UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Towards Dynamic Metaphor Identification: Evaluating GPT O-Series Models on Five Metaphoricity Cues in U.S. Trade Corpora

Bas, Berkay; Bloem, J.; Tan, X.

DOI
10.63317/2zr7e2gvwuo9

Publication date
2026

Document Version
Final published version

Published in
The Fifteenth Language Resources and Evaluation Conference (LREC 2026)

License
CC BY-NC

Link to publication

Citation for published version (APA):
Bas, B., Bloem, J., & Tan, X. (2026). Towards Dynamic Metaphor Identification: Evaluating GPT O-Series Models on Five Metaphoricity Cues in U.S. Trade Corpora. In *The Fifteenth Language Resources and Evaluation Conference (LREC 2026): Main Conference Proceedings : 13-15 May, 2026* (pp. 4547-4559). ELRA Language Resources Association. https://doi.org/10.63317/2zr7e2gvwuo9

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:16 Oct 2026

Towards Dynamic Metaphor Identification: Evaluating GPT O-Series Models on Five Metaphoricity Cues in U.S. Trade Corpora

Berkay Bas[♦], **Jelke Bloem**[♣♠], **Xiaojuan Tan**[♥]

♦ Data Science (Information Studies), University of Amsterdam
♣ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
♥ Amsterdam Center for Language and Communication, University of Amsterdam
b.h.bas@hotmail.com, j.bloem@uva.nl, x.tan@uva.nl

Abstract

Although recent advances have focused on detecting metaphors, existing models generally treat them as static entities. There has been little research into identifying dynamic metaphors in discourse. This article addresses this gap by focusing on metaphoricity cues: Linguistic signals that may indicate the activation of metaphoric meaning in different discourse contexts. This study examines the ability of OpenAI's O-series models (O4-mini, O4-mini-high and O3) in detecting five metaphoricity cues in the U.S. trade discourse, including cues of explicit mapping, emphasis, marking, repetition and novelisation. Research results show that the models performed best on repetition and emphasis, while novelisation was the most difficult cue to detect.

Keywords: Dynamic Metaphors, Metaphoricity Cues, Yin-Yang Dynamics of Metaphoricity, Trade Metaphors, GPT O-series, Metaphor Detection

1. Introduction

Metaphors have become an increasingly prominent research topic within Natural Language Processing (NLP) (Shutova, 2015). According to Lakoff and Johnson's (1980) Conceptual Metaphor Theory (CMT), conceptual metaphors are cognitive structures that enable reasoning about abstract target domains (e.g., IDEAS) through concrete source domains (e.g., FOOD), shown as metaphoric language (e.g., "This plan is half-baked"). Although NLP research on metaphor has advanced considerably, current approaches remain limited in one crucial aspect: They treat metaphors as context-independent lexical or conceptual units. These static NLP methods are in line with the static linguistic approach to metaphor, represented by Pragglejaz Group's (2007) Metaphor Identification Procedure that detects metaphorical usage word for word. Such static views overlook how metaphors evolve over time.

Going beyond static metaphor theories, Tan (2023) and Tan and Cienki (2024, in press) introduce the Yin-Yang Dynamics of Metaphoricity (YYDM), which describes how metaphors shift between inactive and active (viz. between Yin and Yang) states in discourse. This model incorporates rhetorical, cognitive, and political factors to explain how metaphoricity (the property of being a metaphor) in political discourse changes over time. It conceptualises metaphoricity as a continuum modulated by *metaphoricity cues* such as repetitions. For instance, Donald Trump used the verb 'lose' in Example 1, which frames trade as a COMPETITION. The metaphoricity of both the metaphorical word ('lose') and its source domain (COMPETITION) remains static here, as 'lose' appears only once without further elaboration. However, in a later speech (see Example 2), the repeated use of 'lose' in different verb forms functioned as a *metaphoricity cue* that activates the metaphoricity—makes the metaphoricity more salient. This shift illustrates the discourse-level dynamics captured by YYDM but overlooked by static approaches. Expanding on this framework, Tan and Cienki (2024) analyse dynamic metaphors about trade in government discourse during the Trump and Xi presidencies, showing how metaphoricity is not fixed, but develops with political sentiments and changes with political attitudes.

(1) We can't *lose* almost $800 billion on the start of the trade dispute, like has been done for many years.

(2) We were *losing* all our cases in the World Trade Organization. Almost every case, we were *lost*, *lost*, *lost*.

Although the YYDM model offers a robust theory for understanding metaphor activation, its application has been limited to manual discourse analysis. Tan and Bloem (2024) call for this gap to be bridged by integrating insights from the YYDM model into computational approaches. They argue that metaphoricity cues, which serve as observable linguistic markers of metaphor activation, can form the basis for computational detection of dynamic metaphors. To respond to this call, we

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4547–4559
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

evaluate the performance of popular instruction-tuned autoregressive decoder models, specifically the GPT O-series models (O3, O4-mini, O4-mini-high) in detecting five of these metaphoricity cues at the phrase level. We also perform a small follow-up study with open models, including NVIDIA's recently released Nemotron 3 Super.

Our dataset is Tan and Cienki's (2023) Trade Metaphor Corpus developed at Vrije University Amsterdam (VUTMC), which is annotated for such cues. Our experiments rely on theory-informed prompts, as zero-shot or simple one-shot prompts lead to insufficient performance. To validate the prompts and models for metaphor detection, we first compare these O-series models with leading metaphor detection models (MelBERT, MisNet, AdMul) on two benchmark datasets (MOH-X and TroFi), then evaluate them on the VUTMC dataset. By doing so, we seek to operationalise these metaphoricity cues, thereby taking the first step towards theory-informed automatic identification of dynamic metaphors.

2. Related Work

2.1. Computational Metaphor Detection

MelBERT and MisNet are two specialised metaphor detection models. MelBERT uses contextualised embeddings from pre-trained transformers within a Siamese architecture to compare target words and their sentential contexts (Choi et al., 2021). It integrates two theoretical frameworks: The Metaphor Identification Procedure (MIP), distinguishing basic and contextual meaning, and Selectional Preference Violation (SPV), which links the metaphor to semantic incongruity. MisNet builds on the same MIP–SPV foundation, but improves the representation of basic meaning by retrieving dictionary examples and encoding richer word–context interactions (Zhang and Liu, 2022). However, both models focus only on token-level metaphoric patterns.

Zhang and Liu (2023) propose an alternative approach to metaphor detection, introducing AdMul: an adversarial multi-task learning framework. AdMul uses task transfer to improve metaphor detection by aligning it with a newly defined auxiliary task called Basic Sense Discrimination (BSD). The BSD model is designed to distinguish between the literal and metaphorical senses of a word. Notably, it outperforms previous models in zero-shot transfer settings, unlike the prompt-based approach used in the present study (see Section 1).

Some computational frameworks do model conceptual mappings and source domains, in line with aspects of the YYDM model. Ge et al. (2022) propose the first metaphor identification model that is designed to better explain underlying conceptual mappings, based on CMT. This model detects metaphoricity at the word-pair level and infers source-target domain mappings by linking words to WordNet hypernyms. The ability to generate interpretable source–target concept pairs is an advantage of this approach, providing a clear explanation of why a particular expression is metaphorical. Building on Ge et al.'s (2022) framework, Mao et al. (2023) proposes MetaPro, which integrates metaphor identification, literal paraphrasing, and conceptual mapping into an end-to-end processing system. Unlike earlier tools, it offers a web-based interface and supports open-domain English input, making it accessible to non-programmers and useful for computational and cognitive linguistic applications (MetaPro Team, 2025). However, its scope is still limited to isolated adjective–noun and verb–noun pairings.

Wachowiak and Gromann (2023) examine the capacity of generative LLMs in interpreting metaphors by investigating GPT-3's ability to identify the source domains of conceptual metaphors. Unlike traditional approaches based on pre-defined source–target inventories, this method employs few-shot prompting and fine-tuning to predict the source domain from a given sentence and a known target domain. This approach enables the model to produce mappings that are not limited by pre-existing metaphor databases. Although the model achieves an accuracy of over 81% on Lakoff's (1994) Master Metaphor List, it exhibits significant limitations. These include assigning incorrect source domains (hallucinations) and misclassifying metaphorical expressions as literal.

Tian et al. (2024) introduce a theory-guided scaffolding instruction (TSI) framework, which aims to enhance LLMs' metaphorical reasoning capabilities by explicitly incorporating metaphor theory into the prompting process. The TSL framework incorporates well-established theories, such as MIP, SPV and the CMT, to construct a metaphor knowledge graph and a hierarchy of scaffolding questions. Experimental results on the MOH-X and TroFi benchmark datasets demonstrate that the TSI approach, especially when the CMT is employed, outperformed existing LLM-based methods and state-of-the-art models such as MelBERT and AdMul. Subsequently, Lin et al. (2025) introduced a three-part prompting strategy incorporating dynamically generated examples from MelBERT, dictionary definitions as well as self-judgement. This approach achieved state-of-the-art performance with GPT-4o on MOH-X and TroFi.

Although recent models such as AdMul and TSI have made significant advances in detecting and paraphrasing metaphors, they conceptualise metaphors as static linguistic or conceptual units, without modelling how their activation (viz., salience

of metaphoricity) shifts in discourse. This neglects an important aspect of metaphor use: its dynamic nature in real-world communication. Currently, there is no computational framework that targets discourse-level metaphoricity cues. We address this gap by evaluating state-of-the-art reasoning models on their ability to detect metaphoricity cues in political text. This will contribute to developing models that account for dynamic metaphor.

2.2. Dynamic metaphor theories

Cameron (1999) was among the first to reject the static classification of metaphors (metaphorical vs. non-metaphorical). She emphasises the temporal dimensions of metaphor use, shifting the focus from isolated lexical units to evolving discourse elements. She argues that, in conversation, metaphors are context-dependent and interactive, often developing across multiple turns of talk. Drawing on classroom data, she shows how speakers collaborate to adapt metaphorical language in real time. Cameron (2007b) further examines metaphor use in reconciliation dialogues between an IRA bomber and the daughter of his victim. She shows that metaphors emerge and develop throughout the interaction, which help shape social relationships and identities of discourse participants. Gibbs and Cameron (2007) further develop the dynamic view by suggesting a dynamic systems approach to metaphor. They conceptualise metaphors as emergent products of simultaneous interactions of cognitive, emotional, social and cultural processes. By analysing the same reconciliation talks, they demonstrate that metaphorical expressions don't merely reflect cognitive mappings but also serve as real-time adaptive responses to interpersonal dynamics. Cameron (2010) further introduces the Discourse Dynamics Approach (DDA), which offers a systematic way of studying metaphors in social interactions, particularly how they emerge in linguistic, cognitive, emotional, and cultural contexts. The DDA has primarily been used in metaphor-led discourse analysis to trace how recurring metaphoric patterns (viz., systematic metaphors) reflect the thoughts of discourse participants. One of DDA's tenets is metaphor clustering: Metaphors tend to appear in close proximity during moments of heightened interpersonal engagement (Cameron, 2007a).

Developing Cameron's dynamic view, Müller (2008) argues that metaphors exist on a continuum of activation, shifting between 'sleepy' and 'awake' states depending on discourse contexts. She introduces metaphoricity cues, including verbal, pictorial and gestural signals that make a metaphor more salient in interactions. Building on Müller's perspectives, Tan (2023) and Tan and Cienki (2024, in press) propose the YYDM model that accounts for deeper underlying mechanisms of verbal metaphor activation, including changes in source and/or target domains. She argues that changes in the salience of metahoricity can be observed through metaphoricity cues in texts. Despite all these theoretical advancements, dynamic metaphor research still relies on manual annotations. Computational applications in this field have yet to be seen.

3. Research Aims and Questions

We aim to assess the capabilities of LLMs in detecting five metaphoricity cues of the YYDM model at the phrase level, specifically *explicit mapping*, *repetition*, *marking*, *emphasis* and *novelty*, in Tan and Cienki's (2023) VUTMC corpus. We particularly focus on recent GPT O-series models, i.e., O3, O4-mini, and O4-mini-high, due to OpenAI's (2025) significant progress in large-scale reasoning. These models ranked among the top performers on major benchmarks of logical and symbolic reasoning at the time our study was carried out (Vellum, 2025), yet current research provides little insight into their capability for linguistic reasoning. Our study thus contributes to filling this research gap, as metaphoricity cue detection requires a great deal of linguistic and contextual reasoning of LLMs.

We investigate the following question:

Main RQ: How well do GPT O-series models (GPT O3, GPT O4-mini, and GPT O4-mini-high) detect metaphoricity cues at the phrase level in the VUTMC corpus?

To contextualise these results, we also need to investigate two baseline conditions:

Sub-RQ 1: How well do GPT O-series models (O3, O4-mini, O4-mini-high) detect metaphors at the phrase-level compared to the leading models MelBERT, MisNet, and AdMul, on standard benchmarks?

Sub-RQ 2: How well do GPT O-series models (O3, O4-mini, and O4-mini-high) detect metaphors at the phrase-level in the VUTMC corpus, which is annotated for dynamic metaphor cues?

Our prompts, code, and tested dataset can be accessed here.[1]

4. Methodology

4.1. Dataset

Tan and Cienki's (2023) VUTMC dataset contains various types of American governmental texts about trade between the U.S. and China, including statements, press briefings, presidential directives, fact sheets, party manifestos, congressional debates, and news of the White House. The choice for

[1] github.com/berkay5366/Dynamic-Metaphoricity-Cues

the VUTMC corpus is due to the frequent use of metaphors in U.S. political discourse for rhetorical purposes (Mio et al., 2005; Pilyarchuk and Onysko, 2018). The version of the VUTMC dataset used in this study can also be found in the aforementioned GitHub supplementary materials for this article.

To reduce ideological bias from metaphoric framing (Tan et al., 2024), the VUTMC dataset includes American discourses from presidencies across two political orientations. Specifically, it contains a Clinton corpus (573,569 tokens) and a Trump corpus (1,470,409 tokens) of materials released in the eras of Bill Clinton (1993-1997) and Donald Trump (2017-2021), respectively. This ensures a balanced representation of metaphor use across the ideological spectrum of American politics, thereby reducing the likelihood that observed metaphoric patterns are driven solely by partisan rhetoric.

The manual annotation of the dataset follows three main steps. First, all trade metaphors in concordance sentences that have trade-related target words (e.g., trade; trade negotiation) were annotated. Second, neighbouring sentences that share the same source and/or target domains with the concordance sentences were manually grouped into sentence clusters. Third, for each sentence cluster, the metaphoricity cues were coded by two coders based on Tan and Cienki's (2024) coding scheme, yielding 'substantial' intercoder agreement (Cohen's Kappa = .73)

4.2. Metaphoricity Detection prompt

We incorporated a binary decision format into the prompt design for metaphor detection. Each sentence was evaluated as either metaphorical or non-metaphorical, following the three categories of the VisMet Project (2025): Indirect, implicit, and direct. A sentence that satisfied at least one of these criteria was classified as metaphorical.

4.3. Cue Detection Prompt

4.3.1. Prompt Design for Repetition Detection

Defintion of Repetition Repetition involves the repeated use of a metaphorical expression within a single text segment, which reinforces the force of a given linguistic metaphor and its underlying conceptual metaphor. Repetition functions as both a rhetorical intensifier and a cue that increases the metaphoricity of a given linguistic metaphor and its source domain. For example, in the following sentence by Donald Trump: "For decades, I have raised the alarm over unfair foreign trade practices that have *robbed* communities of their wealth and *robbed* our people of their ability to provide for their families", the same metaphorical verb 'robbed' is repeated. This repetition reinforces its metaphority and the conceptual metaphor UNFAIR TRADE PRACTICE IS THEFT, whereby trade practices are personified as criminals.

Scope of Repetition Detection Our prompt design for detecting repetition as a metaphoricity cue focuses on the recurrence of verbs. Although repetition may occur in various grammatical categories, such as nouns or adjectives, we focus only on repetition of verbs, as this constitutes the majority of instances. This includes both identical surface forms (e.g., "robbed ... robbed") and inflectional variants derived from the same stem (e.g., 'rob'; 'robbing'). Nominalisations of the same lexical root (e.g., 'sing'; 'song') are also included. By focusing on verbs, the prompts avoid overloading the model with unrelated forms of repetition, in particular repeated nouns in rhetorical language (e.g., "year after year, decade after decade, trade deficit upon trade deficit") that do not function metaphorically.

4.3.2. Prompting for Novelisation Detection

Definition of Novelisation Novelisation involves introducing a novel metaphorical expression that maps a familiar target domain onto an unusual or unexpected source domain within a given discourse context. Unlike repetition, which reinforces an existing linguistic metaphor and conceptual metaphor through lexical recurrence, novelisation alters the original linguistic metaphor and conceptual mapping to shift the metaphorical framing. A metaphorical expression or source domain is considered novel if it is a rare metaphorical expression or a source domain not typically associated with the target domain of trade in political discourse. For example, while metaphorical words or source domains such as 'war' (in 'trade war') and 'tools' (in 'trade negotiation tools') are commonly used in political texts, other metaphorical words or source domains may be considered novel. For instance, in the sentence "This trade deal is a *fly* on China's face while a full-grown great white shark is eating America's assets", 'fly' is used as an novel metaphorical word and a new source domain to frame the trade deal.

Scope of Novelisation Detection To detect novelisation, we assess whether the metaphorical word is used in a way that differs from its concrete use (i.e., primary meaning). A metaphorical word is classified as novel if (1) its primary meaning is the concrete, literal sense, and (2) its intended contextual meaning is rare and creatively extends its primary meaning. For instance, 'pie' in the expression 'trade pie' is a novelisation because its intended meaning is not identical to its primary sense of a baked dish.

4.3.3. Prompt Design for Marking Detection

Definition of Marking Marking refers to the presence of lexical or grammatical elements that indicate a metaphorical use. Examples include comparative structures such as 'like' or 'as', constructions involving verbs such as 'becomes' or 'represents', and phrases such as 'so to speak' or 'to use an analogy'. While these markers are not metaphorical themselves, they highlight the metaphorical word or mapping being used. An illustrative case from the Trump corpus is the sentence: "Under the leadership of Secretaries Ross, Mnuchin, and their Chinese counterparts, the United States has negotiated intensively to reach consensus in *areas including* agricultural trade, financial services, investment, and energy." Here, the noun 'areas' metaphorically represents abstract policy topics as spatial domains. The grammatical construction 'areas including X, Y, Z' reinforces both the metaphoricity of 'areas' and the metaphorical conceptualisation of agricultural trade as bounded spaces. While the metaphoricity itself arises from the choice of the word 'areas', the syntactic construction and the marker 'including' contribute to its recognisability.

Scope of Marking Detection To detect the marking cue, the model was instructed to identify the lexical or grammatical elements that indicates a metaphorical expression, known as tuning devices (Cameron and Deignan, 2003). The list of marking devices was compiled from Cameron and Deignan's (2003) and Tan's (2023) inventories, including 'like', 'as', 'might be', 'could be', 'so to speak', 'if you like', 'sort of', 'almost' and structural cues such as colons or dashes. To prevent overgeneration, colons were only considered valid if they were immediately followed by a noun functioning as a metaphorical source domain. Importantly, this list was not exhaustive. Our prompt allowed for the recognition of additional lexical items as marking devices, provided they fulfilled a similar marking function. This enabled the model to generalise from the given examples and detect novel marking devices that perform the same role.

4.3.4. Prompt Design for Emphasis Detection

Definition of Emphasis Emphasis is a highlighting of metaphorical words or mappings using syntactic or typographic devices to make them stand out more in the discourse. Unlike novelisation, which introduces new metaphors, emphasis increases the perceptual salience of a metaphor that is already present in discourse. Common mechanisms include the use of quotation marks, italics, capitalisation, contrastive structures such as 'not... but...', and lexical markers such as 'emphasise', 'do/does' or 'it is that'. An illustrative example appears in the Clinton corpus: "*It is not* a question of sneakers or slave labor products or supporting increased trade. *It is* a question of a dirty little secret that we are ignoring today, and that is that China is helping to build and distribute nuclear weapons around the world, and we all know it." The contrastive structure 'It is not... It is...' functions as a syntactic marker of emphasis. This sentence uses contrast to make one idea seem unimportant and emphasise the urgency of another. The speaker asserts that mundane trade matters are not the focus. Instead, the focus is on a much bigger danger: nuclear weapons. This technique causes the listener to pay particular attention to what needs to be dismissed (a trade issue) and what needs to be emphasised ('a dirty little secret').

Scope of Emphasis Detection Unlike the open-ended list of devices used for marking, the list of emphasis indicators, based on Tan (2023), was closed and finite. The model was prompted to classify a sentence as containing emphasis only if it used one or more of the following: quotation marks (single or double), 'not... but...', 'It is X that...', 'do', 'does', 'only', the verb 'emphasise' and its inflectional variants, and the capitalisation or typographic emphasis (e.g. italics) of metaphorical terms.

4.3.5. Prompt for Explicit Mapping Detection

Definition of Explicit Mapping Explicit mapping involves the clear expression of a metaphorical relationship, where the source domain is expressed in the text and the relationship between the source and target domains is clear. Unlike metaphors that rely on implication (e.g., the verb 'robbing' implying a thief), explicit mappings directly signal the underlying conceptual structures, through the use of metaphorical words drawn from their source domains. The following example by Bill Clinton illustrates this clearly: "Ending trade with China is a too sweeping and disruptive *remedy* for the president's purpose." The metaphorical word 'remedy' draws from the medical domain, directly framing trade intervention ('ending trade with China') as a form of treatment. Such mappings are characterised by their limited interpretive ambiguity, making them especially effective in framing.

Scope of Explicit Mapping Detection To detect an explicit mapping as a metaphoricity cue, the model was instructed to identify whether a noun was present in the same sentence as the trade-related keyword to which it could be compared. Unlike implicit mappings, which require inferential reasoning to identify the source domain, explicit mappings directly name the source domain in relation to the target domain related to trade. The

noun could represent either a concrete or an abstract concept. Metaphorical verbs and adjectives are excluded. By limiting the definition to overt, noun-based metaphorical mappings or words, the prompt ensures a clear distinction between explicit and implicit conceptual metaphors.

4.4. Experimental Setup

4.4.1. Detecting Metaphors

To validate the applicability of our metaphor detection prompt, we tested its performance on two established benchmark datasets: MOH-X (Mohammad et al., 2016) and TroFi (Birke and Sarkar, 2005). Tian et al. (2024) tested MelBERT, MisNet and AdMul, using 300 instances from each dataset. Following their methodology, we tested GPT O-series models (O3, O4-mini, O4-mini-high) on samples of the same size using our prompt and compared our results to their reported scores. Each 300-sentence sample was divided into six sets (50 sentences each) to meet OpenAI's session limits. The model's performance in a binary classification setup (metaphorical versus literal) was measured using F1 and accuracy scores.

Four constraints were introduced to test the reasoning capabilities of O-series models. First, although MOH-X and TroFi contain only metaphorical verbs, our prompt did not specify that only verbs were relevant. Second, although MOH-X and TroFi include only indirect metaphors, we defined all three metaphor types (direct, indirect, implicit) based on VisMet. The model must infer that only indirect metaphors are relevant. Third, unlike MOH-X and TroFi with ordered instances, we randomised sentence orders with a fixed seed (49). Finally, unlike the state-of-the-art token-level setups, we did not highlight the target verb for the model. These four constraints required the model to infer metaphoricity from the entire sentence, rather than relying on token-level cues or patterns in the datasets. We then evaluated GPT O-series models (O3, O4-mini, O4-mini-high) on the VUTMC dataset using the same prompt. A sample of 300 sentences, balanced between metaphorical and literal instances, was selected and ordered by the same random seed (49), to ensure comparability with the MOH-X and TroFi benchmark evaluations.

4.4.2. Detecting Metaphoricity Cues

As most metaphoricity cues except repetition did not yield a sufficient number of unique instances to support supervised learning, no fine-tuning or pre-training was performed. Instead, classification was conducted for each metaphoricity cue using all available positive examples (true labels) paired with an equal number of automatically selected negative examples (false labels). This balanced design ensures that the resulting accuracy and F1 scores provide meaningful evaluations, despite negative instances vastly outnumbering positive ones in natural language. An exception was made for explicit mapping. This is because after extensive prompting and evaluation, several instances that had initially been assumed to be explicit mappings had been misclassified. Consequently, the classification for this cue was based on 129 explicit and 72 implicit instances. Table 1 provides an overview of the total number of tested instances for each metaphoricity cue and the total number of tested prompts for each cue.

Metaphoricity Cue	Instances	Prompts
Explicit Mapping	201	4
Novelisation	86	2
Marking	90	2
Emphasis	20	1
Repetition	300	6

Table 1: Distribution of tested instances and prompts for each metaphoricity cue.

For this experiment, we use the same GPT O-series models that we validate on the metaphor detection task. In addition, we perform a follow-up experiment using several more recent and open models to estimate whether our results can generalize to such models. For this follow-up, we only test one out of the five cues, the repetition cue (which has the largest number of instances).

We test three open models that are advertised as having reasoning capabilities in this follow-up. The first is Nemotron 3 Super (NVIDIA, 2026), a 120B parameter open-weight MoE model with 12B active parameters released by NVidia in March 2026. The second is MiniMax-M2.5 (MiniMax, 2026), a 230B parameter open-weight MoE model with 10B active parameters released by the Chinese company MiniMax. The third is Big Pickle (OpenCode, 2026), an experimental 'stealth' model that's not confirmed to be open-weight, but it is currently available with free API access on the OpenCode Zen platform with minimal documentation.

5. Results

5.1. Results of Sub-Research Question 1

Table 2 shows, for both MOH-X and TroFi datasets, the F1 and accuracy scores obtained by the O-series models are higher than the best results achieved by any of the State-of-the-Art (SotA) baselines. This result is close to the state-of-the-art per-

Method	MOH-X		TroFi	
	F1	Acc.	F1	Acc.
SotA metaphor detection models				
MelBERT	77.88	77.89	62.36	62.89
MisNet	77.08	77.11	62.01	62.67
AdMul	79.74	79.89	60.54	62.67
OpenAI O-series models				
O4-mini	85.21	86.00	68.09	65.00
O4-mini-high	82.13	84.33	69.62	**65.67**
O3	**90.26**	**90.00**	**71.66**	64.67

Table 2: Performance comparison of SotA metaphor detection models and GPT O-series models on the MOH-X and TroFi datasets. Best F1 and accuracy scores are **bold** and second–best are underlined in each column.

formance recorded by Lin et al. (2025), who achieve slightly higher scores with GPT-4o using a far more complex prompting strategy. Specifically, they observe a 91.71 F1-score on MOH-X and 73.73 on TroFi. The consistent outperformance of the O-series models across both datasets, together with our prompt, shows that they are methodologically suitable for analysing the VUTMC dataset.

5.2. Results of Sub-Research Question 2

As shown in Table 3, the F1 and accuracy scores of O4-mini, along with the F1 score of O4-mini-high, are the lowest in all three tested datasets (MOH-X in Table 2, TroFi in Table 2 and the VUTMC corpus). This suggests that our political corpus poses a greater challenge for metaphor detection. This raises a methodological concern: If the model fails to recognise that a sentence is metaphorical, it may not attempt to identify metaphoricity cues within that sentence and will instead process the metaphoric expressions as literal. To mitigate this risk, we added a standard disclaimer to each prompt in tasks of metaphoricity cue classification, explicitly stating that the sentence in question is metaphorical and that the only relevant judgement is the presence or absence of metaphoricity cues. This step was necessary to ensure the accuracy of metaphoricity cue detection, because it prevents misclassifications caused by failing to detect a sentence's metaphorical status.

5.3. Results of Main Research Question

Table 4 shows that O3 consistently achieves the highest scores for most metaphorical cues. Of the five cues, repetition and emphasis yield the highest F1 and accuracy scores, though the emphasis cue is infrequent (see Table 1 for cue frequencies). In contrast, novelisation and marking prove to be

Model	F1	Accuracy
O4-mini	51.79	63.76
O4-mini-high	56.52	66.44
O3	**69.94**	**67.11**

Table 3: Performance of GPT O-series models in metaphor detection on the VUTMC corpus. Best F1 and accuracy scores are **bold** and second–best are underlined in each column.

Cue	Model	F1	Accuracy
Explicit Mapping	O4-mini	78.63	72.14
	O4-mini-high	**80.44**	**73.63**
	O3	**80.44**	**73.63**
Novelisation	O4-mini	55.42	56.98
	O4-mini-high	**64.62**	**73.26**
	O3	59.09	58.14
Marking	O4-mini	70.59	**72.22**
	O4-mini-high	69.39	66.67
	O3	**73.79**	70.00
Emphasis	O4-mini	81.82	**80.00**
	O4-mini-high	**83.33**	**80.00**
	O3	81.82	**80.00**
Repetition	O4-mini	82.05	83.67
	O4-mini-high	83.56	84.00
	O3	**84.21**	**85.00**

Table 4: Performance of GPT O-series models in metaphoricity cue detection on the VUTMC corpus. Best F1 and accuracy scores are **bold** and second–best are underlined in each column.

the most challenging. The lowest F1 and accuracy scores for novelisation support recent NLP findings (e.g., Reimann and Scheffler, 2024; Shao et al., 2024) that detecting novel metaphors remains highly difficult.

5.4. Extension to open models

To assess whether our prompts and the observed performance on the task extend beyond the closed GPT O-series models, we also evaluated repetition detection using three open models as shown in Table 5. Repetition was selected for this comparison because it was the most frequent metaphoricity cue in the corpus. The results show substantial variation between models. Big Pickle and MiniMax-M2.5 perform clearly below the O-series models. By contrast, the results of the open-weight Nemotron 3 Super are broadly comparable to those of the O-series models on the same task.

These findings show that there is potential for open reasoning models to perform metaphoricity cue detection as well, even without further prompt engineering. However, this performance is model-

Model	F1	Accuracy
MiniMax-M2.5	61.95	71.33
Nemotron 3 Super	**81.88**	**82.00**
Big Pickle	<u>66.11</u>	<u>73.00</u>

Table 5: Performance of open LLMs in repetition cue detection on the VUTMC corpus. Best F1 and accuracy scores are **bold** and second–best are <u>underlined</u> in each column.

dependent and doesn't clearly correlate with model size (MiniMax-M2.5 is bigger than Nemotron 3 Super). Further prompt engineering would be required to determine whether these performance gaps are due to the reasoning capacities of each model or due to differences in optimal prompting strategies between these models.

While this follow-up experiment shows that open models have potential for the task, a more comprehensive evaluation on additional cues and models, with model-specific prompt engineering is therefore needed before broader claims about generalisability can be made.

6. Discussion

6.1. Explicit Mapping

In the case of explicit mapping, GPT models exhibited mixed reliability. Although the prompts clarified that all sentences contained a metaphor and that the target domain was always related to trade, it did not explicitly instruct the model to recognise multi-word expressions, such as 'trade negotiation' or 'trade conflict', as compound keywords. This created ambiguity in several cases. For example, when presented with the phrase 'trade war' (in sentences such as "trade war is a self-inflicted wound"), some chains of reasoning incorrectly classified the phrase itself as a conceptual metaphor with 'war' as the source domain rather than recognising 'trade war' as the target domain. This suggests that the model sometimes conflates metaphorical vehicles with target concepts when they appear as compound nouns, leading to erroneous classifications.

6.2. Emphasis

The classification of emphasis was particularly sensitive to the formatting irregularities that are typical of transcripts of political texts. Because the VUTMC corpus consists largely of transcribed speech notes and fragments of official statements, many sentences included extensive use of capitalisation, irregular punctuation, and layout features that mimicked structures of emphasis without serving the function of highlighting metaphors. This posed a challenge for all models. For instance, the sentence "CONFRONTING UNFAIR TRADE: President Trump is standing up for the American people and putting an end to decades of bad deals." was incorrectly identified as containing emphasis by all three models, even though the capitalisation merely served as a stylised article headline rather than marking a metaphorical highlight. In contrast, GPT O4-mini correctly identified another ALL-CAPS sentence as 'government-boilerplate legal text', thereby avoiding misclassification. Conversely, in a sentence beginning with 'THE PRESIDENT:', both O4-mini and O3 accurately identified the capitalised speaker tag as not emphasised, but failed to recognise the genuine emphasis in 'do take advantage' that highlights personification. Nevertheless, emphasis is still the second-best detected metaphoricity cue, though it only occurred 20 times in the test set so this F1-score has limited reliability.

6.3. Marking

Identifying markers is challenging due to the open-ended nature of the metadiscursive and grammatical indicators associated. Although a comprehensive list of commonly cited markers (e.g., 'literally', 'like', 'to use an analogy') derived from Cameron and Deignan (2003) and Tan (2023) formed the basis of the detection prompt, the models were also instructed to identify structurally and functionally similar cues using generalisable rules. However, all three models incorrectly identified the phrase 'as I see it' as a marking device when it appears in the sentence "As I see it, the first year of the Bush administration, we had about a $3 million trade deficit with China." This phrase functions as a metadiscourse marker for subjectivity, rather than indicating metaphorical language or mapping, thereby violating functional constraints of the marking cue. In contrast, 'literally', which is also a metadiscourse marker and was included in the list of markers, was correctly identified as a cue. Additionally, general rules had to be introduced to account for hedging expressions such as 'merely', as in 'merely a sponge' or 'merely an invitation', which often function analogously to listed cues, but which are not explicitly enumerated in the original list of markers. However, this dual strategy of list- and rule-based identification produced two conflicting outcomes: (1) high false positives, where formally similar but non-metaphorical expressions were flagged, and (2) high false negatives, where non-listed but semantically valid markers were overlooked. This illustrates that marking is one of the most error-prone cues of metaphoricity in the dataset. This aligns with the findings of Guan et al. (2025), who report poor performance of generative LLMs on metadiscourse marker detection.

6.4. Repetition

To avoid false positives caused by the frequent recurrence of the keyword 'trade', the detection of repetition was primarily constrained to verb repetition. As a secondary rule, derivational shifts between verb and noun forms are permitted to be recognised as valid repetitions. All three models correctly identified 'abuse' and 'abused' as valid instances of metaphorical reinforcement of verb repetition. However, GPT O3 demonstrates a tendency towards overgeneralisation by interpreting instances such as 'trade deal' or 'barriers' as cases of verb-based repetition, erroneously assuming that 'deal' (noun) derives from the verb 'to deal', or that 'barriers' relates to 'to bar'. These false positives arise from excessive inferential reasoning that exceeds the intended syntactic constraint. Paradoxically, however, this same reflective reasoning enables GPT O3 to outperform O4-mini and O4-mini-high in identifying true verb repetitions, which the latter occasionally miss due to their shorter processing chains.

6.5. Novelisation

The novelisation prompt was designed to identify novel metaphorical words whose primary meanings are not their metaphorical meanings but their basic/concrete meanings. However, this criterion excludes an important subclass of novelisation: metaphorical words with a familiar source domain mapped to an unfamiliar target domain. Certain source domains are rarely used to frame trade, but are often used to frame other topics. Because GPT models tended to interpret such metaphorical words in trade discourse as conventional, we excluded this secondary condition from the prompt design. Including both semantic mismatch and frame novelty as criteria would have led the model to overclassify nearly all metaphorical instances as conventional, which would have severely compromised the F1 scores. However, our prompt under the primary condition still results in classification errors. For example, metaphorical expressions using common source domains (e.g., disaster) to frame trade may still be misclassified as novelisation, if their metaphoric/contextual meanings deviates from their primary/basic meanings. For example, the sentence ¨Our trade has been a disaster" is misclassified as a novelisation by two models. Here the primary meaning of 'disaster' is a catastrophic event, whereas its metaphoric meaning (economic failure) draws on a conventional source domain 'disaster' to frame trade.

7. Conclusion

We have shown that OpenAI's (2025) O-series models outperform several state-of-the-art metaphor detection systems on benchmark datasets without complex prompting strategies. Furthermore, these models demonstrate solid accuracy and F1 scores in the novel task we introduce, the detection of metaphoricity cues. Importantly, the models demonstrate strong performance at the phrase level without requiring an explicitly highlighted metaphorical vehicle during inference, unlike most previous metaphor detection systems. However, these results for metaphoricity cues depend on specifically designed, theory-informed prompts. Our follow-up also suggests that open-weight mixture-of-expert models can also perform this novel cue task task with good accuracy, though more comprehensive benchmarking is needed.

Zero-shot or minimally designed one-shot prompts tend to produce highly conservative model behaviour, often leaving metaphors and metaphoricity cues undetected. As with fine-tuning in supervised models, prompt engineering is essential for maximising LLM performance. This process requires expertise in linguistic theory to construct prompts that accurately reflect the complexity of metaphors and the metaphoricity cues that signalise them. Although we focus on the trade discourse, the same metaphoricity cues can be operationalized and generalized to other types of political genres and political discourses.

Despite its exploratory nature, this study serves as as a "first step" toward computational modeling of dynamic metaphors. Future research directions should include evaluation of a broader range of models, particularly open-weight models. The use of open-weight models is more reproducible and more interpretable, and open-weight models are increasingly matching the performance of commercial models at other reasoning-related tasks. Such extension would involve further prompt engineering, as optimal prompting approaches are often model-specific (Sclar et al., 2024). Another direction would be to extend the work to the Chinese part of the VUTMC dataset for cross-linguistic validation of the method. Lastly, our limited operationalization of dynamic metaphor theory should be extended across larger spans of discourse than what was done in the present study. Annotations for this are already present in the dataset we used.

Our computational findings can help to advance empirical research of dynamic metaphors. A natural continuation of this work would be to combine cue detection with the metaphor detection task, supporting future human-in-the-loop dynamic metaphor annotation efforts. A guided reasoning prompt that incorporates detection of both metaphors and metaphoricity cues may lead to improved metaphor detection, as in theory, the cues also point to the presence of linguistic metaphor.

8. Limitations

Our experiments do not cover all possible dynamic metaphor cues. Other more frequent cues discussed by Tan and Bloem (2024) will have to be investigated in future work. Specifically, these are the cues of composing (source domain clustering), diversification, multivalency, and domain mixing. Furthermore, our experiments only test the activation of the five cues within sentence boundaries, whereas the dynamic metaphor model (YYDM) argues that metaphoricity is dynamically activated across groups of sentences and discourses, so our operationalization of the theory is a partial one. Given that our annotated dataset extends beyond the sentence level, future research should test LLMs' performance on detecting metaphoricity cues across sentences, which is likely to be a more difficult version of the task. In doing so, it can better computationally model how metaphor cues accumulate and interact across sentences to shape discourse-level dynamic metaphoricity.

In terms of autoregressive reasoning decoder models, our study is focused on GPT O-series models, as these are widely known and easily accessible to discourse researchers who are less familiar with LLMs. While we did perform a follow-up evaluation on additional open-weight models, this evaluation was not comprehensive. Furthermore, we did not take advantage of the opportunity to perform interpretability research for this task, which is only possible with open-weight models. Evaluation of other reasoning models, particularly open-weight ones, on the metaphoricity cue detection task, might yield more accessible and more interpretable results, e.g. by applying mechanistic interpretability techniques to discover which features drive succesful cue detection. Smaller open-source models require fewer resources and are more environmentally friendly, although our task does not involve generating large amounts of tokens. Using our dataset and code, other researchers are very welcome to benchmark other instruction-tuned LLMs as well, particularly open-weight ones.

The only available corpus annotated for metaphoricity cues is limited to the domain of political texts of international trade. Therefore, we were not able to determine whether our findings can generalize to other discourse domains. At the same time, the prompt instructions for all five metaphoricity cues are designed to be domain-independent. In the study, trade references serve mainly to restrict the model's attention to the relevant target domain, rather than to define the cues themselves. In principle, applying the prompts to another domain would therefore require only replacing trade with the relevant target-domain label, rather than redesigning the cue definitions. Several political genres (e.g., debates, news, speeches, interviews) were covered in the tested VUTMC dataset. However, possible genre-specific differences (Tan et al., 2026) in how these cues are realised nevertheless mean that such transferability still needs to be tested empirically. If this transferability can be confirmed, our approach may aid in the semi-automatic annotation of corpora from other discourse domains.

A methodological limitation is that we were only able to perform the cue detection task with sentences that contain a metaphor. The cue detection approach seems not to be robust to overgeneralization. However, although we restrict the cue detection to pre-identified metaphorical sentences, it does not mean that our method cannot apply to non-annotated metaphorical sentences. In naturalistic real-world settings without any available annotation, the metaphor detection task can be combined with metaphoricity cue detection in a two-stage pipeline, reducing overgeneralization in the second step. Alternatively, novel theoretically informed prompting strategies can be devised that interconnect both tasks in a guided reasoning process, when working with models that have more advanced reasoning simulation capabilities than OpenAI o3.

9. Ethics Statement

Although we tried to reduce the ideological bias of the corpus by using data related to two ideologically different presidents (Democratic vs. Republican), other ideological stances are not represented. Moreover, the VUTMC corpus used in this study is limited to U.S. political texts on U.S.-China trade wars. As a result, the metaphor use and metaphoricity cues represented in the data reflect a narrow political and cultural context.

Given that the data is ideologically charged, there is a possibility that ideological bias of annotators played a role, though Tan and Cienki's (2024) coding scheme is based on MIPVU which was designed to reduce subjectivity. The main sources of subjectivity are in deciding which meanings of a word are basic meanings (and which dictionary to consult in that process), and deciding whether a meaning in context is sufficiently different from a word's basic meaning. These processes could be affected by annotator bias in some cases, though these are more linguistic questions than cultural or ideological questions. Ideological bias related to this procedure has not been noted in previous work.

The use of LLMs also raises ethical concerns about interpretability and environmental cost. Because model decisions are not fully transparent, it is not always possible to explain why a cue was or was not detected. In addition, repeated prompting

increases the use of computational resources. A more resource-efficient setup would group multiple instances within a single prompt, thereby reducing the total number of prompts, although this may involve a trade-off with performance and consistency.

Finally, automatic detection of dynamic metaphors in political discourse could be misused for strategic purposes, including optimisation of political messaging and monitoring of rhetorical framing. Our aim, however, is analytical rather than interventionist: the method is intended to support critical analysis of political discourse, not to influence public perception. Any future application of such systems should therefore take into account the risks of manipulative use.

The VUTMC dataset used in this work, co-authored by the last author of the present study, consists of annotated documents produced by the US government, sourced from governmental websites. These materials were originally published by governmental agencies for the purpose of communicating trade policies to the public. They, therefore, do not contain personal, confidential, or sensitive information. In this study, the original sources are referenced wherever possible to acknowledge the governmental agencies that produced the dataset. By US copyright law, such documents are in the public domain. The linked OSF repository (Tan and Cienki, 2023) contains metadata of this dataset, as well as the extracted annotated data used in the present study, which is licensed under CC-BY 4.0. The extracted annotated data used in the present study is also included in the linked GitHub (`github.com/berkay5366/Dynamic-Metaphoricity-Cues`). Our use of the VUTMC dataset is consistent with its intended purpose of advancing dynamic metaphor research.

10. Acknowledgements

We want to acknowledge the PhD grant offered by VU Amsterdam for developing the VUTMC corpus. We also want to acknowledge the Young Scientists Fund of the National Social Science Fund (Grant No. 24CYY102).

11. References

Julia Birke and Anoop Sarkar. 2005. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Lynne Cameron. 1999. Identifying and describing metaphor in spoken discourse data. In *Researching and Applying Metaphor*, pages 105–132. Cambridge University Press.

Lynne Cameron. 2007a. The affective discourse dynamics of metaphor clustering. *DOAJ (Directory of Open Access Journals)*. Accessed: 2025-06-13.

Lynne Cameron. 2007b. Patterns of metaphor use in reconciliation talk. *Discourse & Society*, 18(2):197–222.

Lynne Cameron. 2010. The discourse dynamics framework for metaphor. In Lynne Cameron and Robert Maslen, editors, *Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities*, pages 77–96. Equinox.

Lynne Cameron and Alice Deignan. 2003. Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol*, 18(3):149–160.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773. Association for Computational Linguistics.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022. Explainable metaphor identification inspired by conceptual metaphor theory. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, Virtual Conference. Association for the Advancement of Artificial Intelligence.

Raymond W. Gibbs and Lynne Cameron. 2007. The social-cognitive dynamics of metaphor performance. *Cognitive Systems Research*, 9(1–2):64–75.

Wenwen Guan, Marijn Alta, and Jelke Bloem. 2025. Enhancing the automatic classification of metadiscourse in low-proficiency learners' spoken and written English texts using XLNet. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 27–41, Suzhou, China. Association for Computational Linguistics.

George Lakoff. 1994. Master metaphor list index. Accessed: 2025-06-13.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Yujie Lin, Jingyao Liu, Yan Gao, Ante Wang, and Jinsong Su. 2025. A dual-perspective metaphor detection framework using large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria. 2023. MetaPro Online: A computational metaphor processing online system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–135. Association for Computational Linguistics.

MetaPro Team. 2025. MetaPro 2.0. Accessed: 2025-06-13.

MiniMax. 2026. Minimax-M2.5: Built for real-world productivity. https://www.minimax.io/news/minimax-m25.

Jeffery Scott Mio, Ronald E. Riggio, Shana Levin, and Renford Reese. 2005. Presidential leadership and charisma: The effects of metaphor. *The Leadership Quarterly*, 16(2):287–294.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*. Association for Computational Linguistics.

Cornelia Müller. 2008. *Metaphors Dead and Alive, Sleeping and Waking: A Dynamic View*. University of Chicago Press, Chicago, IL.

NVIDIA. 2026. Nemotron 3 Super: Open, efficient mixture-of-experts hybrid Mamba-transformer model for agentic reasoning. Technical report, NVIDIA.

OpenAI. 2025. Introducing o3 and o4-mini. Accessed: 2025-06-13.

OpenCode. 2026. Zen. https://opencode.ai/docs/zen/. Accessed: 2026-03-12, OpenCode documentation.

Kateryna Pilyarchuk and Alexander Onysko. 2018. Conceptual metaphors in Donald Trump's political speeches: Framing his topics and (self-)constructing his persona. *Colloquium: New Philologies*, 3(1):98–118.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Sebastian Reimann and Tatjana Scheffler. 2024. When is a metaphor actually novel? Annotating metaphor novelty in the context of automatic metaphor detection. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Wenhao Huang, Ge Zhang, and Jie Fu. 2024. CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3357–3366, Torino, Italia. ELRA and ICCL.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–616.

Xiaojuan Tan. 2023. *Static and Dynamic Metaphoricity in U.S.-China Trade Discourse: A Transdisciplinary Perspective*. Ph.D. thesis, Vrije Universiteit Amsterdam, Amsterdam, Netherlands.

Xiaojuan Tan and Jelke Bloem. 2024. Broadening the coverage of computational representations of metaphor through dynamic metaphor theory. In *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024*, pages 40–50, Torino, Italia. ELRA and ICCL.

Xiaojuan Tan and Alan Cienki. 2023. Dynamic metaphor theory.

Xiaojuan Tan and Alan Cienki. 2024. A dynamic metaphor perspective on Trump and Xi's trade negotiation in governmental discourse. *Language & Communication*, 96:42–53.

Xiaojuan. Tan and Alan Cienki. in press. YinYang Dynamics of Metaphoricity (YYDM) in discourse. *Metaphor & Symbol*.

Xiaojuan Tan, Alan Cienki, and Bertie Kaal. 2024. The diachronic and cross-linguistic use of trade metaphors in US-China governmental discourse: A socio-cognitive approach. *Metaphor and the Social World*, 14(1):130–153.

Xiaojuan Tan, Alan Cienki, and Tina Krennmayr. 2026. A comparative study of trade metaphors across political genres during the US-China trade war. *Language Sciences*, 115:101797.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Vellum. 2025. LLM leaderboard 2025. Accessed: 2025-06-13.

VisMet Project. 2025. Relation to metaphor. Accessed: 2025-06-13.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Long Papers*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1483–1497, Toronto, Canada. Association for Computational Linguistics.