

A Comparative Study of Generative AIs' Capabilities for Metaphor Analysis: A Case Study of Fire Metaphors in English Online Discourse

BARAMEE KHEOVICHAI*
SIWAPORN SUVARNAMANI
JANTIRA KOEDKHAM

Faculty of Arts, Silpakorn University, Thailand

*Corresponding author email: kheovichai_b@g.su.ac.th

Article information	Abstract
Article history: Received: 17 Jan 2026 Last revised: 27 Aug 2026 Accepted: 27 Aug 2026 Available online: 29 Aug 2026	Recent advancements in generative AI technology have increased the role of AI in academic research. However, the capability of different generative AIs in analyzing metaphors has not been adequately investigated. The purpose of this paper is to compare the capability of ChatGPT, Gemini, and DeepSeek in analyzing fire metaphors in online discourse. Data were based on the iWeb Corpus, and 100 concordance lines of three verbs that potentially instantiated fire metaphors, namely, "spark," "fuel," and "dampen," were collected for analysis by the human researcher. The data were then analyzed by each of the generative AIs. The analyses were compared against those of the researcher. Cohen's Kappa was used to determine the intercoder agreement. The analysis reveals 14 target domains associated with the fire metaphors. The results indicate that there is substantial agreement between the researcher and each of the generative AIs. However, a closer look at the analysis of individual words reveals variation from moderate to almost perfect agreement, with DeepSeek scoring the highest. The implications of the roles of AI in metaphor research are discussed.
Keywords: Conceptual metaphor Generative AI Fire metaphors Online discourse Corpus-based analysis	

INTRODUCTION

Metaphor is widely used in everyday life. It has been argued to be a crucial tool of human communication and cognition (Lakoff, 1990). Metaphor offers a way to understand complex and abstract concepts in terms of a more familiar and concrete one. With the inception of cognitive linguistics, metaphor is central to the study of language and cognition, providing insights into the human conceptual system and how it is manifested in language.

Among various metaphors that have been investigated, such as JOURNEY and WAR, FIRE has attracted attention from cognitive linguists due to its close association with human life. Particularly, fire has been shown to conceptualize the emotion of ANGER, leading to several metaphoric expressions such as "inflammatory remarks" (Lakoff, 1990, p. 368) and "deep and burning passion" (Kovecses, 2010, p. 222). Moreover, Charteris-Black (2017) conducted an

extensive investigation into religious discourse and found that fire metaphors are prevalent and play a critical role in constructing religious concepts. These indicate that fire metaphors are indeed of critical importance in language and cognition, thereby deserving further investigation.

Despite tremendous contributions made by the conceptual metaphor theory, it has been criticized on the following grounds. First, traditional conceptual metaphor studies tend to use introspection to generate data, thereby lacking an empirical basis. Furthermore, the conceptual metaphor theory mainly emphasizes cognition, and there is a lack of emphasis on linguistic features associated with metaphor.

In response to these criticisms, research has applied corpus linguistic methods in the study of metaphor (Deignan, 2005). Corpus linguistics can offer a strong empirical basis for metaphor research. With a large sample of balanced and systematic data collection, corpus data can lend strong support for claims derived from corpus investigation. Furthermore, analyzing concordance lines and collocations can cast light on co-occurring linguistic features associated with metaphorical expressions. However, recently, new advancements in AI technology may revolutionize research and offer new avenues for corpus linguistic studies.

The advancement of natural language processing (NLP), machine learning, and large language models (LLMs) has led to the development of generative AI (Generative AI) (Curry et al., 2024). This is a form of artificial intelligence that can perform tasks related to processing and generating language outputs. Generative AI comes in the form of a chatbot, which allows users to interact by chatting and typing commands for the Gen AI to perform the required task. The application of Generative AI is vast, and the technology involved is growing rapidly. Generative AI has capabilities to generate language outputs, images, and other forms of media. It can process natural language, allowing humans to type instructions instead of complicated code. It displays logical reasoning, thereby allowing it to perform data analysis (Sengar et al., 2024).

The potential of Generative AI in the field of linguistics has been tested. Crosthwaite & Baisa (2023) have rightly pointed out that generative AI has a profound impact on language learning. They discussed whether ChatGPT can outperform corpus linguistics, especially when it comes to data-driven learning. They argued that while ChatGPT is more user-friendly than corpus linguistics, it is associated with various limitations and thus corpus linguistics is still needed for research and teaching. Davies started integrating LLMs into the corpus interface, like COCA, to help with the analysis. Yet, he issues caveats about its accuracy and potential hallucinations (Davies, 2020).

There are studies that compare ChatGPT's performance with human analysts. For example, Curry et al. (2024) investigated the capacity of ChatGPT to conduct corpus-based discourse analysis. So far, the results indicate that AI still lacks the ability to perform linguistic research or generate language outputs at the level that parallels human competence. Furthermore, although Generative AI has been trained on large language models, the components and the design of these models have not been described. Crosthwaite and Baisa (2023) criticized that

this lack of information renders it difficult to evaluate, unlike a corpus, which tends to provide description and is designed according to the principles of representativeness and balance. In addition, in the domain of metaphor, research on how different Generative AI models' capabilities to analyze it is still lacking.

In the field of metaphor research, previous studies have investigated the capability of generative AI to identify and interpret metaphors (cf. Geng & Nimehchisalem, 2023; Yang et al., 2024). While Hesse (203) argued that AIs lack the bodily experience necessary for metaphorical conceptualization, previous studies demonstrated the potential of generative AIs to analyze metaphors, albeit with some degrees of inaccuracies and hallucinations. Also currently, a number of AI models have been updated and released as newer versions, which are yet to be investigated. Furthermore, insufficient research has been done to compare different generative AIs' capability to analyze metaphors, not only in terms of identifying metaphors but also assigning the accurate target domains.

As a consequence, this study aims to fill the gaps in the previous study by conducting a systematic investigation on the potential of generative AI models in identifying and analyzing metaphors. In so doing, this study aims to compare the metaphor analysis undertaken by human researchers against the analysis from Generative AI models. The three models selected based on their popularity are ChatGPT, Gemini, and DeepSeek (Gewirt, 2025).

Research questions

This research has the following research questions:

1. What target domains are conceptualized through fire metaphors in online discourse?
2. How accurate are the generative AI tools in identifying fire metaphors in online discourses?
3. How accurate are the generative AI Tool's to analyze the target domains of fire metaphors in online discourse?

LITERATURE REVIEW

Conceptual metaphor

While metaphor was traditionally regarded as a literary device, the advancement in cognitive linguistics has shown that metaphor is a conceptual phenomenon. Metaphoric expressions used in everyday communication are not merely words. Instead, they reflect a thought process where humans think and speak of something in terms of others (Lakoff, 1990). By way of illustration, the following examples indicate that humans think and talk about anger as fire.

Boy, am I burned up!

There is a smouldering anger in the black community throughout the country.

(Kovecses, 2010, p. 142, 235)

In these examples, the phrase “burned up” and the word “smouldering” do not refer to real fire. Instead, they refer to anger. These examples indicate that humans think of anger as fire, thereby enabling them to use expressions related to fire with reference to anger. Thinking metaphorically involves transferring the knowledge and understanding from one domain to another. The domain from which we derive the knowledge and understanding is called the source domain. The domain to which we transfer the knowledge and understanding is called the target domain. For instance, in the example above, the source domain is FIRE, whereas the target domain is ANGER. Generally, source domains tend to be something more concrete and familiar to humans, relating to bodily functions or the surrounding environment. Target domains, in contrast, tend to be associated with abstract concepts or experiences that are complicated to understand. Given the complexity and unfamiliarity of the target domains, humans use knowledge and understanding from the source domains to conceptualize the target domains. In so doing, humans transfer familiar and well-delineated concepts from the source domain to the target domain. This is also called mapping. There are various source domains that are associated with humans’ bodily functions and activities. For instance, the concept of JOURNEY tends to be used to talk about complex life experiences such as LIFE and LOVE (Kovecses, 2010).

Fire is also close to everyday experience, as humans use fire to cook or to keep warm. In fact, it has been widely used to talk about emotions such as ANGER and to describe the degree of intensity (Lakoff, 1990). In addition, Kovecses (2010, pp. 241–242) noted the following conceptual metaphors related to fire as a source domain: ANGER IS FIRE, LOVE IS FIRE, IMAGINATION IS FIRE, CONFLICT IS FIRE, ENERGY IS FUEL FOR THE FIRE, and ENTHUSIASM IS FIRE. These are further corroborated by Kovaliuk (2024) and Stefanowitsch (2009). Kovaliuk (2024) investigated idiomatic expressions from the cognitive linguistics perspective and found that ANGER is conceptualized as FIRE. Furthermore, Stefanowitsch (2006) used a corpus-based approach to investigate metaphors of emotions in English and found that HAPPINESS is conceptualized as FIRE.

In addition, fire metaphors have been used in health communication. Guenier et al. (2025) compared metaphors used in the discourse of traditional Chinese medicine and Western medicine. They found that metaphors played a vital role in conveying complex concepts and that fire metaphors were used differently in traditional Chinese and Western medicine. While fire metaphors are used when referring to inflammation in traditional Chinese medicine, they are used in reference to pain in Western medicine. Furthermore, fire metaphors are used extensively in discourse surrounding COVID-19. Fire metaphors are used to convey the urgency and the severity of the COVID-19 situation (Charteris-Black, 2021). In addition, fire metaphors are used to frame the Covid-19 policies (Liu & Li, 2026), and the impact of the fire and war metaphors conceptualizing Covid-19 (Tao et al., 2023). It was found that war metaphors surpass fire metaphors when it comes to enhancing positive emotions and promoting the understanding of threats posed by Covid-19 (Tao et al., 2023).

In media discourse, fire has been shown to shape the understanding and perception of the media’s role in society (Diep, 2023). The use of fire metaphors in media discourse can legitimize certain actions or responses, such as urgency in addressing issues or justifying

control measures to prevent “spreading fires” of misinformation or panic. Hart (2018) examines the prevalence of fire metaphors in media coverage of the 2011 London Riots, highlighting how such metaphors frame the events in specific ways. Fire metaphors evoked images of uncontrollable and destructive forces, potentially justifying strong police responses. Two experiments were conducted to test the framing effects of fire metaphors in media discourses of disorder. Participants exposed to fire metaphors showed greater support for authoritative measures compared to those exposed to non-fire-related metaphors or neutral descriptions. The experiments demonstrated that fire metaphors in the absence of competing images could legitimize strong police measures during civil unrest. The findings support the relevance of metaphor in CL-CDS and suggest that similar framing effects occur across different semiotic modalities.

Research has also been conducted across languages. By way of illustration, Dahiru and Saidu (2018) examined how anger was conceptualized metaphorically in English and Fulfulde, two typologically unrelated languages. The goal was to determine if these languages shared similar anger metaphors and if they differed in the elaboration of these shared metaphors. They found that both languages share certain conceptual metaphors for anger, such as ANGER IS HOT FLUID IN A CONTAINER, ANGER IS FIRE, ANGER IS AN OPPONENT IN A STRUGGLE, and ANGER IS A DANGEROUS ANIMAL, though with variations in metaphorical elaboration and source domain specifications. Additionally, each language had unique anger metaphors not found in the other.

These research studies yield insights into how fire metaphors are used to conceptualize abstract concepts and achieve communicative purpose in various contexts. However, they tend to be genre-specific, such as media (Diep, 2023). However, a large corpus representing language on websites has not been investigated in terms of fire metaphors. Consequently, this research can fill this gap by shedding light on fire metaphors in the iWeb corpus (Davies & Kim, 2019), which not only represents online discourse but also various varieties of English.

Corpus linguistics and its uses in metaphor research

A corpus is a large collection of naturally occurring language that is stored in an electronic format (Hunston, 2002). A corpus can be designed to represent general language use or language used in a specific context. To access a corpus, corpus analysis software is required to process and manipulate data in a way that can facilitate analysis. Some common tools in corpus analysis software are concordance lines, collocation, n-grams, and keyword analysis. Corpus linguistics methodology allows researchers to analyze metaphors in large corpora, thereby offering a strong empirical basis. To identify metaphors in large corpora, corpus tools and appropriate search terms can be used to access words or expressions that are potentially used metaphorically.

Metaphors can be searched for in a corpus in various ways. For example, words that can potentially instantiate the source domains or the target domain under investigation could be searched for and presented in the format of concordance lines. Researchers can then read and analyze the surrounding context to determine if the word is metaphorically used. In addition,

Stefanowitsch (2006) proposes the metaphor pattern analysis approach to investigate metaphors in a corpus. In this case, the investigation identifies both words indicating the source and the target domains co-occurring near each other in context for the analysis of conceptual metaphors. Furthermore, a list of collocations could be analyzed to identify the collocations that indicate metaphorical uses. In addition, if the corpus is annotated with semantic tagging such as USAs (Hardie et al., 2008), semantic tags that are relevant to the source or the target domains of interest can be investigated.

Corpus linguistics can provide a strong empirical basis for and a sophisticated methodology to study metaphor. Various studies used different methods to identify metaphoric expressions and cast further insights into metaphorical expressions and the conceptualization of various domains of experience. As for the present study, given that the iWeb Corpus is not annotated with semantic tagging, metaphor pattern analysis seems to be the most viable option. Having both source and target domain terms present provides ample context for both human analysts and generative AIs to accurately analyze the potential cross-domain mappings.

Studies on generative AI's capabilities to identify and analyze metaphors

Recent studies have started to investigate the capabilities of ChatGPT to identify and analyze metaphors. Studies have analyzed several kinds of data and partly supported the claim that ChatGPT displayed the capacity to analyze metaphors, in terms of identifying metaphors, interpreting the meaning of metaphors in context, and analyzing source and target domains.

Geng and Nimehchisalem (2023) investigated metaphors used in short stories from a book about language assessment. The metaphors were identified using ChatGPT (GPT-4), and they were then validated by experts. The book "Tests and Us: A Collection of Real Stories (Volume 2)" was analyzed in terms of types of metaphors and the connection between metaphors and the themes related to key concepts in language assessment. It was found that structural metaphors constitute the majority of metaphors identified. Moreover, metaphors were found to deepen the depiction of language assessment, especially the injustice and dysfunction of the educational system, as well as the complexity and multifaceted nature of language education. This study offers insights into metaphors and their crucial roles in the conceptualization of language education and assessment. What's more, it showcases the capacity of ChatGPT to analyze conceptual metaphors. However, the analysis was based on one literary genre. Also, it might be argued that metaphors in literary genres tend to be creative and thus they might be obviously distinctive from conventionalized metaphorical expressions used in everyday communications. Regarding the use of AI to find metaphors followed by experts' validation, this approach might limit the findings to metaphorical expressions flagged by the program. There is a possibility that the program might incorrectly analyze metaphors as literal language, which has not been addressed by this study.

Ichien et al. (2023) conducted research into ChatGPT's ability to analyze and interpret metaphors in Serbian poetry. Data were Serbian poems recently published to avoid them from being used as training data. Metaphors were given to humans and ChatGPT (GPT 4.0) to interpret the meaning. The interpretation was then evaluated by a panel of experts. The results

indicated that ChatGPT outperformed human subjects (psychology undergraduate students). In addition, it can interpret the symbolic meaning of selected poems accurately. This indicates ChatGPT's capabilities to understand and interpret figurative language. However, this study has not investigated whether ChatGPT can analyze the cross-domain mapping, assigning the source and the target domains of metaphorical expressions. In addition, other models of generative AI have not been compared.

Hesse (2023) discussed the possibility for AI to detect and interpret metaphors. However, he noted the complexity of interpreting the metaphorical meanings in context, especially when both literal and metaphorical meanings were possible. He argued that in such cases, the knowledge of social and communicative contexts was essential, and AI might not be privy to such knowledge. Furthermore, emotion and artistic value in human minds are crucial to the use and interpretation of metaphor. With that said, it remains to be seen whether the advancement in NLP and AI in general will overcome these hurdles, leading to the accurate identification, comprehension, and generation of metaphors by AI.

Yang et al. (2024) developed methods to identify metaphors using ChatGPT. This method employed computational approaches to automatically identify metaphors in corpora of various kinds, namely, the Vienna University Corpus, which was annotated with metaphors, the Trofi corpus containing Wall Street Journal articles, and the Molt-X corpus, which consisted of sentences from crowd-sourcing. They found that ChatGPT, with an appropriate prompt and supplementary computational techniques, could effectively identify metaphors. However, the researchers acknowledged that there were errors and inaccuracies. The analysis focused only on the verb-noun combinations.

As can be seen from the studies above, ChatGPT has shown the capabilities to identify and analyze metaphors. Yet, the cross-domain mappings of source and target domains have not been adequately investigated. In addition, these studies have investigated various genres but not online discourse due to the assumption that ChatGPT might have already been trained on that data. However, it remains unclear whether ChatGPT has been trained to analyze metaphors in the training data. What's more, other AI models have not been compared against ChatGPT, even though nowadays several AI models have been developed and used widely. Therefore, this paper aims to compare the generative AIs' ability to identify and analyze the target domains of metaphoric expressions in corpus data.

METHODOLOGY

The methodology of this research consisted of two parts: analyses conducted by humans and generative AI models. Data were collected from the iWeb Corpus (<https://www.english-corpora.org/iweb/>) (Davies & Jong-Bok Kim, 2019). This was one of the largest corpora with 14 billion words of data from 22 million web pages from 6 countries. The websites were systematically selected. Each website contained an average of 240 webpages. The selection was based on Alexa.com, which provided information about the top 1,000,000 websites around the world. Then, only webpages with most users from the US, Canada, Ireland, the UK, Australia, and New Zealand were included to ensure that the websites were in English. This

resulted in 200,000 websites. After that, the URLs for the websites were obtained by searching through Google and Bing. The websites that were transactional, pornographic, or inaccessible were excluded. Making sure that these websites contained sufficient language data to be representative of the language used on websites, only those with at least 10,000 words and at least 30 webpages were included. This corpus aimed to be a representative of website language, a major platform for online communication.

To arrive at the findings based on human research, the procedures were as follows. After accessing the iWeb corpus, 3 words that potentially instantiated fire metaphors were selected, namely, “spark,” “fuel,” and “dampen,” based on the study by Charteris-Black (2017, p. 31). These terms represented 3 aspects of organic fires, namely, starting a fire, causing a fire to grow, and causing a fire to stop. These selected verbs had the highest frequency in comparison with other verbs according to the iWeb corpus, as shown in Table 1 below.

Table 1
Frequencies of lexical items indicating fire metaphors in the iWeb Corpus

Aspects of Organic Fire	Lexical Item	Frequency
Starting a fire	spark	168,791
	kindle	111,156
	ignite	78,642
Causing a fire to grow	fuel	88,989
	inflamm	16,351
Causing a fire to stop or end	dampen	34,527
	extinguish	34,516
	snuff out	4,260

The 3 words were used as a basis for generating search terms in iWeb, comprising [spark]_v, [fuel]_v, and [dampen]_v, respectively. 100 concordance lines were randomly selected for each of these words. However, only the lines where the target domains could be established from the context of the concordance lines were included. Additional concordance lines were added if the initial searches resulted in concordance lines with unclear contexts for metaphor analysis. These lines were then analyzed to determine whether they were metaphorically used or not, and what target domains could be identified in each instance. The analysis was based on the Metaphor Identification Procedure (MIP), which has been proposed by the Pragglejaz Group as follows (Pragglejaz Group, 2007):

1. Read the entire text–discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text–discourse
3. (a) For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
- (b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be

- *More concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];*
- *Related to bodily action;*
- *More precise (as opposed to vague);*
- *Historically older; Basic meanings are not necessarily the most frequent meanings of the lexical unit.*
- (c) *If the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.*
- 4. *If yes, mark the lexical unit as metaphorical.*

(Pragglejaz Group, 2007, p. 3)

The analysis was then cross-checked with an intercoder, an English lecturer with a Ph.D. in English as an International Language, specializing in metaphor study. Before analyzing these data, the researcher explained the coding protocols, and the intercoder practiced analyzing 5 concordance lines first. After that, the intercoder analyzed 10% of the data, that is, 30 randomly selected concordance lines from the actual data. Cohen's Kappa (Landis & Koch, 1977) statistic was used to determine the intercoder agreement, as it had been shown that this statistic could calculate the intercoder agreement when analyzing nominal data. In addition, this statistic had been used to calculate agreement between the analyses of two researchers in metaphor research (Reijnierse, 2018). The value of Cohen's Kappa ranged from -1 to 1, and the interpretation of the value could be summarized in Table 2 below based on Landis and Koch (1977). This interpretation had been used by metaphor research, such as Reijnierse (2018). The analyses of the researchers and the intercoder were compared, and the Cohen's Kappa score was 0.964, indicating almost perfect agreement. This degree of agreement supported the validity of the analysis.

Table 2
Strength of agreements associated with Kappa statistics (Landis & Koch, 1977, p. 165)

Value of Kappa	Level of Agreement
< 0	poor
0.00–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	almost perfect

In the next phase, generative AI programs were used to analyze the data, and their capabilities were compared. While there are several generative AI programs, this study focused on the most widely used ones that can be used for data analysis. The selection was based on the programs that were ranked in the top ten according to the popularity index (Gewirt, 2025). However, in the ranks, there were programs that were used for translating or editing, such as Quillbot and DeepL, so these were excluded. In addition, Ms Copilot used the same AI model as ChatGPT; therefore, it was also excluded. The resulting generative AI programs were ChatGPT, Gemini, and DeepSeek. The researchers used ChatGPT 5.2 Plus, Gemini Pro 3.0, and DeepSeek to analyze each set of data.

To arrive at an effective prompt for study, the researchers piloted the prompts with data from another project belonging to the main author (Kheovichai, 2025). The project investigated food metaphors in the Corpus of Contemporary American English. Data were in the concordance line format, which was compatible with this paper, even though the source domain was different. A different dataset was used so that the congruency between the analysis of the researcher and the generative AIs could be determined based on the whole dataset.

First, the author asked the AI to identify the target domain associated with the node word as follows:

“You are a specialist in the field of conceptual metaphor theory. Analyze the concordance lines in terms of what target domains are conceptualized through the metaphoric expressions spark.”

It turned out the program could, to some degree, assign the target domains, but the labelling was different from that of the author. In many cases, the AIs simply expressed the contextual meaning of the source domain words in each concordance line. As Bowen and Watson Todd (2025) argued that ChatGPT was better at deductive coding, the researchers provided these generative AIs with the coding protocols, the very same ones given to the human intercoder, and asked them to analyze the data.

Therefore, the researcher provided the classification in the target domain. At first, the classification was provided in the prompt, rendering it very wordy. Subsequently, the researcher provided the classification in the first attachment instead, and attached the concordance lines as a second attached file. This way, the analysis was more in line with that of a human researcher. However, the researcher would like the AI to code the concordance lines in the table, not just to provide the coded number in its output. Therefore, the researcher revised the prompt, and the final version was as follows.

“You are a specialist in the field of conceptual metaphor theory. Analyze the data in the first attachment. These are concordance lines of the words “spark,” “fuel,” and “dampen.” Analyze if they are metaphorical or not, and assign the target domain associated with these words based on the target domain classification protocol in the second attachment. Write your analysis in the “target domain” column in the first Excel file.”

To familiarize the programs with the data and the analytical process, 5 concordance lines of the words “spark,” “fuel,” and “dampen” from the iWeb Corpus were provided to the generative AIs to practice analyzing. These lines were not part of the data for the actual analysis. Feedback was provided to the AIs when necessary. Furthermore, Bowen and Watson Todd (2025) noted that ChatGPT only provided limited answers despite the prompt commanding it to provide exhaustive coding. In their research on using ChatGPT to analyze logical fallacies in students’ writings, the program only provided 10–15 logical fallacies at a time, even though there were more instances of logical fallacies in the data. In line with their findings, this study used the generative AIs to analyze 10 concordance lines at a time to prevent exhaustion, thereby enhancing accuracy. The results were subsequently merged.

Finally, the analysis conducted by each program was then analyzed in terms of congruency with the researchers' analysis. Cohen's Kappa was used to determine the agreement. As noted above, Cohen's Kappa was a robust statistical test used to determine intercoder reliability. The results were compared across the three generative AI programs to determine their accuracy in metaphor identification and analysis.

RESULTS

This section presents the results from data analysis. First, the analysis of the metaphorical target domains is presented. The section presents the target domains that were identified in the dataset and the frequency information of these target domains. After that, the performance of generative AIs in identifying metaphorical uses and assigning target domains is compared and discussed. The section begins with a comparison of the analysis of all the datasets. After that, it discusses the comparison based on individual words to see if they have an impact on the congruencies or incongruencies between the analysis of human researchers and that of each generative AI. Furthermore, since Gemini and DeepSeek showed the thought process involved in the target domain assignment, the section presents and discusses illustrative cases where there were discrepancies between human analysts and the generative AIs.

Target domain identification

This section describes the researcher's identification of the target domains associated with the metaphorical expressions "fuel," "spark," and "dampen." The analysis reveals 14 target domains as well as literal uses. These 14 target domains and an excerpt from the iWeb Corpus are shown in Table 3. In what follows, the target domains are explained, and examples of collocations indicating the target domains are elaborated.

Table 3
Target domains and excerpts from the corpus

No.	Left	Node	Right	Target
1	aside social inhibitions and boundaries for a moment . They	spark	an instant of joy and delight and inspire a long-term hope for	emotion
2	or complementary hues as part of a thematic display designed to	spark	impulse purchases and multiple sales. # A retailer will	business
3	right now ? These are the types of questions that can	spark	ideas that lead to new inventions in the future. If you	ideas
4	military personnel who served in Syria , Putin also sought to	dampen	any talk of a rift with Damascus , saying Moscow 's partial	discussion
5	The question of whether obesity should be called a disease has	sparked	controversy for most of the last century . In their Obesity	conflict
6	# You 'll also learn how to refactor your diet to properly	fuel	your body without gaining weight or feeling hungry . Then ,	body
7	such an approach would be a far more effective way to	dampen	an overheated economy than today 's blunt tool of a rise in	economy
8	a pretty handy item to have in your garage . It	dampens	engine noise by improving ignition performance of diesel fuel ,	sound

No.	Left	Node	Right	Target
9	despite recent government initiatives that many thought would	dampen	activity. Indeed , more mortgages are available , rental yields	activity
10	's Latino population surged 63 percent in 10 years , largely	fueling	the state 's 12 percent growth since 2000 , according to U.S.	development
11	walls and I would have thought that ribs would tend to	dampen	vibrations in a horns body . Whilst an old Conn may well	vibration
12	body has never been found . # The schoolgirl 's disappearance	sparked	one of the biggest cases Essex Police has ever seen and involved	crime
13	well as raise money , the bank measures were designed to	dampen	Labor calls for a royal commission . # " If the Government	politics
14	Somewhat weighted toward the risky but with over 150 holdings to	dampen	out individual disasters , the T. Rowe Price Health Sciences	harms

1. Emotion: The target domain of emotion can be seen from words indicating emotions co-occurring with the fire metaphors, such as “joy,” “interest,” “anger,” and “concern.” An example is shown in concordance line 1. Here, the target domain of EMOTION can be seen in the word “joy.”

2. Business: This target domain refers to business organizations, business people, and business activities. These can be seen from words such as “purchases,” “market,” “company,” “industry,” and “product.” From concordance line 2, the fire metaphor “spark” co-occurs with “purchases,” which indicates the target domain of BUSINESS.

3. Ideas: This target domain refers to ideas, stories, beliefs, thinking processes, or information. Example words indicating this target domain are “ideas,” “propaganda,” “creativity,” “imagination,” and “memories.” In concordance line 3, the fire metaphor “spark” co-occurs with “ideas,” thereby indicating the target domain IDEAS.

4. Discussion: The target domain of discussion refers to the act of conversation, debate, and discussion. Examples of words indicating this target domain are “talk,” “conversation,” “debate,” and “discussion.” The example can be seen in concordance line 4. The fire metaphor “dampen” co-occurs with “talk,” which suggests the target domain of DISCUSSION.

5. Conflict: The conflict target domain refers to verbal, physical conflict, or disagreement between two parties or more. This can be seen in words such as “controversy,” “row,” “war,” “protest,” and “conflict.” In the table, concordance line 5 shows an instance where the fire metaphor “spark” co-occurs with “controversy,” suggesting the target domain of CONFLICT.

6. Body: This target domain refers to the human or animal body, body parts, as well as bodily functions. This may be gleaned from words such as “body,” “immune,” “brain,” and “nervous tension.” Concordance line 6 provides an example where the fire metaphor “fuel” co-occurs with “body,” which indicates the target domain of BODY.

7. Economy: This target domain refers to the economic system or circumstance. This is linguistically realized in words such as “economy,” “economic,” and “economically.” In concordance line 7, the fire metaphor “dampen” co-occurs with “economy,” suggesting the target domain of ECONOMY.

8. Sound: The target domain concerns various kinds of sound and is instantiated in words such as “sound,” “noise,” and “chants.” This is exemplified in concordance line 8, where the fire metaphor “dampen” co-occurs with a target domain word “noise.”

9. Activity: This target domain is about various kinds of human activities and it can be seen from words such as “activity,” “training,” “turnout,” “ride,” and “usage.” In concordance line 9, the fire metaphor “dampen” occurs with “activity,” which is indicative of the target domain ACTIVITY.

10. Development: This refers to the development or increase in resources, technology, and others. Example words are “growth,” “evolution,” “spread,” and “development.” Concordance line 10 shows an instance of the target domain word “growth,” which refers to the increase in population.

11. Vibration: This target domain concerns the physical force or impact that causes vibration, and this is lexicalized in the word “vibration.” This is shown in example 11, where “dampen” occurs with “vibration,” indicating this target domain VIBRATION.

12. Crime: This target domain is associated with criminal activities and this can be seen from the co-occurring words such as “cases,” “human trafficking,” and “crash” (referring to a car crash). As can be seen from concordance line 12, the fire metaphor “sparked” co-occurs with the “one of the biggest cases,” referring to crime.

13. Politics: This target domain encompasses politicians, political movements, and campaigns, and example words are “calls for” and “campaign,” which are used in political contexts. This is shown in line 13, where the fire metaphor “dampen” co-occurs with “Labor calls,” which refers to political activities.

14. Harm: This target domain refers to danger, threat, or something that can cause a deleterious effect. Example words are “disaster,” “storm,” and “harmful effect.” In concordance line 14, “dampen” co-occurs with “disasters,” instantiating the target domain of HARM.

Apart from metaphorical uses, literal uses of these words can be found where they are used with reference to fire or engines that require fuel for combustion.

The frequencies of the target domains can be shown in Table 4 below. From the table, it can be seen that the five most frequent target domains include EMOTION, BUSINESS, IDEAS, DISCUSSION, and CONFLICT. EMOTION is the most dominant target domain, accounting for almost one-third of the data. This is consistent with previous studies, which identified that the source domain of FIRE is used to conceptualize emotions such as anger (Kovaliuk, 2024) and happiness (Stefanowitsch, 2006).

Table 4
Frequencies of the target domains

Target	Frequency
Emotion	<u>94</u>
Business	<u>28</u>
Ideas	<u>24</u>
Discussion	<u>21</u>
Conflict	<u>20</u>
Body	<u>14</u>
Economy	<u>11</u>
Sound	<u>10</u>
Activity	<u>10</u>
Development	<u>7</u>
Vibration	<u>5</u>
Crime	<u>4</u>
Politics	<u>3</u>
Harms	<u>3</u>
Literal	<u>46</u>
Total	<u>300</u>

Comparative analysis of generative AIs' ability to analyze metaphors

This section discusses the analysis of metaphors by the generative AIs. The results are derived from a comparison between the analysis of human researchers and generative AIs. The comparison was based on Cohen's Kappa Coefficient. In what follows, the overall analysis of the three metaphorical expressions is discussed. After that, the analysis of each word is discussed. Table 5 shows the intercoder reliability test results based on all of the terms.

Table 5
Statistical results of the agreement between human researchers and generative AIs

Generative AIs	Cohen's Kappa	Interpretation
ChatGPT	0.727*	substantial
Gemini	0.768*	substantial
DeepSeek	0.791*	substantial

* $p > 0.01$

According to Table 5, there is a significant agreement between the analysis of human researchers and generative AIs, and the Cohen's Kappa values are higher than 0.7, indicating a substantial agreement. However, DeepSeek has the highest value (0.791), followed by Gemini (0.768) and ChatGPT (0.727), respectively. It seems then that Gemini and DeepSeek are relatively close in terms of their capacities to identify the target domains.

While in general, it seems that the generative AIs have shown the capacity to identify the target domains, a closer look at individual words reveals a more complex picture. Table 6 shows the results from the intercoder reliability tests based on individual words.

Table 6
Statistical analysis of the agreement between human researchers and generative AI for individual words

Words	Generative AIs	Cohen's Kappa	Interpretation
Spark	ChatGPT	0.729*	substantial
	Gemini	0.813*	almost perfect
	DeepSeek	0.814*	almost perfect
Fuel	ChatGPT	0.583*	moderate
	Gemini	0.630*	substantial
	DeepSeek	0.733*	substantial
Dampen	ChatGPT	0.847*	almost perfect
	Gemini	0.846*	almost perfect
	DeepSeek	0.806*	almost perfect

* $p > 0.01$

From Table 6, statistical analysis suggests that GenAIs have varying ability to identify the target domains, and the levels of agreement vary across different source domain words. Such nuances may impact the reliability of generative AIs to assign an accurate target domain. As for “spark”, DeepSeek has the highest Kappa value (0.814), followed by Gemini (0.813), and ChatGPT (0.729). The first two achieve almost perfect agreement, while ChatGPT achieves substantial agreement. With regard to “fuel,” each GenAI seems to perform less accurately. According to Cohen’s Kappa, there is only a moderate agreement between the human analysts and ChatGPT, whereas the other two GenAIs have substantial agreement with the human analyst. Considering “dampen,” the performance of GenAIs is better. Based on Cohen’s Kappa, there is almost perfect agreement between the human analyst and each of the GenAIs, further consolidating the GenAIs’ capacities of assigning target domains associated with “dampen.” As a consequence, while there is a strong potential for GenAIs to analyze metaphors, the discrepancies in their capabilities and variations due to individual words can be observed. Gemini and DeepSeek show the thinking process, but ChatGPT does not automatically provide the explanation. This can provide insights into the incongruencies. A closer look reveals some issues that lead to the discrepancy. Table 7 shows illustrative examples of concordance lines where the analysis diverged.

Table 7
Example concordances with different analyses by human researchers and generative AIs

No.	Left Context	Node	Right Context
1	technologies . # Additionally , hemp grown for biomass could	fuel	a trillion-dollar-per-year energy industry , while improving
2	. Knowledge is power , and when properly harnessed , it	fuels	a successful company . # Technology has had a tremendous

In the first line, the word indicating the target domain is “industry,” and DeepSeek also identifies this word as a potential lexicalization of the target domain. However, the analyst assigned the target domain of business in accordance with the coding protocol because this refers to a domain of business, not the economic system. DeepSeek, in contrast, assigned the target domain of the economy despite the coding protocol.

Another issue is due to the selection of words signaling the target domain. In line 2, the verb “fuel” is surrounded by two nouns, one a grammatical subject “it” referring to knowledge and the other a grammatical object “company.” Given that DeepSeek coded this as the target domain IDEA, it seems that DeepSeek chose the grammatical subject “it” as the instantiation of the target domain IDEA. However, the researcher analyzed the grammatical object “company” as the manifestation of the target domain BUSINESS. The researcher further investigated the word “fuel” when it co-occurs with fire in a non-metaphorical sense, and examples are shown in Table 8.

Table 8
Concordance lines with “fuel” co-occurring with “fire” and “flame”

No.	Left Context	Node	Right Context
1	Just about everything in your home can	fuel	a fire . As a fire burns, the heat created
2	a strong, steady white flame and was	fueled	by gasoline. The standard lamp of the
3	The oven uses the gas as a method of	fueling	the flame . Another kind of oven would

From the table, it can be seen that “fire” and “flame” occur with “fuel” as their object. Thus, it is more likely that the grammatical object of the verb “fuel” instantiates the target domain, given the similar syntactic position. Thus, it seems that when two nouns occur in the vicinity of “fuel,” the generative AIs can be confused and choose an inaccurate lexical item.

It seems that Gemini and DeepSeek are more efficient than ChatGPT for overall analysis and target domain identification of individual words. In addition, all of the GenAIs reach almost perfect agreement with the human analysts when analyzing “dampen” and “spark.” On the other hand, the level of agreement decreases substantially when it comes to the analysis of “fuel.” This may be attributable to the surrounding linguistic contexts where two nouns that potentially instantiate the target domain occur near the source domain word, resulting in GenAIs selecting a different word than the researchers.

CONCLUSION

This research aims to compare generative AIs’ ability to analyze the conceptual metaphors in online discourse. Compiling data from the iWeb Corpus based on the potentially metaphorically used words “spark,” “fuel,” and “dampen,” the researcher compared the agreement between the analysis of the human researcher and each of the generative AIs, namely, ChatGPT 5.2 Plus, Gemini Pro 3, and DeepSeek. The statistical comparison was based on Cohen’s Kappa coefficients.

The data analysis indicates that there are 14 target domains. These are: 1) EMOTION, 2) BUSINESS, 3) IDEAS, 4) DISCUSSION, 5) CONFLICT, 6) BODY, 7) ECONOMY, 8) SOUND, 9) ACTIVITY, 10) DEVELOPMENT, 11) VIBRATION, 12) CRIME, 13) POLITICS, and 14) HARMS. There are also instances of non-metaphorical uses of the fire terms. The EMOTION target domain has the highest frequency, accounting for almost one-third of the data.

This study then used the generative AIs to analyze the concordance lines, and the analyses were compared against those of the human analyst. The results indicate substantial agreement for all of the generative AIs, with DeepSeek having the highest level of agreement. Thus, the generative AIs seem to display a strong potential for analyzing metaphors and identifying target domains. However, a closer look at the analysis of individual words revealed variations and inconsistencies. While there was almost perfect agreement for the analysis of “dampen,” moderate congruencies were observed in the analysis of “fuel.”

Further investigation into the thought processes of the generative AIs provides an explanation for the disagreement. It seems then that the generative AIs sometimes assign a different target domain, despite the instructions in the coding protocol. In some cases, when there are two nouns co-occurring with the fire metaphors, the program selected a different noun from the human analyst. Although instances of this kind were not the majority, they resulted in variation in the analysis and inconsistencies.

Overall, given the statistical results, the generative AIs seem to exhibit a potential to accurately analyze metaphors, especially in terms of target domain assignment. It might be possible to envision generative AIs used as a research assistant or as tools for intercoder reliability tests. While it is more common to collaborate with a co-researcher or invite an expert to be an intercoder, not every researcher can find a collaborator, and the intercoder reliability test tends to be done on some of the data, not in its entirety. The generative AIs, on the other hand, can be used on the whole dataset, further strengthening the analysis. However, there are still some inconsistencies in the analysis conducted by the generative AIs and thus, research on a larger scale is required to further determine the accuracy and inconsistencies in generative AIs’ analytical processes.

This study sheds light on the generative AIs’ ability to analyze metaphors in online discourse. While it shows the potential of generative AIs to analyze metaphors, certain limitations can be observed. First, this study only focused on verbs, and thus future studies can investigate other word classes, such as nouns and adjectives. Second, this study analyzed data from the iWeb Corpus, which represents language used on websites. Further research can explore other genres to determine if the influence of genres will have an impact on the accuracy of generative AIs. Lastly, future research can compare generative AIs’ ability to assign source domains to the metaphorically used words. These avenues of research will undoubtedly broaden our knowledge of generative AIs’ roles and capabilities in linguistic research. Ultimately, the authors believe that generative AIs have the potential to be an assistant that enhances research processes, resulting in the rapid advancement of knowledge more effectively.

ACKNOWLEDGEMENTS

This research received research funding from the Research, Innovation, and Creative Fund, Faculty of Arts, Silpakorn University, in the fiscal year 2025. We are grateful to the reviewers who provided constructive feedback on this paper.

THE AUTHORS

Baramee Kheovichai is an Assistant Professor of English at the Department of English, Faculty of Arts, Silpakorn University, Thailand. He holds a Ph.D. in English Language and Applied Linguistics from the University of Birmingham, UK. His areas of research interest include corpus linguistics, discourse analysis, metaphors, and the use of AI in linguistic research.

kheovichai_b@g.su.ac.th

Siwaporn Suvarnamani is an English lecturer at the Department of English, Faculty of Arts, Silpakorn University, Thailand. She holds an M.A. in English for Business and Industry from King Mongkut's Institute of Technology North Bangkok, Thailand. Her current research interests include business English, language in advertising, and teaching.

note173@yahoo.com

Jantira Koedkham is an English lecturer at the Department of English, Faculty of Arts, Silpakorn University, Thailand. She holds an M.A. in English from the Faculty of Arts, Chulalongkorn University, Thailand. Her current research interests include teaching English as a second language, using AI or digital tools in the English classroom, and English literature.

juntira.k@gmail.com

REFERENCES

- Bowen, N. E. J. A., & Watson Todd, R. (2025). Enhancing ChatGPT-based writing research through effective prompt use. *Teaching English with Technology*, 25(1), 26–40.
- Charteris-Black, J. (2017). *Fire metaphors: Discourses of awe and authority*. Bloomsbury.
- Charteris-Black, J. (2021). Metaphors of the pandemic: Fire and force of nature. In J. Charteris-Black (Ed.), *Metaphors of Coronavirus: Invisible enemy or zombie apocalypse?* (pp. 61–91). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-85106-4_3
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3), 1–4. <https://doi.org/10.1016/j.acorp.2023.100066>
- Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, 4(1), 1–9. <https://doi.org/10.1016/j.acorp.2023.100082>
- Dahiru, S., & Saidu, A. (2018). Metaphorical conceptualizations of ANGER in English and Fulfulde. *European Journal of Literature, Language and Linguistic Studies*, 2(2), 9–21. <https://doi.org/10.5281/zenodo.1203355>
- Davies, M. (2020). *The Corpus of Contemporary American English (COCA)*. <http://corpus.byu.edu/coca/>
- Deignan, A. (2005). *Metaphor and corpus linguistics* (Vol. 6). John Benjamins Publishing.
- Diep, N. L. (2023). Conceptual metaphor “Media is Fire”. *Proceedings of the AsiaCall International Conference*, 4, 63–74.
- Geng, H., & Nimehchisalem, V. (2023). Can ChatGPT analyse textual data? The sub-themes reflected by typical conceptual metaphors in short stories of language assessment. *ASEAN Journal of Applied Linguistics*, 2(December), 16–31.
- Gewirt, D. (2025). *The top 20 AI tools of 2025 - and the #1 thing to remember when you use them*. ZD Net. <https://www.zdnet.com/article/the-top-20-ai-tools-of-2025-and-the-1-thing-to-remember-when-you-use-them/>

- Guenier, A. W., Wang, B., Li, M., & Xing, M. (2025). How metaphors facilitate intercultural health communication: Insights from traditional Chinese medicine doctors in UK clinics. *Journal of World Languages*, 11(2), 319–342. <https://doi.org/10.1515/jwl-2025-0025>
- Hardie, A., Koller, V., Rayson, P., & Semino, E. (2008). Exploiting a semantic annotation tool for metaphor analysis. *Metaphoriké*, 15, 141–160. http://comp.eprints.lancs.ac.uk/1527/1/49_Paper.pdf
- Hart, C. (2018). ‘Riots engulfed the city’: An experimental study investigating the legitimating effects of fire metaphors in discourses of disorder. *Discourse and Society*, 29(3), 279–298. <https://doi.org/10.1177/0957926517734663>
- Hesse, J. (2023). Machines and metaphors: Challenges for the detection, interpretation and production of metaphors by computer programs. *Theoria (Sweden)*, 89(5), 607–624. <https://doi.org/10.1111/theo.12481>
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- Ichien, N., Stamenković, D., & Holyoak, K. J. (2023). Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4), 296–309. <https://doi.org/10.1080/10926488.2024.2380348>
- Kheovichai, B. (2025). A corpus-based analysis of eating metaphors in American English from the perspective of conceptual metaphor and lexical priming theories. *LEARN Journal: Language Education and Acquisition Research Network*, 18(1), 1–22.
- Kovaliuk, Y. (2024). Connecting idioms and metaphors: Where cognitive linguistics meets cognitive stylistics. *B.A.S. British and American Studies*, 30, 157–170. <https://doi.org/10.35923/BAS.30.15>
- Kovecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed.). Oxford University Press. <https://doi.org/10.1023/A:1023919116538>
- Lakoff, G. (1990). *Women, fire and dangerous things*. Chicago University Press.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Liu, Y., & Li, D. (2026). Metaphors and China’s COVID-19 policies: A diachronic analysis of Chinese-English bilingual news editorials. *Text & Talk*, 46 (1), 53–78. <https://doi.org/10.1515/text-2024-0068>
- Davies, M., & Kim, J.-B. (2019). The advantages and challenges of “big data”: Insights from the 14 billion word iWeb corpus. *Linguistic Research*, 36(1), 1–34. <https://doi.org/10.17250/khisli.36.1.201903.001>
- Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1), 1–39. https://doi.org/10.1207/s15327868ms2201_1
- Reijnierse, W. G. (2018). DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2), 129–147. <https://doi.org/10.1007/s41701-017-0026-7>
- Sengar, S. S., Hasan, A. Bin, Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*, 84, 23661–23700. <https://doi.org/10.1007/s11042-024-20016-1>
- Stefanowitsch, A. (2006). Words and their metaphor: A corpus-based approach. In A. Stefanowitsch & S. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 64–105). Mouton de Gruyter.
- Tao, R., Kim, S. J., Lu, L., Kang, J., & McLeod, D. (2023). Fighting fire or fighting war: Examining the framing effects of COVID-19 metaphors. *Health Communication*, 39(10), 2026–2040. <https://doi.org/10.1080/10410236.2023.2253398>
- Yang, C., Chen, P., & Huang, Q. (2024). Can ChatGPT’s performance be improved on verb metaphor detection tasks? Bootstrapping and combining tacit knowledge. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1, 1016–1027. <https://aclanthology.org/2024.acl-long.57>