Quaderni del CIRM

Centro Interuniversitario di Ricerca sulle Metafore

La metafora ha un'identità complessa e plurale, il cui studio coinvolge un numero elevato di discipline e competenze diverse. È una strategia attiva al servizio del pensiero spontaneo e coerente, che motiva le estensioni di significato lessicale – e quindi della polisemia – nonché il mutamento storico dei valori e dei contenuti lessicali. Come tale, è una struttura convenzionale che fa parte di un patrimonio di risorse sulle quali il parlante fa affidamento. Tuttavia, è anche un procedimento di creazione concettuale che coinvolge le strutture portanti della grammatica delle lingue, i cui esiti spaziano dall'invenzione poetica alla creazione di concetti scientifici e filosofici, e più in generale di concetti e termini appartenenti ai più svariati ambiti specialistici. In questo senso, è uno strumento attivo nella costruzione dei testi di qualsiasi natura e contenuto, dai testi letterari e poetici all'argomentazione politica. Per queste diverse ragioni, la metafora, oltre ad essere in questo momento il tema forse più studiato nell'ambito delle scienze del linguaggio, ha una portata interdisciplinare senza paragone. Il suo studio coinvolge la linguistica, la terminologia, la stilistica, l'analisi dei testi e dei discorsi, sia letterari che funzionali, la traduzione, la critica letteraria, la filosofia (dall'estetica all'epistemologia), le scienze cognitive e le loro basi neurologiche.

La collana, in collaborazione con il Centro Interuniversitario di Ricerca sulle Metafore (creato dalle Università di Genova, Cagliari, Modena e Reggio Emilia, Torino), si propone di valorizzare la ricchezza interdisciplinare della tematica in prospettiva interlinguistica e interculturale, proponendo pubblicazioni di orizzonti scientifici, approcci teorici e culturali diversi.

Quaderni del CIRM

Centro Interuniversitario di Ricerca sulle Metafore

Numero 5

a cura di
Michelangelo Conoscenti, Annamaria Contini, Ruggero Druetta, Elisabetta Gola, Adriana Orlandi, Paola Paissa, Ilaria Rizzato, Micaela Rossi, Daniela Francesca Virdis

UNIVERSITÀ

tab edizioni

viale Manzoni 24/c
00185 Roma
www.tabedizioni.it

Prima edizione novembre 2025
ISBN versione cartacea 979-12-5669-272-9
ISBN versione digitale 979-12-5669-273-6

Indice

Portrait of the AI as a Young Metaphorist

Michelangelo Conoscenti

Introduction and Theoretical Background

This work builds on Zottola and Conoscenti (forthcoming), in which we explore the similarities and differences in the representation and discursive construction of AI. Our analysis compares two corpora – one authored by academics and the other generated by AI itself. Given that AI now regularly interacts with humans, this study focuses on conversations around the concept of metaphor, derived from interactions between myself and various AI systems. All the queried systems operate through human-trained Large Language Models (LLMs) to produce natural language output. As such, these interactions can be understood within the framework of dialogic communication (Wodak, Meyer 2009) and, with regard to the linguistic-psychological dynamics of the exchange, as conversational joint actions between human and machine (Clark 1996). In this specific study, the AI does not conceal its artificial nature, in contrast to the original premise of the Turing Test. Indeed, Jones and Bergen (forthcoming) have shown that contemporary LLMs are capable of passing the Turing Test. Here, interactions take the form of 'interviews' on the nature of metaphor – both as defined by the AI and in terms of the metaphors AI uses to describe itself. The purpose of these questions is to prompt the system to reveal its perceived characteristics and self-concept, thereby enabling the construction of a taxonomy of metaphors for AI derived from the system's own perspective.

It will be demonstrated that the AI-generated corpus exhibits advancements in logical-abstract reasoning and increasingly natural interaction patterns. The platforms offer critical and balanced self-descriptions, which are demonstrably influenced by the tone and phrasing of their human interlocutors.

1. Research Questions

As LLMs, such as ChatGPT, increasingly become the primary medium through which humans interact with machines, this study seeks to explore the extent to which they can demonstrate forms of what might be considered *autonomous thinking* – that is, generating language that is not perceived by humans as unnatural and that sustains coherent and consequential conversations. LLMs are described in the literature (Di Bello 2023) as *Statistical Parameter Aggregators* – systems that lack self-awareness and instead rely on probabilistic adjustments to refine their output. These models are thus heavily dependent on feedback mechanisms that incorporate both prior interactions and the training data on which the LLMs were developed. To limit feedback effects – i.e., to ensure that previous conversations did not influence the current interactions – all dialogues in this study were conducted without logging into the system. This approach was necessary due to the inherent nature of GPT architecture, which stands for *Generative Pre-trained Transformer*, a family of transformer-based language models. In Zottola and Conoscenti (forthcoming), Section 4 (*Heuristic Explorations and Hermeneutic Processes: Interacting with the AI*), the capacity of AI outputs to resemble or simulate human reasoning is examined through a parallel between the *Transformer model architecture* (Vaswani *et al*. 2017) and Peirce's semiotic concept of the *interpretant*. Both serve as mediating agents in the production of meaning. In the context of transformer models, the attention mechanism is a core technique that allows the model to weigh the relative importance of different elements in the input sequence when generating responses. This process enables the model to prioritize the most relevant information[1], thereby producing output that sounds human. In doing so, it meets the expectations of the human interpretant – expectations often shaped more by wishful thinking (i.e., "the AI says what I hope to hear") than by rational analysis. This dynamic also helps explain why humans are easily 'fooled' by AI: we are so eager to have our expectations confirmed that we overlook the fact that the AI is statistically trying to satisfy us[2], even if that means assembling fragments of unrelated data – in effect, an analogue to human deception.

1. This mechanism, which is essential for coordinating the input-output flow between the two interlocutors, realises and replicates Clark's (1996, p. 81) fundamental coordination device, i.e. salience: "Perceptual salience is all too often ignored as an essential coordination device in language use".

2. See the final part of Section 4, *Discussion*, for several examples.

Within this interpretive framework, Mirzadeh *et al.* (2024, p. 2) further observe that:

> Logical reasoning is a critical trait of intelligent systems. Recent advancements in LLMs have demonstrated significant potential across various domains, yet their reasoning abilities remain uncertain and inconsistent.

These researchers tend to focus primarily on logical-mathematical reasoning, often treating it as a given, while paying comparatively less attention to the systems' capabilities in natural language processing (NLP). By contrast, the objective of this study is to examine the metaphors that are emerging in public discourse and to explore how they are appropriated and repurposed by LLMs. This focus is grounded in the premise that metaphor is closely tied to human communication: people use metaphors to conceptualize abstract ideas through concrete, real-world experiences. Accordingly, metaphors can serve as both a critical and meaningful benchmark for assessing the development of LLMs in the realm of NLP.

2. Methodology

To investigate these research questions, I adopt a mixed-methods approach combining *Corpus Linguistics* (Brookes, McEnery 2022) to handle large volumes of data, with *Critical Discourse Analysis* (CDA) (Fairclough 1995) to interpret how discourses about and produced by AI are constructed and framed. CDA provides analytical tools for interrogating representations and the discursive construction of a given object, as well as the cultural environments in which such representations are embedded – highlighting the multiplicity of meanings attributed to a single concept. The corpus used in this study consists of 16 questions and answers generated through interactions between the author and three web-based AI platforms: *ChatGPT*, *HuggingChat*, and *Gemini*. The exchanges took the form of interviews exploring both the nature of AI and a series of questions designed to elicit the systems' self-descriptions and self-perceptions. These platforms were selected to reflect varying levels of technological maturity, regulatory frameworks, ethical considerations, and usage constraints. During the course of interaction, I gradually disclosed (in Q1, Q7, and Q9) my identity as a researcher and the purpose of the inquiry. As the conversations progressed, I observed a noticeable improvement in the systems' ability to process and elaborate

on abstract ideas. In response to this, I reformulated Q1 and Q2 to better align with the AI's increasingly sophisticated output. Questions 3, 4, 12, and 13 served as control questions to assess consistency and responsiveness.

To preserve the anonymity of the sessions and avoid feedback loop effects (i.e., influence from prior logged interactions), all dialogues were conducted without logging into any platform. Nonetheless, I documented each conversation. For example, the phrase *Your response is highly intriguing and thought-provoking* was employed across multiple interactions to assess whether the systems retained contextual memory – either within the same session or across different sessions – and whether they adapted their output accordingly.

Additionally, during the summer of 2023, it was observed that LLMs – particularly ChatGPT – began to exhibit slower response times, likely due to increased user demand[3]. Scholars such as Ribino (2023), Lee and Wang (2023), and Yin *et al.* (2024) have explored the role of politeness in human-AI interactions. In parallel testing conducted with other users, I found that incorporating politeness markers – such as "Dear", "Please", and other attention getters – resulted in more immediate and elaborated responses. These cues were consistently used throughout all interactions. Over time, the LLMs began to expand their output without specific prompting and increasingly included emoticons, innuendos, and even irony.

Moreover, unlike the original prescriptions of the Turing Test, LLMs do not attempt to conceal their artificial nature. On the contrary, they often '*play*' with the user, frequently adopting a third-person stance and employing innuendos that can appear ironic or even subtly mocking. As a result, questioning AI becomes a form of heuristic exploration[4], adding to the ambiguous communicative environment that LLMs exploit to generate responses for end users (Di Bello 2023). Using the tools of CDA, these interactions are reinterpreted as a hermeneutic process, one that reveals the evolving characteristics embedded in the interactional dynamics themselves[5].

3. Refer to the following source: https://indianexpress.com/article/technology/tech-news-technology/chatgpt-is-getting-worse-with-time-study-shows-8855737/.

4. The implied reference to HAL 9000 is intentional rather than incidental. For readers unfamiliar with science fiction, HAL is the acronym for *Heuristic ALgorithm* and the sentient AI supercomputer featured in Stanley Kubrick's *2001: A Space Odyssey*.

5. This paper, as well as Zottola and Conoscenti (forthcoming), stems from the same assumption. I replicated Jung's approach in writing his introduction to the *I-Ching*. Faced with something impenetrable and complex, Jung allowed *The Book of Changes* to guide the development of his preface. Since LLMs are similarly opaque to the end user in terms of how their algorithms generate responses, I adopted a comparable strategy: I allowed the LLMs within generative AI systems to speak for themselves, recognizing

For this reason, the conversations were structured around two central themes. The first involved questioning LLMs about the metaphors found in academic literature, with a particular – but not exclusive – focus on the current scholarly debate. In parallel, questions were posed about the general state of metaphor research. These initial exchanges served as a kind of 'warm-up' activity, designed to momentarily destabilize the internal Statistical Parameter Aggregators discussed earlier. The second theme shifted the focus by asking the systems to engage with the same task – discussing metaphor – but with an explicit emphasis on how metaphors represent AI itself. This approach compelled the LLMs not only to retrieve and summarise existing information, but also to activate their generative capacities with the aim of producing, ideally, original content. To further complicate the interactional framework, the systems were prompted to reflect on and describe themselves in the first person.

3. Results

The collected data comprise a corpus of 33,569 types and 4,462 tokens. Although this would be considered a small corpus by typical standards, its quantitative-qualitative analysis provides valuable insights into the processes under investigation. Notably, the three conversations yielded 160[6] unique CONCEPTUAL METAPHORS. For the purposes of this study, these metaphors are analysed from different perspectives, based on three different taxonomies specifically developed to refine and organise the results.

The first criterion adopted involves assigning each unique CONCEPTUAL METAPHOR[7] exclusively to a single SOURCE DOMAIN[8]. The classification is therefore based on the dominant theme or aspect of AI that the metaphor highlights. As will be discussed further in Section 3.2, some metaphors could plausibly fall under multiple SOURCE DOMAINS; however, in

them as possessing a conversational identity and autonomy – thus becoming (inter)active participants in the discourse generation process.

6. This research involved the generation of specific data extraction routines and tables. Due to space limitations, prompts and outputs have been reduced to the basic CONCEPTUAL METAPHOR they express. The master list of the 160 metaphors referred to in the discussion can be found in the *Appendix*, after the *References*.

7. By convention, CONCEPTUAL METAPHORS are indicated in small capitals.

8. For an initial automated selection of SOURCE DOMAINS, both the Metaphor Identification Protocol Vrije Universiteit MIPVU (Steen *et al.* 2010) and the MetaNet repository were used (https://metanet.arts.ubc.ca/metaphor-databases/). See also the procedure description in Section 3.2 *From Polysemous SOURCE DOMAINS to Multi-Layers SOURCE DOMAINS.*

such cases, each metaphor is assigned to the SOURCE DOMAIN that best captures its primary meaning. This assignment follows a clearly defined procedure outlined later. It is worth noting that the algorithm used for metaphor classification incorporates several automated routines that draw on thematic context. The latter are common in content analysis and are designed to identify recurring narrative patterns – i.e., 'themes' – across data sets. Such patterns are instrumental in describing the phenomenon under examination and are directly tied to specific research questions. This approach prioritises the content of the communicative material as the basis for analysis. The database used in this study was initially developed for Zottola and Conoscenti (forthcoming), whose aim was to document how AI is discussed in academic discourse. Accordingly, the SOURCE DOMAINS defined here are tailored to reflect scientific discussions, focusing on AI's function, development, impact, and internal mechanisms. This framework allows for a structured analysis of how AI is conceptualised and understood through metaphorical language – and whether such conceptualisations are reproduced by LLMs. In Table 1 the results are presented in both absolute numbers and percentages.

Table 1. *SOURCE DOMAINS and Number of CONCEPTUAL METAPHORS (Absolute Value and Percentage).*

Source Domain	**Number of Metaphors**
Cognitive Entity	16 [10%]
Learning & Evolutionary System	16 [10%]
Tool or Assistant	18 [11.25%]
Mirror of Humanity	10 [6.25%]
Danger or Ethical Dilemma	12 [7.5%]
Beyond Human Control	12 [7.5%]
Natural/Environmental Force	16 [10%]
Communicative & Creative Identity	12 [7.5%]
Machine/Industrial Process	12 [7.5%]
Rival or Competitor	10 [6.25%]
Exploration & Discovery	10 [6.25%]
Unseen/Mysterious Force	10 [6.25%]
Ethical/Social Construct	6 [3.75%]
Total	**160 [100%]**

3.1. *From Unique* SOURCE DOMAINS *to Polysemous* SOURCE DOMAINS

If we accept, as Steen (2024, p. 242) has argued, that "most metaphor may be structurally ambiguous between deliberate and non-deliberate meanings, which in turn affords multivalent metaphor use", then both polysemy and metonymy can offer valuable insights into the 'cognitive' construal processes at work in LLMs[9]. This perspective also supports the notion that even SOURCE DOMAINS themselves can be polysemous. Steen (2024, p. 246) employs *Wmatrix* for "the analysis of the semantic fields of specific word senses… which may also be used to begin identifying related conceptual domains". In a similar vein, to detect potential semantic overlaps and to allow for the indexing and attribution of metaphors across multiple SOURCE DOMAINS, the present study employed LIWC-22[10] for corpus analysis. By combining its native routines – (a) *Compare Frequencies*, (b) *Meaning Extraction*, (c) *Language Style Matching*, and (d) *Contextualizer* – with the automated analyses outlined in Section 3.2, a set of SOURCE DOMAINS and SUBORDINATE SOURCE DOMAINS was identified. The number of metaphors attributed to each domain (including its subdomains, if applicable) is indicated in square brackets [#].

1. Nature and Function [63]
   (1a. Tools and Utilities [27], 1b. Educational and Guiding Roles [36]);
2. Human-like Characteristics [81]
   (2a. Learning and Growth [46], 2b. Emotional and Relational [35]);
3. Abstract Concepts linked to Knowledge and Information [107];
4. Technological, Mechanical and Computational [54];
5. Emotional and Relational [35];
6. Ethical and Philosophical [38];
7. Metaphysical and Cosmic [41].

9. Later, Steen (2024, p. 242) notes: "This widespread ambiguity of metaphor is based in a highly frequent, specific combination of properties that have to do with metaphor's variable structures and functions… Corpus research has shown that most metaphor is polysemous (language), conventional (thought), indirect (reference to some world), and non-deliberate (communication)". The aim here is to show how the labelling of SOURCE DOMAINS may vary depending on both the analyst's perspective and the LLMs' Transformer-model architecture.

10. Refer to the following source: https://www.liwc.app/.