA)*. The VUA corpus consists of 117 text fragments extracted from the British National Corpus (BNC) and represents four different genres: academic texts, news, fiction and colloquial speech. This genre diversity ensures broad coverage of different language styles and contexts. All words in the corpus were annotated according to the criterion of metaphoricity in accordance with the MIPVU (Metaphor Identification Procedure VU University) procedure, which is a systematic protocol that ensured a high level of agreement between annotators (Cohen's coefficient $\kappa > 0.8$). According to MIPVU, a word is considered metaphorical if it is used in a non-literal sense.

A subset of data from NAACL 2020 was used for the study. In particular, *1500 sentences* from the training part of the corpus were selected for model

tuning, and an official test set of *203 sentences* was used to evaluate effectiveness. This test set contains *708 words annotated as metaphors.* This sample size was chosen to ensure the practical feasibility of the experiments while maintaining the representativeness of the entire corpus.

Models and tools. The study evaluated a number of leading large language models. Using the zero-shot prompt method, the models were used in their original and publicly available state to test their basic capabilities. Four artificial intelligence models were selected for the study:

• *GPT-4o (OpenAI)* is *OpenAI's* flagship (as of July 2025) multimodal model, known for its powerful reasoning capabilities and human-level performance on many professional and academic benchmarks. In the experiments, the model was used via the official ChatGPT in text mode. (https://chatgpt.com/)

• *Claude Sonnet 4 (Anthropic)* – an advanced model from Anthropic, designed for reliable and 'hybrid' thinking in large-scale tasks. The model has an extended context window (up to 200,000 tokens) and is optimised to achieve a balance between speed and cost. Access to the model was provided through its official website for testing using the zero-shot prompt method. (https://claude.ai/)

• *Gemini 2.5 Pro (Google DeepMind)* is the most powerful version (as of July 2025) in the Gemini series, positioned as a 'thinking model' with advanced capabilities in logical thinking, coding, and multimodal understanding. In her work, it is used as a fundamental model via the Gemini website. (https://gemini.google.com/app)

• *Gemini 2.5 Flash (Google DeepMind)* – a version of Gemini optimised for speed and cost efficiency, providing the best compromise between price and performance. (https://gemini.google.com/app)

GPT-4o and *Gemini 2.5 Flash* were selected for the fine-tuning experiments. GPT-4o was fine-tuned using the OpenAI API, while Gemini 2.5 Flash was fine-tuned using the Google Cloud Vertex AI platform. It is important to note that no fine-tuning was performed for *Claude Sonnet 4* due to the lack of relevant public functionality in the Anthropic API at the time of the study, nor was any fine-tuning planned for Gemini 2.5 Pro (as of July 2025).

Experiment procedure. Zero-Shot prompt testing strategy. Using a zero-shot prompting method, a single unified prompt was developed for queries to each model. The prompt was designed to clearly define the task (identifying metaphorical words in a sentence) and establish a strict output format. Providing a precise definition and example of a metaphor was intended to minimise ambiguity for the models. The full text of the prompt used in the study is provided below:

You are a precise and methodical linguistic analyzer. Your sole task is to identify words used metaphorically in the provided sentence.

A metaphor is a word used in a non-literal sense to create an analogy or suggest a resemblance. For example, in the sentence "He navigated a sea of troubles," the word "sea" is a metaphor.

Your output must strictly follow these rules:

List ONLY the words that are used metaphorically.

Separate the words with a single comma and a space.

Do not include any punctuation attached to the words (e.g., for "track," you must write "track").

If you find no metaphors, you must output the single word: none.

Do not include any introductory text, explanations, or summaries.

Sentence to Analyze:

Each test sentence was added after the line 'Sentence to Analyse:'. This procedure was applied to all four models without any additional tuning, which made it possible to evaluate their basic ability to understand instructions and metaphorical language.

Model fine-tuning procedure. Specialised training datasets were created from the VUA corpus for fine-tuning, formatted according to the requirements of each platform.

• *OpenAI fine-tuning (GPT-4o):* A training file was prepared in *JSONL* format with the *Chat Completion* structure expected by the OpenAI API. Each training example is a JSON object with an array of messages containing messages with the roles system, user, and assistant, where the assistant provides the correct answer (metaphors). This ensures that the model learns from pairs of 'query→response'.

{"messages": [{"role": "system", "content": "You are a linguistic analyzer. Your task is to identify metaphorical words in a sentence. List them separated by a comma, or output 'none' if there are no metaphors."}, {"role": "user", "content": "Sentence: \"\"There are other things he has, on his own admission, not fully investigated, like the value of the DRG properties, or which part of the DRG business he would keep after the break up .\"\"\";;;;"}, {"role": "assistant", "content": "things, on, admission, part, keep, after"}]}

• *Google Vertex AI fine-tuning (Gemini Flash):* For Gemini, in accordance with the requirements of the Google Cloud Vertex AI platform, a separate file was created where each line is a JSON object with the *contents* key. This key contains an array of two objects with the roles *user* (containing the request) and *model* (containing the reference response).

{“contents”: [{“role”: “user”, “parts”: [{“text”: “Identify all metaphorical words in the sentence. Sentence: \”\”There are other things he has, on his own admission, not fully investigated, like the value of the DRG properties, or which part of the DRG business he would keep after the break up .\”\”\”;;;;”}]}, {“role”: “model”, “parts”: [{“text”: “things, on, admission, part, keep, after”}]}]}

Gemini 2.5 Flash was fine-tuned using the Tuning graphical interface on the Google Cloud Vertex AI platform, where the relevant data file was uploaded. After completing the fine-tuning, the resulting custom models were tested on the same test set using a unified prompt to ensure consistency.

Assessment metrics. Standard binary classification metrics were used for quantitative measurement and comparison of model performance in the metaphor identification task: *Precision, Recall* and *F1-Score.*

• *Precision* – the proportion of words that the model correctly identified as metaphors among all words it labelled as metaphors. The formula: *Precision = TP / (TP + FP)*, where TP is true positives and FP is false positives.

• *Recall* – the proportion of real metaphors that the model successfully identified. The formula: *Recall = TP / (TP + FN)*, where FN – false negative results.

• *F1-Score* – the average harmonic between accuracy and precision, providing a single balanced assessment of overall performance. The formula: *F1 = 2 × (Precision × Recall) / (Precision + Recall).*

In the above formulas, TP *(True Positive)* is a word that is a metaphor and has been correctly identified by the model. *FP (False Positive)* – a word that the model incorrectly identified as a metaphor. *FN (False Negative)* – a metaphorical word that the model was unable to detect. All metrics were used to identify metaphors, with the main focus on F1-Score as the most balanced indicator.

Results and discussion. *The effectiveness of fundamental models (Zero-Shot Performance).* In the first stage, four models without prior specialisation were tested to determine their ‘innate’ ability to identify metaphors. The results presented in Table 1 indicate not only low overall efficiency, but also the existence of significantly different, yet equally unbalanced strategies applied by the models.

Data analysis revealed three approaches. *Gemini* family models act as *‘highly accurate but cautious classifiers.’* They demonstrate the highest precision among all tested models (Pro – 69.38%, Flash – a record 80.68%), which proves their ability to avoid false positives. However, this strategy comes at the cost of low recall (Pro – 30.08%, Flash – 20.06%), which causes the models to ignore the vast majority (70-80%) of real metaphors.

At the opposite end of the spectrum is *Claude Sonnet 4*, which employs a strategy of ‘*maximum but chaotic coverage*’ and shows the highest recall (55.23%), identifying more than half of the total number of metaphors. However, as evidenced by the record high false positive rate (FP=1078), this is achieved at the expense of extremely low precision (26.62%). A qualitative analysis of errors showed that the model tends to label almost any emotionally charged, abstract, or even neutral words as metaphors, which makes its results practically unusable without careful manual filtering.

The *GPT-4o* model did not demonstrate a clear advantage in either precision or recall, achieving a mediocre F1-Score. Therefore, it can be concluded that none of the basic models is a reliable tool for automatic metaphor analysis, as each of them suffers from a critical imbalance between key metrics.

The impact of fine-tuning. The fine-tuning procedure on a relatively small sample of 1500 examples radically changed the performance of both models, confirming its critical importance for adapting LLM to specialised linguistic tasks.

GPT-4o demonstrated the most impressive increase in efficiency, with its overall F1-Score soaring from *39.79%* to *64.20%* – a jump of 24.4 percentage points. The key transformation was the increase in recall (from *32.49% to 67.37%*): the fine-tuned model began to find twice as many metaphors as its basic version. This indicates that the fine-tuning allowed the model to move from an uncertain strategy to a balanced and highly effective approach.

Gemini 2.5 Flash has also undergone significant changes (F1-Score has almost doubled, from 32.13% to 61.05%). Remarkably, the model has learned to

Table 1

Results of testing basic models (using Zero-Shot prompting)

Model	TP	FP	FN	Precision	Recall	F1-Score
GPT-4o	230	218	478	51,34%	32,49%	39,79%
Claude Sonnet 4	391	1078	317	26,62%	55,23%	35,92%
Gemini 2.5 Pro	213	94	495	69,38%	30,08%	41,97%
Gemini 2.5 Flash	142	34	566	80,68%	20,06%	32,13%

Note: The total number of metaphors in the test sample (Ground Truth) is 708.

Table 2

Results of testing fine-tuned models

Model	TP	FP	FN	Precision	Recall	F1-Score
GPT-4o	477	301	231	61,31%	67,31%	64,2%
Gemini 2.5 Flash	420	248	288	62,87%	59,32%	61,05%

find a compromise: it 'sacrificed' its precision (from 80.68% to 62.87%) to significantly increase recall (from 20.06% to 59.32%).

Qualitative analysis of errors and strategies. A deeper understanding of the differences in model behaviour is provided by a qualitative analysis of their typical errors.

An analysis of false positives revealed that the base *Claude Sonnet 4* was overly liberal, often mistakenly labelling almost any vivid or abstract words as metaphors. For example, in the sentence *"her romance with the young king was quashed for political reasons,"* the model incorrectly labelled *political* and *romance*, whereas only the word *quashed* was metaphorical. It also often confused metaphor with metonymy (labelling *Whitehall* as a metaphor in the expression '*throws a spanner in the Whitehall machinery*') or marked entire idiomatic expressions, including obviously literal words such as *friend* and *sue*. *The fine-tuned GPT-4o*, on the other hand, exhibited a different pattern of errors: it recognised metaphorical constructions better, but sometimes expanded their boundaries excessively, including auxiliary words in the answer (e.g., *about, shaped* in the phrase "h*ugged about by a brood of... shaped like candle snuffers*") or even technical terms such as *articulated* in the phrase *"articulated or double-decker trams"*.

The False Negatives analysis showed that the basic models, especially Gemini, often missed subtle or conventionalised metaphors. They ignored the metaphorical use of common verbs and prepositions. For example, the fundamental Gemini model did not recognise the metaphorical use of the word *incoming* in the expression "*an incoming Labour government would turn large areas of Whitehall upside down*". The models also had difficulty recognising creative, unusual comparisons, such as *candle snuffers* in the phrase '*hugged about by a brood of smaller roofs shaped like candle snuffers*', which Gemini missed. The fine-tuning improved this situation significantly: *The fine-tuning of Gemini 2.5 Flash* has enabled it to successfully recognise both implicit verbal metaphors (e.g., *exert* in '*exert a fascination*') and more creative constructions such as *snuffers*, indicating a significant increase in its sensitivity to less obvious cases.

A direct comparison clearly demonstrates the effect of fine-tuning. In the sentence '*That is the first of many... quango-like bodies recommended by the review,*' the base GPT-4o found nothing, while the fine-tuned version correctly identified *That*, *bodies*, and *recommended*. This indicates that fine-tuning not only improves performance but also fundamentally changes the model's ability to perform deep semantic analysis.

Discussion and methodological conclusions. A comparative analysis of the adjusted models shows that, although both achieved high and similar results, they retained their unique 'character'. *GPT-4o is the formal winner* according to the F1-Score summary metric and the absolute leader in terms of completeness. This makes it an optimal tool for research aimed at maximising the detection of metaphorical expressions, where the researcher is prepared to perform further manual verification to eliminate a certain number of false positives. Instead, *Gemini 2.5 Flash* retained a slight advantage in precision, which partly makes it a reliable choice when minimising 'noise' in the results is a priority.

It is also important to note that at the time of the study (July 2025), Anthropic's public API did not provide accessible functionality for fine-tuning, which made it impossible to include the Claude model in the second stage of the experiment. This is an important practical limitation for individual researchers and indicates varying levels of openness and accessibility of tools across leading AI platforms, which can significantly influence the directions and opportunities for academic research in this field.

Conclusions. The study aimed to conduct a comprehensive comparative analysis of the effectiveness of leading large language models in the task of metaphor identification using zero-shot prompting and fine-tuning methods. The experiments conducted provided answers to the research questions and led to a number of important conclusions.

First, it was found that base LLMs without special training demonstrate low and unbalanced performance for reliably solving the task. Models either show excessive caution, achieving high precision at the expense of extremely low recall (such as Gemini 2.5 Flash with an F1-Score of 32.13%) or, alternatively, generate a large number of false positives due to their pursuit of recall (such as Claude Sonnet 4 with an F1-Score of 35.92%).

Secondly, it has been experimentally proven that the fine-tuning procedure on a relatively small sample (1500 examples) is an extremely effective method of specialising models. Both pre-trained models

demonstrated a significant increase in the F1-Score performance metric: GPT-4o – by 24.4% (to 64.20%), and Gemini 2.5 Flash – by 28.9% (to 61.05%), which indicates a high potential for fine-tuning to adapt LLM to highly specialised linguistic tasks.

Thirdly, a comparative analysis of fine-tuned models demonstrated that *GPT-4o* achieves a better balance between precision and recall, displaying the highest F1-Score and significantly higher recall (67.37%), making it an optimal tool for research aimed at maximising metaphor detection. *Gemini 2.5 Flash*, on the other hand, retains a slight advantage in precision (62.87%), which can be useful in tasks that require minimising false positives.

Fourth, significant practical limitations were identified that characterise the differences in the ecosystems of leading AI developers. Unlike OpenAI and Google, at the time of the study, the Anthropic platform did not provide public access to fine-tuning tools, which is an important factor for the scientific community.

Prospects for further research include expanding the amount of training data for fine-tuning, testing models on corpora from other languages and genres, and moving from token-level analysis to more complex tasks, such as identifying conceptual metaphors and their frames. However, the results obtained already demonstrate that properly adapted large language models can become a powerful tool in the hands of linguists and researchers of various types of discourse.

BIBLIOGRAPHY

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). *Language Models Are Few-Shot Learners*. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
2. Charteris-Black, J. (2004). *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave Macmillan. https://doi.org/10.1057/9780230000612
3. Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
4. Leong, C., Beigman Klebanov, B., & Shutova, E. (2020). Report on the 2020 Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.3
5. Musolff, A. (2006). Metaphor scenarios in public discourse. *Metaphor and Symbol*. https://doi.org/10.1207/s15327868ms2101_2
6. Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*. https://doi.org/10.1080/10926480709336752
7. Shutova, E. (2010). Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 688–697). Association for Computational Linguistics. https://dl.acm.org/doi/10.5555/1858681.1858752
8. Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T., & Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/celcr.14
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., & Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems*, 30. Curran Associates. https://doi.org/10.48550/arXiv.1706.03762
10. OpenAI API Fine-tuning Documentation https://platform.openai.com/docs/guides/supervised-fine-tuning
11. Google Cloud Vertex AI Fine-tuning Guide. Google Cloud Documentation. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning
12. Anthropic Claude Model Overview. Anthropic. https://www.anthropic.com/claude
13. OpenAI GPT-4 Product Page. OpenAI. https://openai.com/research/gpt-4
14. Google Gemini Model Overview. Google DeepMind Blog. Retrieved from https://deepmind.google/models/gemini/
15. YU-NLPLab. (n.d.). *VU Amsterdam Metaphor Corpus – training subset (first 1500 sentences)* [Data set]. GitHub. https://github.com/YU-NLPLab/DeepMet/blob/master/corpora/VUA/vuamc_corpus_train.csv
16. Jin, G. (n.d.). *VU Amsterdam Metaphor Corpus – test subset (203 sentences)* [Data set]. GitHub. https://github.com/jin530/MelBERT/blob/main/data_sample/VUAtok_sample/test.tsv
17. Пасічник, В., Яромич, М. (2025). Особливості жанрової класифікації літератури за допомогою великих мовних моделей. *Folium,* 6, 132–143. https://doi.org/10.32782/folium/2025.6.19.

Отримано: 25.08.2025
Прийнято: 29.09.2025
Опубліковано: 30.10.2025