Technology, Knowledge and Learning (2026) 31:751–788
https://doi.org/10.1007/s10758-026-09956-9

ORIGINAL RESEARCH

Unraveling Creativity Through Variability: A Comparison of LLMs and Humans in an Educational Q&A Scenario

Michele Braccini[1] · **Gianluca Aguzzi**[1] · **Paolo Baldini**[1]

Received: 6 May 2025 / Revised: 9 January 2026 / Accepted: 19 January 2026 /
Published online: 10 February 2026

Abstract
Large Language Models (LLMs) have demonstrated a remarkable capacity for generating human-quality text across diverse applications. In light of this, the use of these models is becoming increasingly widespread. Education is no exception and so literature is starting to explore the potential of LLMs in educational contexts: ranging from teaching assistant chatbot to personalized tutoring. Although the versatility of these models is evident, the critical question arises whether they can truly replicate the creativity and nuances of human languages, giving students the variability and richness of expression necessary for effective learning, and for nurturing critical and divergent thinking. This is particularly important in the pre-school or primary education context, where novel and insightful content is crucial for effective engagement. Framing this problem from an abstract linguistic perspective, it translates into wondering how effectively the "creativity" of LLM-generated text can be quantified and how it compares to human one. So, recognizing the inherent ambiguity in defining creativity this paper investigates the variability of LLM outputs as a potential proxy, and compares it with observed text variability in a large human-authored dataset. Specifically focusing on early childhood education (ECE), we compare the semantic and syntactic diversity of responses from LLMs and humans prompted to answer like 5-year-olds. Our empirical results show LLMs have lower variability than humans, particularly on repeated questions, though the gap narrows for different questions. This limitation raises concerns for educational suitability, where diverse explanations of concepts are crucial, suggesting a need to enhance LLM expressive range and diversity.

Keywords Large language models · Variability · Computational creativity · Criticality

These authors contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

Through language, human beings interact, express themselves and their emotions and think. Natural language, and in particular its richness and variability, shapes the landscape of thinking, influencing the creation of new ideas and concept. In general, the complexity of language correlates with the complexity of concepts that can be produced (Lewis & Frank, 2016), described and stimulated by it. Language is embodied as a product of human physical experience, which has been renewed for thousands of years. It has an intrinsic socio-cultural connection, it is more the product of traditions, customs and habits than of logical rules. Its complexity and richness are, indeed, the distinguishing feature between us and other living species and, for now, even with non-living ones, such as artificial intelligences. With the advent of the large language models (LLMs), however, even this human bastion seems destined to fall, with machines apparently capable of producing artifacts, such as responses, texts and poems, sometimes different but sometimes even indistinguishable from human ones.

Natural language, together with visual language, is the main medium through which culture and education of all degrees and levels are imparted. Hence, through it not only the notional and methodological corpus of a discipline is conveyed, but as a by-product—more or less implicitly and to a greater or lesser degree depending on the nature of the discipline itself—creativity, critical thinking and divergent thinking are foraged. Interestingly, creativity is an indispensable tool in education and occupies a central role, along with play, in early childhood education (ECE) as opposed to higher levels of education, where precision, method and technicality take over, and convergent thinking overtakes divergent thinking (Beghetto, 2005; Kim, 2011). Creativity, as defined by James Melvin Rhodes in 1961 (Rhodes, 1961), is a feature not only of the final product (product), but also of the mental process that led to it (process), of the characteristics of the individual (person) and of the environment (press) in which it is reified. Thus, in educational contexts, it manifests itself in natural language in many forms: from the educator's use of metaphors to express a concept, to having different interpretations of the same concept (reformulation), to expressing ideas and concepts through figurative language and imagery, to name but a few examples. Especially in the early stages of education, being immersed or not in contexts with different degrees of (linguistic) creativity can have an impact on the creativity that students are likely to develop. Given the increasing use of large language models (LLMs) in education and growing concerns related to their possible implications (Gan et al., 2023; Kasneci et al., 2023; Yan et al., 2024), it becomes increasingly urgent to determine how much their linguistic capabilities can influence the learning process and, ultimately, the cognitive processes of students. That requires starting to measure the linguistic properties of these artificial models, and thus also their linguistic creativity.

However, characterizing a product of the (human or artificial) intellect is notoriously a difficult task. The complexity increases further when trying to develop artificial systems that closely mimic the outputs of the human mind, such as creative and artistic works. In the particular context of natural language, this is primarily due to the fact that it requires on the one hand the measurement of properties such as aforementioned creativity, which are subjective and elusive by their very nature, and on the other hand properties such as complexity, which are difficult to measure without the use of large-scale text datasets, literary collections and corpora for text analysis. Another property, typical of both living and non-living systems,

which is interesting but difficult to grasp due to the finiteness of the system under investigation, is the ability to produce behaviors that are interesting but not chaotic. This property—that is often the prerogative of systems operating in a critical dynamical regime—in natural language translates into the system's ability to generate coherent responses while exhibiting variability that allows it to surprise and engage the interlocutor.

Therefore, recognizing the importance of creativity, complexity and criticality in the evaluation and generation of natural language, while acknowledging the difficulty of measuring them, this paper focuses on the variability of large language models (LLMs) outputs, in their out-of-the-box default configuration. In particular, we compare the variability of LLMs with human variability in a question-reply context. Variability is, indeed, a prerequisite for the aforementioned properties and offers the advantage of being directly measurable with statistical tools, even on relatively short texts, as in our application scenario. The use of output variability as a proxy for creativity is consistent with the literature interested in defining evaluation criteria for creativity support tools (Calderwood et al., 2020; Kreminski et al., 2022). The long-term goal of this work is to deepen our understanding of the distinctive features of human language and its use and to bring artificial systems closer to them. To achieve this goal, we will use semantic and syntactic measures to better characterize the creative and dynamical aspects of generated texts.

A greater richness of LLMs's language, comparable to that of human beings, is a desirable feature and has multiple advantages, especially in educational contexts where they are used as surrogate artifacts of their human counterparts. Firstly, it can improve engagement and interaction with students (Heung & Chiu, 2025). Moreover, as already mentioned, it can have a positive impact on students' cognitive abilities and particularly on their creativity. Last but not least, it counteracts the phenomenon of *deskilling*, which also concerns the linguistic sphere and which is becoming more and more pressing and worrying with the increase adoption of these tools and, above all, with the increase in the frequency and continuity of student-LLM reciprocal interaction and feedback loops. If the latter involves a linguistically poor LLM, e.g., one with limited variability, and persists over a prolonged period, it may lead to an impoverishment of the language, reinforcing syntactically and semantically poor structures that may hinder the development of the learner's linguistic abilities.

Therefore, this work proposes a multidimensional methodological framework to quantitatively assess the difference in variability in LLM responses compared to those generated by humans, under a simulated Q&A educational scenario. The generic linguistic framework developed—encompassing both syntactic and semantic dimensions—is then tested on many families of LLMs of varying sizes in order to achieve the most generalisable results.

The analysis framework developed aims to lay the groundwork for the future definition of quantitative metrics and the development of novel architectures or mechanisms for constructing LLMs. Indeed, although these aspects are beyond the immediate scope of this work, they could be integrated into their development processes and, thus, improve the training and fine-tuning of next-generation LLMs, making them as faithful as possible to the human counterparts they are inspired by.

The paper is structured as follows: in Sect. 2, we provide an overview of the background for our study. In Sect. 3 we describe the research questions that capture the aims of this work. Sect. 4 delves into the details of the dataset, evaluation metrics and methodological procedures employed in our analysis. Sect. 5 presents the results obtained. In Sect. 6 we

discuss the implications of our results in light of the research questions raised. Finally, the Sect. 7 concludes the paper and outlines potential future research directions.

2 Background and Motivation

2.1 Education & LLMs

In a domain as sensitive as education, it is not enough to make a general assessment of the quality, equity and validity of LLM (Chang et al., 2024) outcomes; instead, it is crucial to apply domain-specific experimental approaches and evaluation metrics. The literature presents many use cases exploring the integration of LLMs in educational contexts. For example, they are employed as evaluators (Guo et al., 2024; Seo et al., 2025), as teaching assistants (Dan et al., 2023; Hu et al., 2025) and as personalized tutors. For an exhaustive list of actual and potential applications of LLMs in education, we refer the reader to (Wang et al., 2024).

The widespread use of LLMs in education is accompanied by—given the delicate nature of the domain—a rapid increase in concerns and critical discussion about the potential risks and future implications associated with their application, whether regulated or not. The concerns range from the risk of deskilling and cognitive biases, due to the the risk of over-reliance on LLMs, to the potential lose of creativity and critical thinking skills. The literature on the discussion of these issues and best practices for the design of AI-driven educational systems is flourishing and includes a wide range of topics (Chiu et al., 2023; Heung & Chiu, 2025; Yang, 2022; Yilmaz & Karaoglan Yilmaz, 2023).

Given the scope of this paper, in the following sections we limit our focus to the impact on students' creativity resulting from the use of LLM. In particular, we examine studies investigating the potential benefits, limitations and implications of LLM used in (co-)creative educational and learning contexts.

2.2 Computational Creativity & LLMs

Evaluating creativity in computational systems remains challenging due to its complex, multidisciplinary nature (Lamb et al., 2018). Despite extensive research across psychology, philosophy, and computer science, no universally accepted definition of creativity exists (Loughran, 2017). Rhodes' influential framework divides creativity into four "P's" (Rhodes, 1961): **person** (individual characteristics), **process** (cognitive mechanisms), **product** (output novelty and quality), and **press** (environmental influence). From the process perspective, Boden (Boden, 2004) categorizes creativity hierarchically from combinatorial (recombining existing ideas) to exploratory and ultimately transformational creativity (profoundly altering conceptual space).

Since our study examines differences between LLM-generated and human language, we focus on the **product** dimension. Here, *variability* becomes critical as it directly influences novelty and quality. Research supports this approach: studies have compared artist and non-artist creativity by analyzing drawing variability (Fayena-Tawil et al., 2011), while Ritchie emphasizes that creative systems should generate output that balances novelty with domain

typicality (Ritchie, 2007). Others contextualize novelty historically (Elgammal, 2015) or apply modified Turing tests to assess creative text (Elgammal et al., 2017).

Variability thus represents a fundamental aspect of creativity, encompassing diversity and unexpectedness in generated content. While related to novelty, it also includes elements of controlled randomness. In this paper, we analyze text variability as a quantifiable proxy for creativity, focusing specifically on the product dimension.

Large Language Models (LLMs) have demonstrated impressive text generation capabilities, raising questions about their potential for creativity. While LLMs excel at mimicking existing styles and combining familiar concepts, their ability to generate truly novel and surprising outputs, cornerstones of creativity as defined by Boden, remains a subject of debate. Literature has examined LLMs' performance on traditional creativity tests, such as the Alternative Uses Test (AUT) or the Torrance Test of Creative Thinking (TTCT) (Chakrabarty et al., 2024), where state-of-the-art models approach human-level abilities in some tasks (Stevenson et al., 2022). From a more theoretical and philosophical perspective, Franceschelli and Musolesi (Franceschelli, 2023) analyze LLMs under the lens of creativity theories, highlighting challenges in achieving value, novelty and surprise. They argue that LLMs, since LLMs are trained on vast datasets of existing text, their outputs are often statistical interpolations of learned patterns, potentially limiting their capacity for Boden's transformational creativity. This "probabilistic" approach to producing text is inherently limited in expressively when it comes to creativity.

Recent experimental work provides insights into LLMs' creative capabilities and their inherent variability. Doshi and Hauser (Doshi, 2024) found that while LLM-generated suggestions enhanced the perceived creativity and quality of individual stories (especially for less inherently creative writers), they simultaneously reduced the *collective diversity* of the output. The AI-assisted stories exhibited significantly higher similarity to each other compared to human-only creations, highlighting a tension between individual enhancement and collective novelty.

Sun et al. (2024) instead conducted a comprehensive evaluation across 13 creative tasks (including divergent thinking, problem solving, and writing), finding that state-of-the-art LLMs achieve human-average *individual* creativity (approximately the 52nd percentile) and substantial *collective* creativity (with one LLM queried 10 times matching the output of 8-10 humans). While LLMs excelled in certain domains, they still lagged in creative writing tasks. Unlike their focus on rated creativity scores, our work specifically investigates response *variability* as a distinct, measurable proxy for creative potential. Moreover, these studies do not specifically investigate the relationship between semantic and syntactic variability in generated texts. Understanding this relationship is crucial, as it reveals how lexical diversity (syntactic variation) corresponds to conceptual diversity (semantic variation) and vice versa.

Finally, Anderson et al. (2024) compared ChatGPT with a non-AI tool (Oblique Strategies deck) through semantic analysis of human-generated ideas. Their research revealed that ChatGPT use resulted in *more homogeneous* (less diverse) ideas *collectively* across users, though not within individual users' idea sets. This suggests LLMs may constrain group-level diversity while maintaining individual variation–a finding that complements our investigation into LLM output variability as a fundamental aspect of their creative capabilities.

2.3 Dynamical criticality & LLMs

From the point of view of dynamical systems, the concept of creativity and, more generally, of a process capable of producing an interesting, coherent but not boring output, can be underpinned by the concept of dynamical criticality (Kauffman, 2000, 2016; Sawyer, 1999). Dynamical criticality is a concept taken from the theory of phase transitions and describes an abrupt change in the observable characteristics of a system when specific control parameters (temperature, pressure, and so on) undergo a change (Roli et al., 2018). Systems exhibit interesting properties at the critical point, such as *(i)* universality *(ii)* maximum information exchange *(iii)* the power law of relevant quantities and *(iv)* critical slowing down.

The interesting properties of critical systems led some authors (Kauffman, 1993) to hypothesize that systems operating in a critical dynamical regime, between order and chaos, optimally balance robustness and adaptability, being able to respond consistently but maintaining variability of response if necessary. In addition, its features led to the conjecture of "computation at the edge of chaos" (Langton, 1990; Packard, 1988), which suggests that systems operating at the edge of chaos exhibit the highest computational capabilities. Therefore, critical systems show the highest trade-off between flexibility and reliability. Evidences indicates that the human brain functions near criticality (Beggs, 2008; Beggs & Timme, 2012; Chialvo, 2010), and its hallmarks can be found even in natural language, one of its primary cognitive products, (Ebeling & Pöschel, 1994; Piantadosi, 2014). From a systemic perspective this could provide an explanation of the brain's remarkable capability of adaptation and learning, and maybe creativity, in complex environments.

Few works have started to attempt to answer the question of whether LLMs operates in critical dynamical regime. The authors of the work Zhang et al. (2024) tested pre-trained LLMs on reasoning and prediction tasks of increasing complexity. They found that performance on these downstream tasks correlates with the complexity to which they are exposed during the training phase. In particular, LLMs trained on overly simple or chaotic sequences perform poorly, whereas those exposed to structured yet challenging patterns perform better, suggesting the existence of a level of complexity— which lies between order and chaos—that lead to the emergence of intelligent behaviors. Instead, Nakaishi et al. (2024) investigate whether LLMs exhibit critical features from a linguistic perspective. They have analyzed the part-of-speech correlations in GPT-2's generated text at varying temperatures. They observe power law correlations at a temperature of 1—a hallmark of criticality—suggesting that LLMs, like other complex adaptive systems, may operate near a critical regime where flexibility and coherence are optimally balanced.

3 Research Questions

There is a large body of literature in which computational linguistics measures are used to characterise the degree of creativity in a text. Noteworthy examples include the works of Orwig (2024) and Orwig et al. (2024), where semantic diversity of texts has been found to positively correlate with human evaluation of creativity. The work of Beaty and Johnson (2021) proposes an automated creativity assessment that relies on multiple semantic distance measures. Instead, in an educational context, the work of Kandemirci et al. (2025) employed several syntactic and semantic linguistic measures in an attempt to find predictors

of the creative content of children's writing. The authors demonstrated that these quantitative measures were able to predict, to some extent, human creativity scores.

This overview shows that computational linguistic measures are certainly not sufficient to capture all the complex facets that make up creativity—for which human evaluations remain essential—but it demonstrates that they can be considered objective, scalable and efficient proxies for measuring some of its aspects, as they have been found to correlate with the definition of creativity given by human evaluators.

Furthermore, in an educational context, in addition to quantitative measures for assessing creativity, it is necessary to take into account the variability of verbal/written interactions between a student and a teacher, both in syntactic and semantic terms, see also Sect. 2.2. This is essential in order to capture the originality dimension that characterises creativity. More concretely, if we consider an LLM that acts, for example, as an assistant chatbot or virtual teacher, the experimental protocol to be employed to measure its creativity, and to compare it to human one, must begin to address questions such as: *"Is its response variability such that it would be able to re-explain the same concept in different words to a child who did not understand it the first time?"*; or *"In the long-term, is the linguistic imprinting derived from LLMs as rich as that of human teachers?"*. With this in mind, we can summarize the abstract research questions that this paper aims to answer in this way:

- **RQ1** How does the variability of LLM-generated responses to the same question compare with human one?
- **RQ2** How does the variability of LLM-generated responses vary between different questions compared to human one?
- **RQ3**: To what extent the observed semantic variability is underpinned by syntactic variability, and vice versa?

The multidimensional analysis framework developed to answer these research questions is presented in the following sections.

4 Methods

4.1 Dataset

For this study, we utilised the H3C (Guo et al., 2023) dataset, a comprehensive collection of human-authored Q&A pairs from different domains. Table 1 reports its composition.

The choice of this dataset is motivated by the presence of a large number of human-generated answers from the "r/explainlikeimfive" (ELI5) subreddit. This forum is dedicated to breaking down complex topics into simple and easily understandable explanations, as if the reader were 5 years old. At the end of the questions in this category we find the phrase "Explain like I'm five". This is therefore a good starting point to check the variability of the responses of LLM graduates acting as teaching assistants or as actual teachers for preschool or primary school students. In this case, LLMs will assume the role of teachers, trying to provide answers that are as detailed and accessible to this age group as possible.

We selected questions that received at least 3 human replies to assess the variability among human responses. From the total of 17,000 messages, we randomly selected 1,000

Table 1 Composition of the HC3 dataset, showing the number of human questions and answers in its different subdatasets

	# Questions	# Human Answers	Source
reddit_eli5	17112	51336	ELI5 dataset
open_qa	1187	1187	WikiQA dataset
wiki_csai	842	842	Wikipedia Computer Science Q&A
medicine	1248	1248	Medical Dialog dataset
finance	3933	3933	FiQA dataset

The subdatasets are: *reddit_eli5* (ELI5, a large Reddit-based dataset of simple explanations (Fan et al., 2019)), *open_qa* (WikiQA, open-domain question answering from Wikipedia (Yang et al., 2015)), *wiki_csai* (Wikipedia Computer Science Q&A, technical questions from Wikipedia (Guo et al., 2023)), *medicine* (MedDialog, medical dialogue dataset (He et al., 2020)), and *finance* (FiQA, financial opinion mining (Maia et al., 2018))

Table 2 Overview of the language models used in this study, including their parameter counts and references

Model	References	Params
Large Models (>8B)		
Gemini Flash 2.0	(Google DeepMind, n.d.)	???
Gemma2 27b	Team et al. (2024)	27B
Mistral small	(Mistral AI, n.d.)	22B
Qwen2.5 14b	Yang et al. (2024)	14B
Phi 4	Abdin et al. (2024b)	14B
Medium Models (>4B)		
LLaMA−3.1-8B	(Meta AI, n.d.-a)	8B
Qwen2.5	Yang et al. (2024)	7B
Mistral	Jiang et al. (2023)	7B
DeepSeek R1	Deepseek et al. (2025)	7B
Small Models (<4B)		
Phi−3.5	Abdin et al. (2024a)	3.8B
LLaMA 3.2-3B	(Meta AI, n.d.-d)	3B
Gemma2 2b	Team et al. (2024)	2B
StableLM 1.6b	(Stability AI, n.d.)	1.6B
Smollm 1.6b	Allal et al. (2025)	1.6B

messages for our analysis, ensuring a representative sample of the dataset's semantic and syntactic variability. For each selected message, we generated three responses using several LLMs, both open-source and proprietary. Specifically, Gemini 2.0 (as a reference for current state-of-the-art models), LLaMa 3.1, QWen, among others. The complete list of models considered in this work is presented in Table 2.

4.1.1 Dataset Suitability for Education

Since our analysis focuses on the emerging variability of responses in a simulated educational context, we must first ensure that the human responses contained in HC3's ELI5 dataset possess statistical properties suitable for this purpose. At the same time, we would like to stress that the chosen dataset is not proposed in this work as a gold standard for the

development of educational systems, such as teaching assistants or similar. Rather, it should be intended as a starting point from which to pursue the objectives of this scientific investigation, namely the assessment of emerging variability between LLMs and humans with an educational constraint ("Explain like I'm five") through the development of a multidimensional quantitative framework.

The large number of answers contained in HC3 and its multi-domain composition gives us the opportunity to perform a comparative analysis of ELI5 with respect to other subdatasets. The first comparison is reported in the original HC3 paper (Guo et al., 2023), where the authors compared the *density* of answers—number of different words per normalized answer length—in the different subdatasets. The results show that human responses in ELI5 constitute the *least dense* subset of data, i.e., the one with the fewest different words for the same response length. This is a first evidence that the answers in ELI5 are simpler and more accessible than those in other subsets and therefore, at least with regard to this criterion, they are in line with the premises of the data subset and are consistent with the requirements of the educational scenario considered in this work.

In this work, we take a further step forward by extending and generalising the previous analysis, using several state-of-the-art readability metrics to achieve a more robust result, better evaluate the readability of ELI5 human responses and, at the same time, attempting to approximate the level of education required to understand them. We therefore used the following metrics on 800 randomly selected response samples: *Flesch Reading Ease*, *Flesch-Kincaid Grade Level*, *Gunning Fog Index*, *Dale-Chall Readability Score*, and *Difficult Words*. The metrics were computed using the *textstat* Python package (https://github.com/textstat/textstat). We reported the results of this analysis in Table 8. Overall, the results showed that ELI5 is the easiest to read among the subsets in HC3. Interestingly, the distribution of Flesch Reading Ease values has a mean and median above 60, a third quartile above 70, and outliers even above 100. This means that, according to these criteria, the texts produced by humans in ELI5 tend to be fairly easy to read, in many cases by children in the age group covered by the ECE, which is the the scope of the paper. However, we do not use them to pre-filter the dataset in order to avoid introducing unwanted bias into the study. Indeed, these readability metrics have well-known limitations, as they tend, among other things, to overlook reader diversity and disregard any qualitative differences (Redish, 2000).

4.2 Multidimensional Analysis Framework

To rigorously quantify and compare the response behaviors of humans and LLMs, focusing in particular on their inherent phenomenological variability, we employed a multidimensional analytical framework. This framework provides a structured approach to examine responses from multiple orthogonal perspectives.

The core of the framework consists of the three main similarity macro-analyses described here, which form the structural backbone of the entire work.

- **Coherence Analysis:** This analysis tries to evaluate human and LLM-generated coherence of the replies with respect to the questions prompted. For each question, we compute the pairwise similarity between a question and each of its three responses (see Fig. 1), both human and LLM-generated. This is repeated for all questions in the dataset. The goal is to assess the degree to which responses align with the questions. In addition,

we integrate the previous analysis with a higher level "LLM-as-a-judge" approach in order to obtain a more qualitative assessment of the coherence of the responses.

- **Intra-Question Analysis:** This analysis assesses the degree of agreement among different responses to the same question. So, it takes into account for each question all three answers that are present in the dataset. Operationally, this involves computing pairwise similarities among the three responses to each question (see Fig. 1), producing a total of 3000 similarity values for each model.
- **Inter-Questions Analysis:** This analysis investigates the variability of responses between the different questions. For this analysis, we consider only a single representative response for each question (the first response, precisely). This involves calculating pairwise similarities between the first response of each question (see Fig. 1), for all questions, so a total of $\frac{N\times(N-1)}{2}$ with $N = 1000$ similarity values for each model are collected.

We represent the three types of analysis described above in Fig. 1.

4.2.1 Additional Analyses

To further enrich our understanding of the response behaviors of humans and LLMs, we supplement the analytical framework with additional analyses which, depending on their type, are applicable to Intra-Question, Inter-Question, or both cases. These additional

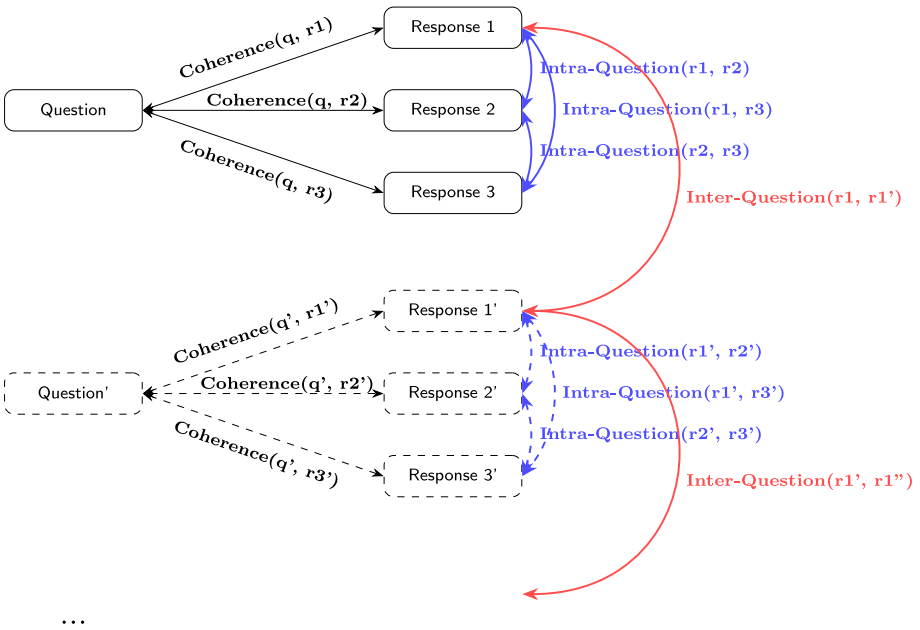


Fig. 1 Schematic representation of the three main types of analysis performed in this study: Coherence analysis (black arrows), Intra-Question analysis (blue arrows) and Inter-Question analysis (red arrows). The dotted boxes represent the repetition of the same structure for all 1000 samples of questions and answers extracted from the dataset

dimensions allow us to gain a more comprehensive and phenomenological understanding of the response behaviour of humans and artificial models.

The additional analyses are:

- **Questions vs. Replies:** This analysis explores the relationship between the similarity of questions and the similarity of their corresponding responses. This analysis allows us to verify if the similarity of responses—semantically and syntactically—correlates with the similarity between the questions that triggered them. This kind of analysis only applies to the *Inter-Questions* case.
- **Semantic vs. Syntactic:** This analysis puts in relation the semantic and syntactic dimensions of the responses, independently from the questions that originated them. It aims to reveal the underlying linguistic patterns and structures that characterize human and LLM responses.

4.2.2 Similarity Metrics

As similarity measures, we employed both semantic and syntactic metrics, including:

- **Cosine Similarity** (Semantic Measure): A standard measure to measure the similarity between two vectors, often used in NLP tasks. If it is computed on the embeddings of the sentences, it can be used to measure the semantic similarity between two sentences. The formula is:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

where the dot represents the inner product operation and $\|\cdot\|$ the norm of the vector.

- **Jaccard Similarity** (Syntactic Measure): A measure of the similarity between two sets. It is defined as the size of the intersection divided by the size of the union of the two sets. It can be used to measure the similarity between two sentences by considering the set of words in each sentence, therefore focusing on the syntactic similarity. The formula is:

$$\text{jaccard\_similarity}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

The results section will present analyses of some relevant combinations of the cases belonging to the two orthogonal axes presented.

For the cosine similarity computation, we employ the Python library *scikit-learn*.[1] For embedding generation, we employ the `text-embedding-04` model,[2] which is widely adopted and consistently delivers strong performance across a range of NLP tasks. For the

[1] https://scikit-learn.org/stable/

[2] https://ai.google.dev/gemini-api/docs/models#text-embedding

Jaccard syntactic measure computation, we used the OpenAI tokenizer `tiktoken`[3] to create the set of tokens for each sentence, allowing for accurate measurement of lexical overlap.

5 Results

5.1 Coherence Analysis

The Fig. 2 shows the distribution of similarities between all questions and their corresponding replies, capturing how closely the responses relate to their prompting questions.

Looking at the semantic dimension (cosine similarity), we observe that most LLM models exhibit a median similarity value of approximately 0.8 with their prompting questions, while human responses show a lower median of around 0.7, and even a lower mean. This difference suggests that LLM-generated responses tend to stay more semantically aligned with the original question, potentially indicating a more direct question-answering approach. The phi3.5 model stands as a notable exception with a median of only 0.3, which upon manual inspection confirms that this model frequently generates incoherent or hallucinatory content unrelated to the questions posed.

The interquartile range (IQR) is significantly wider for human responses than for most LLMs, demonstrating that humans exhibit greater variation in how closely they adhere to the question's semantic content. This could reflect humans' tendency to incorporate personal experiences, tangential thoughts, or creative interpretations when responding to questions—a behavior less pronounced in most LLMs.

From the syntactic perspective, we observe lower overall similarity values across all sources, indicating greater lexical divergence between questions and answers than semantic divergence. All distributions, both human and AI, exhibit a median syntactic similarity of approximately 0.1, except for phi3.5, which shows a median lower than 0.05. This suggests

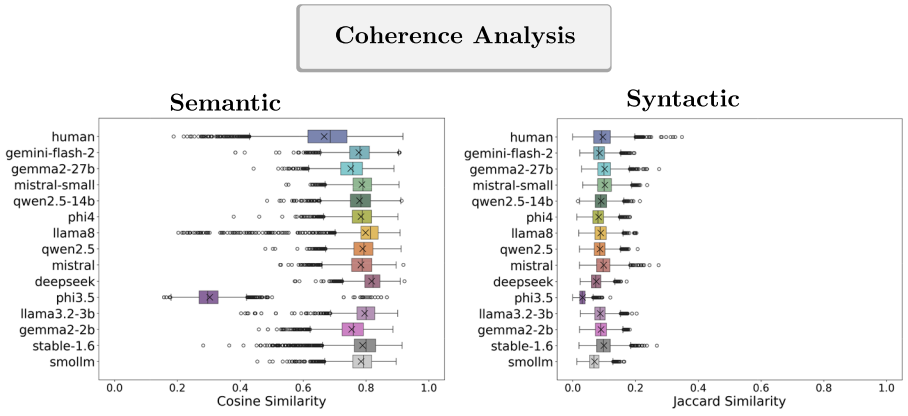


Fig. 2 Coherence Analysis. The distributions show the semantic and syntactic similarities between all pairs of replies to the same question, for all questions. The black X sign represents the mean of the distribution. The cosine similarity is shown on the left and the Jaccard similarity on the right

[3] https://github.com/openai/tiktoken

that in all cases, human and non-human, there is a tendency not to reuse the vocabulary presented in the question, and therefore new vocabulary often emerges in their answers. These results were somewhat predictable with regard to humans, as the metrics capture the diversity and richness that could be found with interaction with different humans. However, the syntactic difference between question and answer for LLMs is interesting. A partial explanation for their tendency to be distant from the syntactic content of the question is probably due to their inclination to maintain contextual relevance or their induced tendency to maintain conversational type interactions with the user. This last hypothesis is partly supported by the examples of replies reported in Table 5.

In general, the results of the coherence analysis suggest that both the responses generated by humans and those generated by the LLMs—with the exception of phi3.5—are not completely random or chaotic, but rather show a certain degree of consistency with the questions prompting them, albeit using very different sets of tokens.

Although quantitative coherence analyses aim to detect any deviations in the response to a question, they may not capture all aspects of natural language and, in particular, the way humans use it. Indeed, humans resort to rhetorical expedients such as metaphors, analogies, abstractions and digressions to answer even a simple question. Therefore, while maintaining an automated and objective approach, we integrate the above analysis with an "LLM-as-a-judge" approach to assess the relevancy of responses provided by both artificial models and humans to a given question. This high-level approach is supported by scientific literature (Bavaresco et al., 2025) and is applied when the amount of data is high and/or it is not possible to involve human evaluators in the process of assessing question-answer coherence. To this end, we employed the answer relevancy metrics provided by LlamaIndex (Liu, 2022), namely the method *AnswerRelevancyEvaluator*. The relevancy metrics returns a score in [0, 1], higher score means higher relevancy. We performed these experiments using the *gpt-oss:120b* model as the judge, OpenAI's open-weight model with 120 billion parameters. The choice of a relatively large model is motivated by the need to ensure that the evaluation model has sufficient capacity to understand and assess the relevance of the responses provided, while at the same time ensuring a reasonable response time, given that the same evaluation must be repeated for multiple artificial models and multiple pairs of questions and answers. For these last reasons, we computed relevance metrics for 30 pairs of sampled questions and answers randomly extracted from the 1000 pairs used for all other statistics. The results of this analysis are shown in Table 4 and show that all LLMs and humans have fairly high average relevance scores, except *phi3.5*, which has a relevance score of 0 for all samples. This provides proof, albeit partial, of the validity of the computational procedure used to evaluate the relevancy. Some models stand out because their average relevance score is equal to 1, e.g., *gemini-flash-2*, *mistral-small*, etc. It is also worth noting that humans are very close, for the two statistical parameters presented, namely the mean and standard deviation, to the values of the *stable−1.6* and *smollm* models, which share the common characteristic of being relatively small models. Overall, this means that regardless of the actual content of the responses, and therefore the possible use of metaphors, digressions and other narrative devices that both humans and models may have employed, statistically the response given to the questions was deemed relevant by the artificial judge. This means, therefore, that both the dataset of human responses used in the first instance and the responses given by LLMs in the second instance can be considered adequate for the purposes and objectives pursued in our work.

5.2 Intra-Question Analysis

As detailed in the methods section, the Intra-Question Analysis focuses on the pairwise similarity between answers to the same question, for all questions. Therefore, the analyses in this section take into account all three answers for each question.

5.2.1 Replies Similarity Distribution

The Fig. 3 shows the distribution of similarities between all pairs of answers to the same question, for all questions.

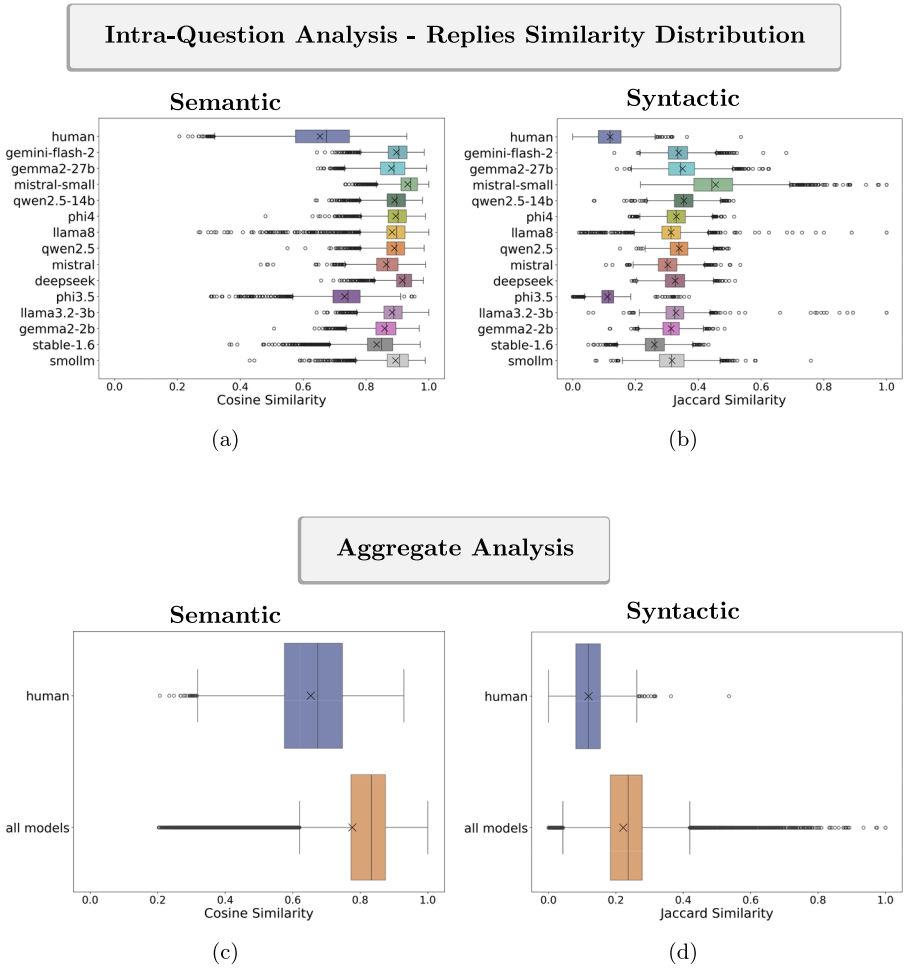


Fig. 3 Intra-Question Analysis - Replies Similarity Distribution. (Top) Comparison of semantic (**a**) and syntactic similarity (**b**) distributions for all pairs of answers to the same question, for all questions. (Bottom) Aggregate comparison of semantic (**c**) and syntactic similarity (**d**) for the *Intra-Question Analysis* case, comparing humans against all LLM models under investigation. The black X sign represents the mean of the distribution. The cosine similarity is shown on the left and the Jaccard similarity on the right

It can be noted that the human responses exhibit a different trend of replies compared to LLM-generated ones. In particular, for both semantic and syntactic measures employed, the distributions of human replies are statistically lower—in terms of central tendency—than those generated by LLM. Furthermore, the interquartile range (IQR) is wider for human responses in the semantic case, confirming a greater dispersion in the data and, consequently, higher variability. This result captures the human ability to provide pairs of answers characterized by both large and small differences in meaning, with a prevalence of the former over the latter. This result—combined with the one presented in the previous section—leads us to conclude that humans manage, unlike LLMs, to combine variability and creativity of response and at the same time coherence with the prompted question, and hence no completely random or chaotic behavior is observed.

Only the phi3.5 model appears comparable to the human distributions; however, further analysis of its responses revealed hallucinatory behavior. Therefore, its behavior cannot be considered similar to human behavior, despite the similarity values collected.

Another interesting case is the mistral-small model, which, probably due to its small size, shows high cosine values and also—in relation to the other models—Jaccard values.

To assess the statistical significance of the differences between all models tested, the Wilcoxon test on pairwise similarity distributions was performed. The results, reported in Fig. 12, show that the differences between almost all models are significant, and especially the human distribution is statistically significant ($p < 0.05$) compared to all artificial models for both semantic and syntactic measures. Heatmaps reported in Fig. 15 shows the one-to-one model comparison for both the semantic and syntactic measures. For each question in the dataset, we computed the mean similarity between the first response generated by each model pair. The results show that—with the exception of phi3.5—the differences between humans and each of the other models are the most dissimilar, represented in the figure by the darker color. As for the other models, we can observe a color gradient of similarity reflecting the size of the models.

To easily compare the trends of the different models, we compare human pairwise similarities against all the LLM-generated responses. This aggregate analysis facilitates numerical comparison of trends and mitigates the selection bias present in the dataset and, in general, inherent in human–machine comparisons. Human response variability, indeed, could arise from factors like age, socio-economic status, and education. Combining diverse LLM responses contribute to reduce this bias, enabling a fairer evaluation. This analysis is showed in Fig. 4c and d. The procedure used for the aggregate analysis is as follows:

1. For each question, we collect all the responses generated by the LLM, i.e., a total of 3 times the number of artificial models replies for each question.
2. Then, we calculate the similarity between all the answers produced by all the artificial models.
3. We repeat the procedure described above for all questions and collect all pairwise similarity values, whose distribution is represented by a boxplot.

This aggregate analysis confirms the previous trends for both semantic and syntactic measures. Noteworthy is the spread of the distribution and thus variability of human responses for the semantic measure, which is greater than that of LLM-generated responses even when aggregated.

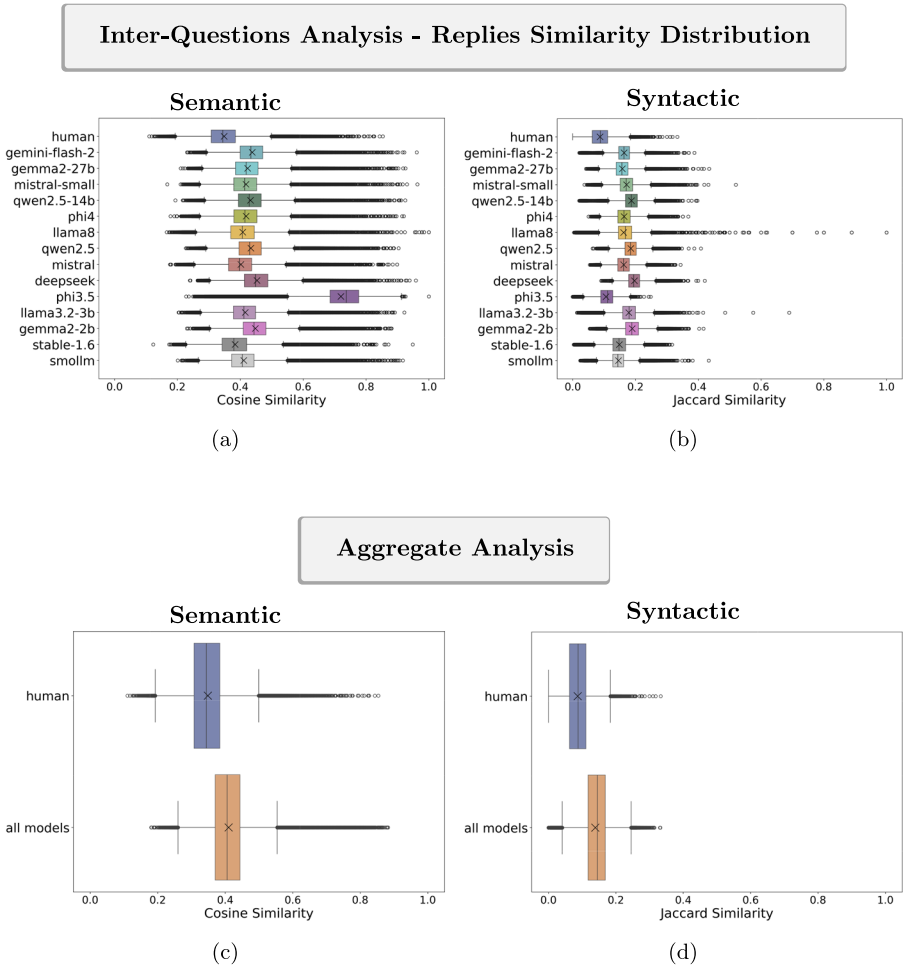


Fig. 4 Inter-Questions Analysis - Replies Similarity Distribution. (Top) Comparison of semantic (**a**) and syntactic (**b**) similarity distributions for all pairs of answers to different questions. (Bottom) Aggregate comparison of semantic (**c**) and syntactic (**d**) similarity for the *Inter-Questions Analysis* case, comparing humans against all LLM models under investigation. The black X sign represents the mean of the distribution. The cosine similarity is shown on the left and the Jaccard similarity on the right

5.3 Inter-Questions Analysis

In this section, we analyze the variability of answers between the different questions, focusing on the semantic and syntactic similarity between the first answer to each question.

5.3.1 Replies Similarity Distribution

The Fig. 4 shows the distribution of similarities between all pairs of answers in the Inter-Questions case. As in the Intra-Question case, this type of analysis is useful to understand

how similar the answers to two different questions are in general, irrespective of the similarity between the questions that elicited them.

Observing Fig. 4 it is firstly noticeable that in terms of absolute values, the values are much lower for both the cosine and Jaccard measure than in the previous case. This, not surprisingly, tells us that in general answers to different questions are more dissimilar than answers to the same question. However, it is interesting to note that when different questions are asked, the gaps between the distributions of human and LLM responses are less pronounced than in the Intra-Question case.

Once again, it is worth noting the degenerative behavior of the phi3.5 model, which in this case shows a very peculiar behavior: high cosine values and low Jaccard values. This is further evidence of its hallucinatory behavior, perhaps caused by the model's excessive tendency to generalize. This could explain both the high semantic similarity values and the low syntactic similarity, as it does not generate tokens completely randomly as a white noise process would, but generates sentences with a certain semantic structure, replicated in all responses.

For the inter-question case we also produced the aggregate analysis to try to mitigate the selection bias implicit in the dataset. In particular, since considering all LLMs answers is on the one hand computationally intensive and on the other hand would produce a set of pairwise similarities with much higher cardinality than human, we decided to follow the following procedure:

1. For each question, we randomly select one LLM model and take the first answer it generates.
2. We replicate this procedure for all the 1000 questions.
3. Next, we calculated pairwise similarities between all replies, as done for the non-aggregate analysis.

This aggregate analysis, shown in Fig. 5c and d, puts the results between artificial and human models even closer for both the semantic and syntactic dimensions. This seems to suggest that not only do LLMs, in their out-of-the-box configurations, come close to what is the response behavior we observe in humans, but, furthermore, that if one treats different LLMs as different humans, the differences become even narrower, statistically.

5.3.2 Questions vs Replies

In this section, the similarity of answers is put in relation to the similarity of the questions that generated them. This type of analysis only applies to the Inter-Question case and aims to check whether the similarity of the answers—from a semantic and syntactic point of view—correlates with the similarity of the underlying questions. Furthermore, it provides an insight into the distribution of pairwise similarities between questions along the x-axis, a dimension that had not been considered in previous analyses.

The analysis is reported in Fig. 5, in the form of a hexagonal binning plot (hex bin). The color maps represent the densities of the hexagonal cells: the more intense the colours, the denser the corresponding regions are. A number of 30 cells in the x-direction are used and, too ensure that all bins containing at least one data point are visible, a minimum count threshold equals to 1 is set.

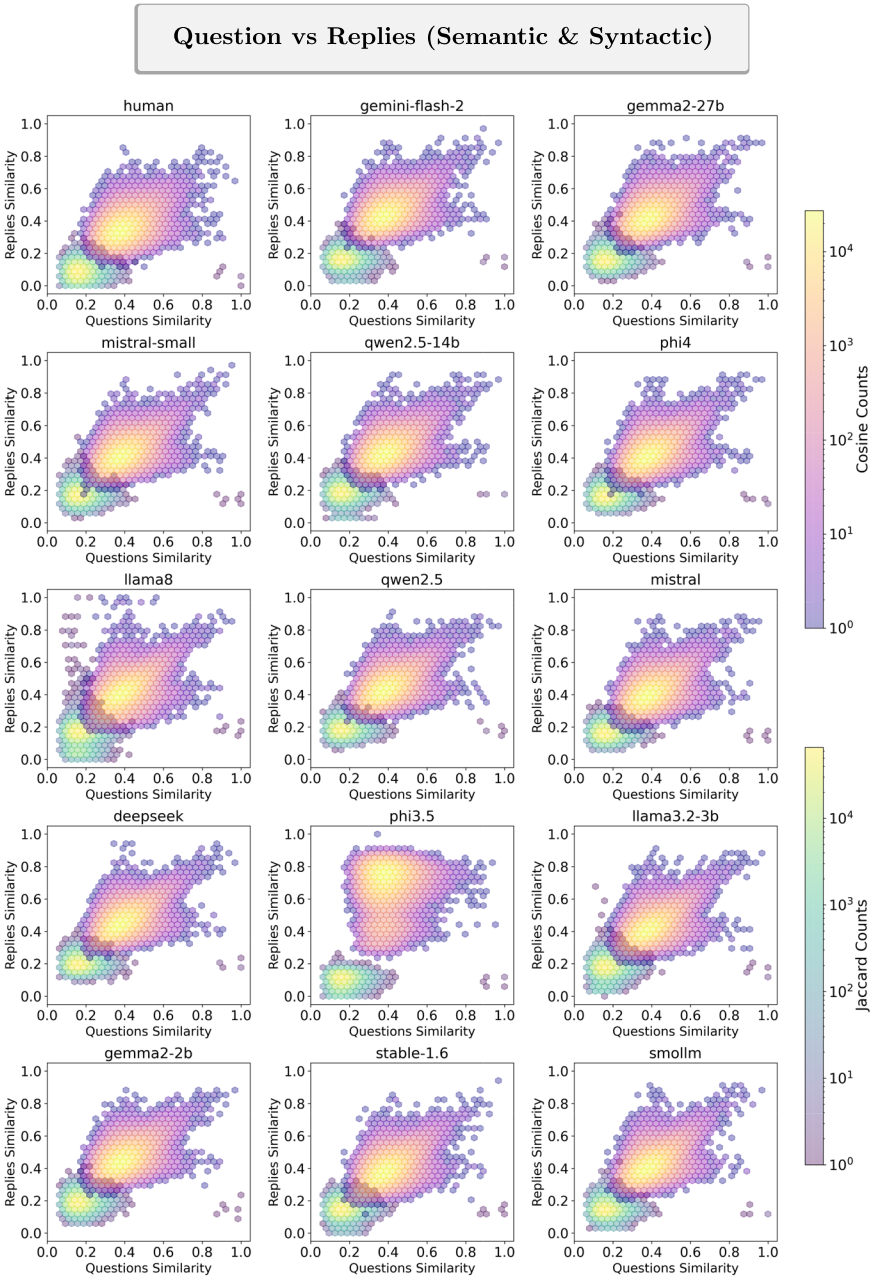


Fig. 5 Semantic (cosine) hexagonal binning plot (with a number of 30 hexagons in the x-direction and a minimum number of points per cell of 1) of the relationship between cosine similarity of the pairs of questions and the cosine similarity of their corresponding answers. The plot employs hexagonal bins with color intensity representing the density of points

Although there are no major differences between the models, some interesting trends can be observed. The hexagonal cosine and Jaccard cells for the human case are placed at slightly lower y-values than the other models, as already noted in Sect. 5.3. The cloud of points relative to phi3.5 is further proof of its difficulty in producing semantically different answer pairs, even when the questions are very different. For the semantic perspective, in general, in almost all cases a positive correlation between the similarity of questions and answers is observable. However, some slight differences between artificial models and humans are found in the x-axis values around 0.8: indeed, humans show greater variability, whereas LLMs exhibit a bimodal distribution of points with peaks on the replies similarities (y-axis) around 0.4 and 0.9.

From a syntactic point of view, the same considerations made in Sect. 5.3 remain valid. The only behavior worth mentioning is that of llama8, which shows several categories of syntactic similarity between answers, albeit at low cardinality, in the face of dissimilar questions.

5.4 Semantic vs Syntactic Analysis (Inter-Questions & Intra-Question)

We relate the semantic and syntactic dimensions of pairwise similarities to try to begin to reveal the phenomenological relationships of language, independently of the questions or contexts that originated them. The Fig. 6 shows the semantic-syntactic relationship of similarities between all pairs of answers for both the Inter-Questions and Intra-Question cases—by means of a hexbin density distribution of cosine similarity (on the x-axis) with respect to Jaccard similarity (on the y-axis). These plots show the joint distributions of the marginal distributions reported in Figs. 3 and 4.

By analyzing the plot some interesting trends emerge. First of all, LLMs feature a clear shift in semantic-syntactic behavior passing from the Inter-Question to the Intra-Question compared to what is observed in the human case. In particular, with the degree of variability varying between models, the Intra-Question points show a tendency towards regions of high syntactic similarity as cosine similarity increases, forming a shape resembling a right-angled triangle. Generally speaking, all LLMs, with the exception of phi3.5 which, intriguingly, in its hallucinatory behavior shows a semantic-syntactic spectrum more similar to that of humans (more pronounced for the Intra-Question case), have distributions placed slightly higher and further to the right than those of humans: this confirms their aforementioned difficulty in producing truly different responses, semantically and syntactically. In the human case, on the other hand, the Intra- and Inter-Questions points are distributed more similarly, with distributions whose shapes are mostly overlapping.

The aggregate plots reported in Fig. 7 confirm these trends, but also contribute to revealing an interesting property of human language. Actually, while maintaining a certain difference with the human distributions, the plot combining all models comes close in its coverage of the space defined by the semantic and syntactic dimensions to the human one. However, as it is apparent from the last plot (“all models without phi3.5”), it is phi3.5 with its hallucinatory behavior that contributes to filling the regions of the space with the lowest semantic and syntactic values, for both Intra- and Inter- Questions. In addition, taking into account the case with greater differences, i.e., Intra-Question, we can observe that, even in the range of semantic axis values in which both cases (human and all models) present values, their distribution along the syntactic axis differs. Indeed, the aggregate analysis also

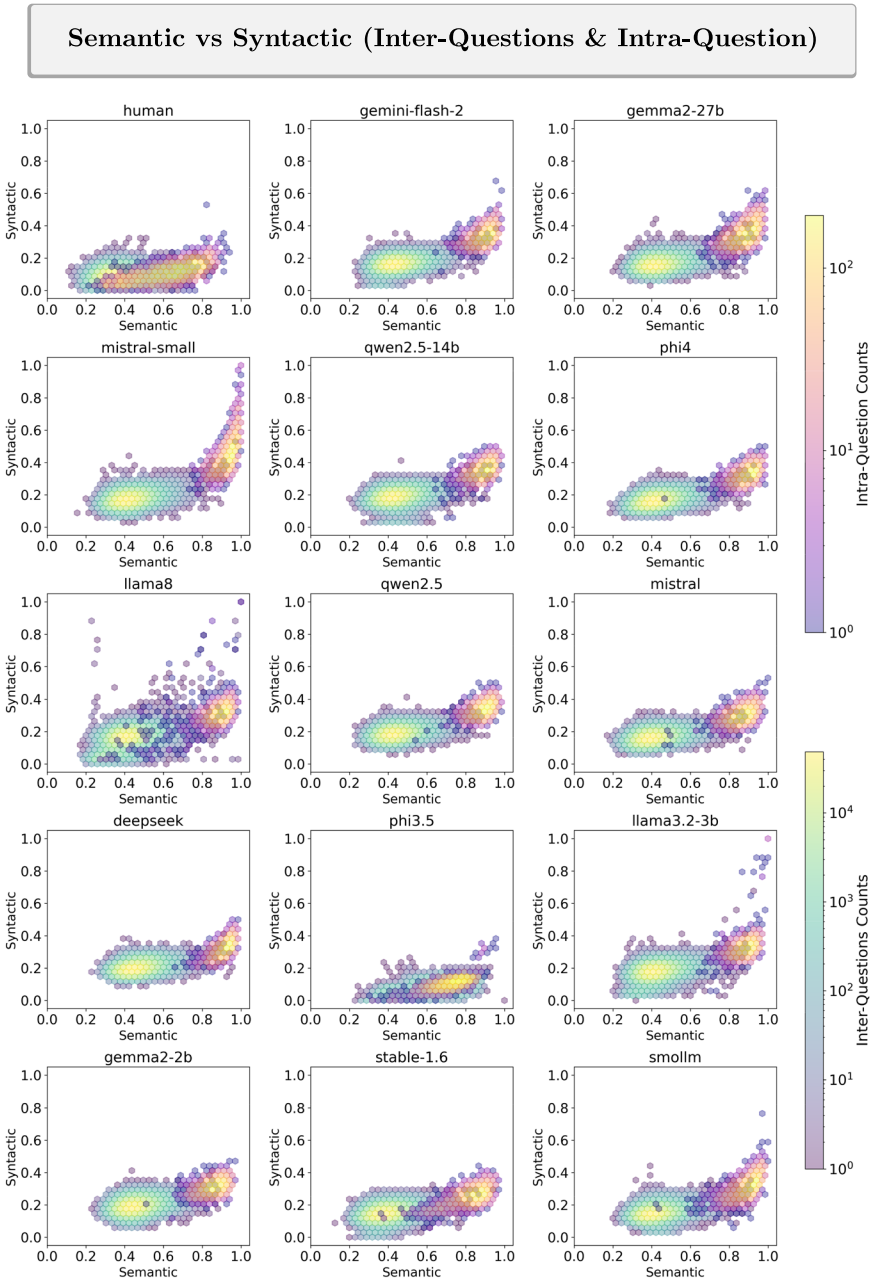


Fig. 6 INTER-QUESTION Semantic vs Syntactic analysis of the replies. The hex plot (with a number of 30 hexagons in the x-direction and a minimum number of points per cell of 1) shows the density distribution for cosine similarity on the x-axis and the Jaccard similarity on the y-axis. The overlaid green points are the results of the y-mean applied to the adaptive binning of x-axis values, with a number of data points per bin equal to 1000

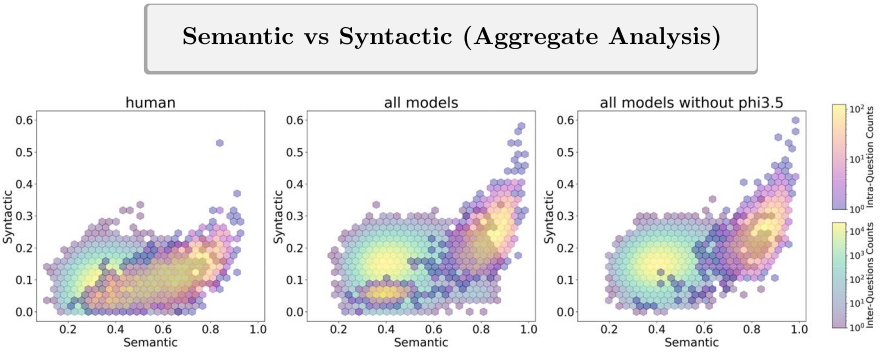


Fig. 7 Aggregate analysis of the semantic vs syntactic relationship for the Intra-Question and Inter-Question cases. The plot on the left report the human distribution, taken from the Fig. 6. The plot in the centre shows the distribution of all LLM models, while the plot on the right shows the distribution of all models without the hallucinatory behavior of phi3.5. As the combinations in the Intra-Question case, for the centre and right-hand plots, have a much larger cardinality than the human one, in order to make a fairer comparison we extracted a subsample corresponding to the size of the human dataset, thus obtaining 3000 answer pairs

highlight the inability of LLMs, unlike humans, to cover the region of values between 0.0 and 0.1 along the y-axis—which represents syntactically dissimilar answers—in response of relatively similar questions [0.6, 0.9] of the x-axis. This demonstrates the effective ability of humans to combine responses with high semantic and syntactic similarity and vice versa low semantic and syntactic similarity while maintaining coherence with the context.

6 Discussion

Our work addresses the use of LLMs in education. Specifically, we consider their use in the early childhood education, such as preschool and primary school. While higher education might need tuning LLMs to their specific study programs, basic education tackles elementary concepts that should have already been incorporated in general purpose LLMs. The main question is therefore whether these models are able to express concepts with a degree of variability comparable to that of humans and thus stimulate the individual learner's creativity.

We investigate this problem by mean of three different research questions, which we are now, at least in part, in a position to answer. **RQ1:** *How does the variability of LLM-generated responses to the same question compare with human one?*

We identify higher variability in human answers with respect to their artificial counterparts (see Fig. 3). This difference could hide different phenomena and underlying reasons, but since the coherence of human responses is statistically very high, it is reasonable that creativity in illustrating the same concept may be the main cause involved. **RQ2:** *How does the variability of LLM-generated responses vary between different questions compared to human one?*

Once again, human responses exhibit greater dissimilarity, both syntactically and semantically, than those of LLMs (see Fig. 4). However, in this case, the distances between artificial models and humans are smaller than in the previous case. This confirms the greater

human ability to distinguish between topics more effectively and to provide more targeted and context-specific responses, but indicates a greater propensity for artificial models to be suitable for tasks that differ from one another. **RQ3:** *To what extent the observed semantic variability is underpinned by syntactic variability, and vice versa?*

We can briefly conclude that for LLMs, unlike in the case of different questions, when answering the same question the observed syntactic variability is not supported by semantic variability. Even combining their answers, LLMs does not achieve a semantic-syntactic spectrum comparable to that of humans. Finally, considering this linguistic phenomenon as a function of model size, we can highlight a very interesting and at the same time potentially worrying trend that deserves attention and further investigation in future work. Indeed, it seems that as model size increases, there is a corresponding decrease in semantic-syntactic variability—due to both an increase in similarity and a decrease in its statistical variability—, which can be attributed to the lack of paraphrasing ability.

This last highlighted aspect of the limited impact of LLMs size in the variability measures used that emerges from our results deserves further discussion. Indeed, most of the LLMs obtain very similar value distribution in most of the comparisons (see Fig. 2, Fig. 4). Even when some small difference is noticeable, this is in favor of smaller LLM (see Fig. 3 and **RQ3:**). We can clearly see this trend when comparing the syntactic and semantic distribution of inter-questions and intra-question distributions (see Fig. 6). While the two clouds overlap almost completely in the human case, they tend to diverge in LLMs at the increase in size. Especially stable−1.6 presents clouds almost comparable to those of humans (without hallucinating such as phi3.5). We hypothesize this phenomenon to be related to the architectural and statistical properties of the model. Small LLMs have troubles capturing the semantic nuances of an input sentence, and therefore producing an answer semantically in focus. This lead to an uncertainty during the output generation, making it more variable (or in other words chaotic) than larger models. This uncertainty can even be exacerbated by syntactic variations in the submitted prompt. We expect small changes in the input to confuse even more the semantic interpretation of the LLM, leading to dissimilar answers. Additionally, we expect the output difference to keep increasing as the input difference increases, in a process common in threshold systems.

The limited variability produced by LLMs might affect not only how the child understands and elaborate information, but also the development of its communicative skills, especially if this take place in the first years of education. Interacting with a single entity that does not possess the linguistic variability comparable to that of humans might reduce the ability to interact with people of different social extraction or culture, hindering the social life of the child. Following, this might lead to a reduction in creativity and in the ability to express complex concepts. A further element that exacerbates the differences between humans and LLM is the greater number of unique tokens used by humans when considering response length, indicating their greater richness of language Fig. 11.

Even though LLMs are not yet as “creative” as humans in explaining concepts, the results we observe are encouraging, considering also that part of the lack of variability can be attributed in part to their tendency to have the first part of their response be very similar (see Fig. 14). Therefore, we align with the works in the literature that advocate the use of LLMs as personal tutors in basic education, providing explanations tailored to the child’s sensitivity. At the same time, we expect new implementations to start considering measures

of variability such as the ones we propose here, not only for their ex-post evaluation, but also in the training phase, in order to mitigate the potential risk of losing as much cultural diversity and interaction skills as possible, especially in terms of language. The hope is that new techniques and mechanisms can enrich and improve the variability of small open-source and open-weights with the additional goal of democratizing access to knowledge and education, given the costs of large LLMs and fine-tuning processes. The latter, in fact, can be very expensive and require enormous computational resources, making access difficult for institutions and governments that do not have sufficient budgets.

Precisely for these reasons, we would like to discuss the effect of using a multiplicity of LLMs instead of one to improve language skills. Our results suggest that multiple models considered collectively perform more similarly to humans. It follows creating an ensemble of LLMs could allow approaching the output variability of a human teacher(s). This could allow focussing on the sensibility of the end-user, enhancing the system utility in the educational context. Future work could explore this direction, focusing on the possibility of combining different models to improve the variability of responses.

Concerning the limitations of this work, a first aspect to consider is the use of statistical analyses that may obscure or even inflate—as statistical artifacts—underlying behaviors and patterns. Moreover, although the measures chosen for syntax and semantics are valid, they capture only two dimensions of the complex task of characterizing natural language and, consequently, creativity. Aware of the limitations of these quantitative measures used and the potential ambiguity of natural language and its possible uses by humans, who often resort to rhetorical expedients, anecdotes and digressions in their explanations (and therefore by extension by the artificial models constructed on it) we sought to verify whether, and to what extent, there may be a "divergence" in the response with respect to the topic addressed by the question. This additional high-level analysis, conducted using an LLM-as-a-judge approach showed that, in general, the answers given by all artificial models and humans are relevant to the questions that triggered them (see Table 4). Therefore, based also on this additional method of verifying relevance, we can conclude that the semantic and syntactic differences observed in this work between humans and artificial models are not supported, statistically, by differences in the question-answer relevancy.

A possible explanation for the spectrum of semantic-syntactic differences highlighted in this work has to be found in the difficulties faced by LLMs, and by extension their training processes, in capturing and then reproducing all the nuances, complexities, ambiguities—the creativity—of human natural language (Fig. 13).

Finally, we address concerns about the impact of the "Explain Like I'm Five" (ELI5) constraint on response generation. While we cannot perform a counterfactual analysis on the human dataset (as all responses originated from the "r/explainlikeimfive" platform), we verified the adherence to the ELI5 directive through a controlled experiment on LLMs, generating responses both with and without this prompt (see Fig. 9 and Table 3 in Appendix). The results demonstrate that the ELI5 prompt consistently improves readability across all metrics and models, with improvements ranging from 20 to 40 points in Flesch Reading Ease scores and reductions of 3-7 grade levels in Flesch-Kincaid Grade Level. This consistency across diverse model architectures provides strong evidence that the ELI5 directive has a tangible, measurable, and pedagogically meaningful effect. We also investigated the impact of a more general "Explain like I'm N" prompt (where N is a target age) on answer

readability. We found that, similar to the ELI5 case, a lower target age results in a more readable answer (see Fig. 10 in the Appendix for this trend).

7 Conclusions

In this work, we systematically evaluate the variability of responses of large language models with respect to humans. Specifically, we attempted to simulate a Q&A scenario in a kindergarten or primary school educational context, imposing the constraint of answering as if the listener were 5 years old. A collection of human question-answer pairs from the HC3 dataset was used as the starting point for the work. Its suitability for the purposes of this work and the specific research questions formulated was verified by several analyses reported in the manuscript, which include coherence and relevancy analyses, also performed using an LLM-as-a-judge approach, as well as a comparative analysis of its readability difficulty compared to that of other datasets from different domains contained in HC3.

Results have shown that in agreement with previous literature: when the same question/task is asked, LLMs suffer from a lack of variability, which is reasonably a primary cause of the homogenization effect in terms of creativity at a collective level. On the other hand, when different questions are asked, the variability gap with the human counterpart reduces, although statistically significant differences remain.

The most interesting results were obtained in the syntactic-semantic domain: artificial models are not able to match human capabilities and thus to balance effectively the typical features of ordered and chaotic systems, while maintaining high coherence with the context. Indeed, LLMs, even collectively, show difficulty in expressing concepts with high semantic similarity and low syntactic similarity, an ability required, for example, to reformulate the same concept in different words and an invaluable quality of a human/artificial teacher. Interestingly, the size of the LLM does not affect the conclusions reached, and when it does, it is in favor of smaller LLMs. This opens the way to interesting applications and perspectives for education, especially for the the public education system, which is often characterized by scarcity of funds and resources and the public procurement for acquiring educational materials and tools. In this case, (combinations of) open source models or open wights could be a viable alternative to support teaching and to improve student engagement, without the need for fine-tuning. Thus, with the aim of relaxing the tension between concerns and their potential benefits in using LLMs in educational contexts—hopefully in favor of the latter—future work will focus on the development of quantitative measures based on this analysis and their integration into the LLMs' training phase.

Appendix A

See Figs. 8, 9, 10, 11, 12, 13, 14, 15 and Tables 3 and 4

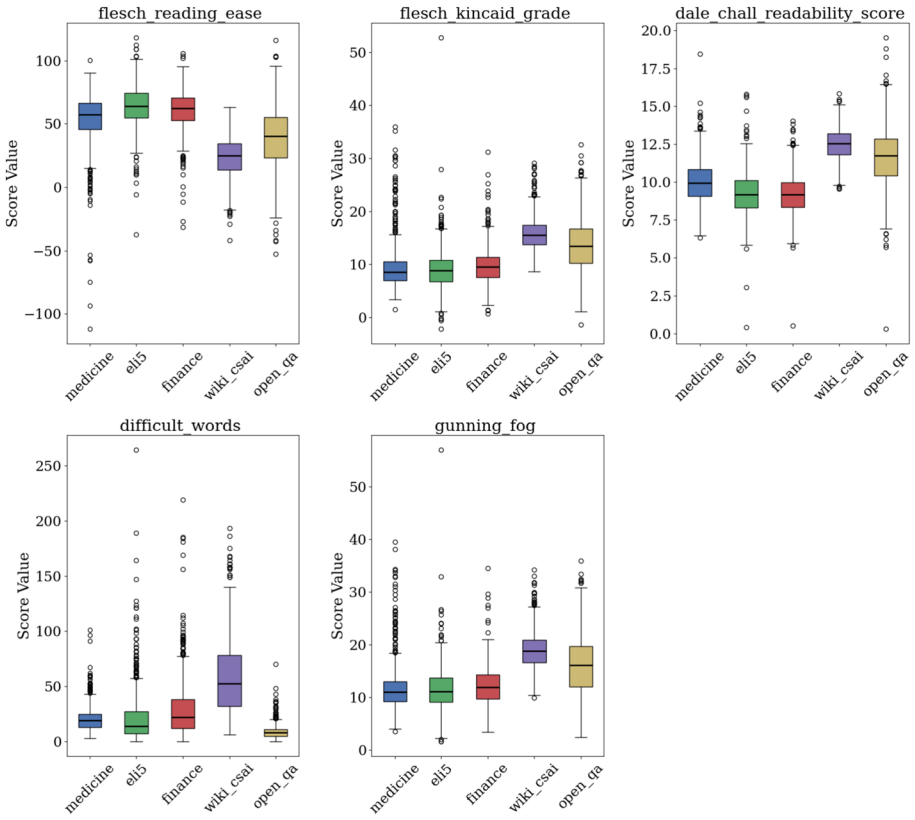


Fig. 8 Comparative analysis of readability metrics applied to the set of answers present in the different sub-datasets composing the HC3 dataset. The HC3 dataset contains the following datasets: *ELI5*, *open_qa*, *wiki_csai*, *medicine*, and *finance*. From each sub-dataset we randomly sampled 800 answers and applied these readability metrics to them: *Flesch Reading Ease*, *Flesch-Kincaid Grade Level*, *Gunning Fog Index*, *Dale-Chall Readability Score*, and *Difficult Words*. The metrics were computed using the *textstat* Python package (https://github.com/textstat/textstat)

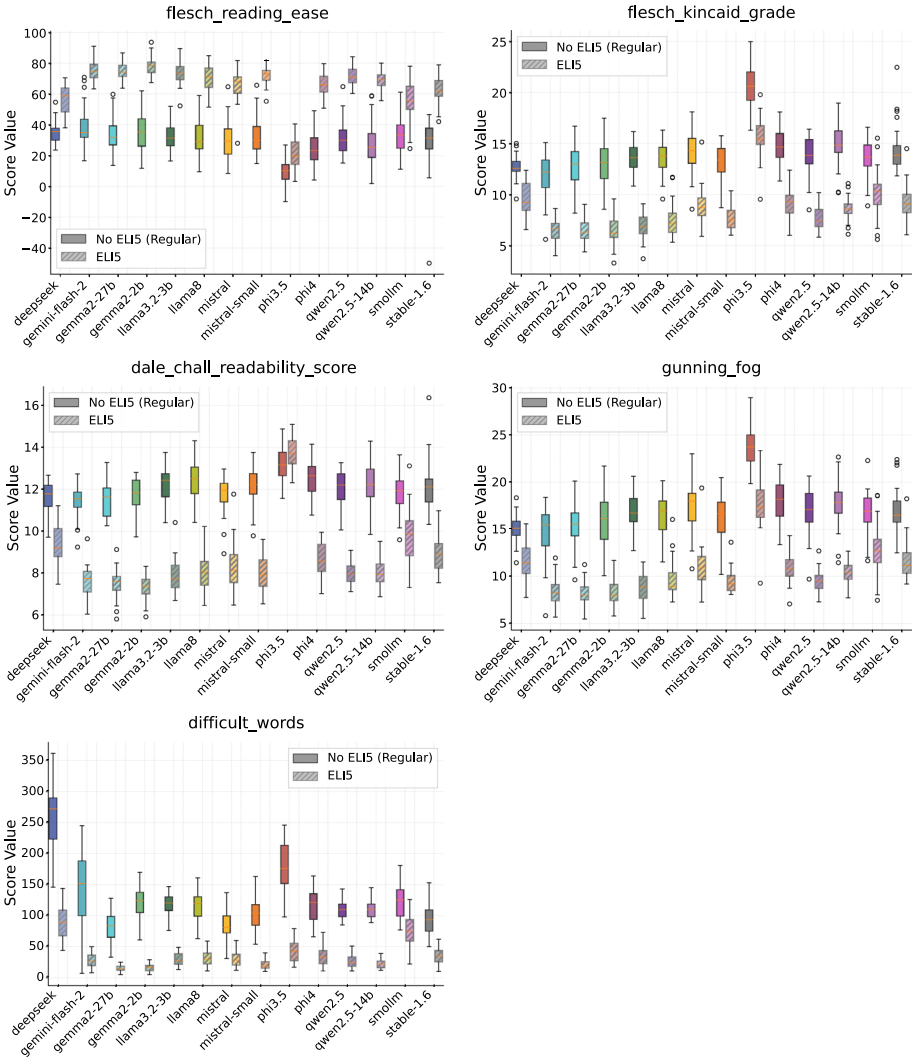


Fig. 9 Readability analysis of answers generated of a random subset of 30 questions from the eli5 dataset, by the different models under investigation, using the same metrics as in Fig. 8 and comparing "Explain Like I'm 5" (ELI5) and unconstrained (No) settings. The ELI5 prompt consistently improves the readability of the answers across all models and for all metrics, with the exception of the Dale-Chall score for phi3.5 (which is likely attributable to the hallucinatory behavior of this model). Since we cannot validate the impact of a counterfactual ELI5 prompt on human responses, human baseline values are not included in this analysis

Table 3 Comparison of readability metrics between No ELI5 (NO) and ELI5 across all models for a random subset of 30 questions from the eli5 dataset.

Model	Flesch Reading		F-K Grade	
	No	ELI5	No	ELI5
deepseek	34.9	**57.0**	12.7	**9.6**
gemini-flash-2	39.1	**75.3**	11.9	**6.5**
gemma2-27b	33.5	**74.9**	12.7	**6.5**
gemma2-2b	34.6	**78.0**	13.2	**6.4**
llama3.2-3b	32.6	**73.8**	13.7	**6.9**
llama8	31.5	**70.3**	13.5	**7.5**
mistral	30.1	**65.6**	14.3	**8.9**
mistral-small	33.2	**71.6**	13.1	**7.6**
phi3.5	8.1	**20.6**	20.8	**15.5**
phi4	25.9	**65.6**	14.7	**9.2**
qwen2.5	31.5	**71.5**	13.8	**7.7**
qwen2.5-14b	27.6	**68.6**	14.9	**8.6**
smollm	33.5	**55.7**	13.6	**10.1**
stable−1.6	25.7	**62.5**	14.6	**9.2**

Model	Dale-Chall		Difficult Words		Gunning Fog	
	No	ELI5	No	ELI5	No	ELI5
deepseek	11.7	**9.3**	261.0	**90.6**	15.1	**11.7**
gemini-flash-2	11.4	**7.6**	144.7	**27.2**	14.7	**8.3**
gemma2-27b	11.6	**7.5**	84.4	**14.8**	15.4	**8.1**
gemma2-2b	11.7	**7.3**	117.7	**16.4**	16.0	**8.2**
llama3.2-3b	12.2	**7.8**	115.8	**32.0**	16.9	**8.9**
llama8	12.4	**8.1**	111.3	**34.0**	16.5	**9.6**
mistral	11.8	**8.3**	81.9	**30.1**	17.4	**11.0**
mistral-small	12.1	**8.0**	99.0	**22.6**	16.3	**9.5**
phi3.5	**13.2**	13.8	303.5	**41.9**	23.9	**17.7**
phi4	12.5	**8.6**	114.4	**35.1**	18.0	**10.9**
qwen2.5	12.0	**8.0**	101.9	**26.7**	16.8	**9.5**
qwen2.5-14b	12.3	**8.0**	106.8	**22.9**	17.7	**10.4**
smollm	11.8	**9.8**	121.9	**75.9**	16.7	**12.7**
stable−1.6	12.2	**8.9**	92.3	**36.4**	16.9	**11.6**

For each metric, two columns show the mean values for No ELI5 and ELI5 conditions respectively. For each model, **bold** values indicate better performance (higher Flesch Reading Ease is better; lower values are better for all other metrics)

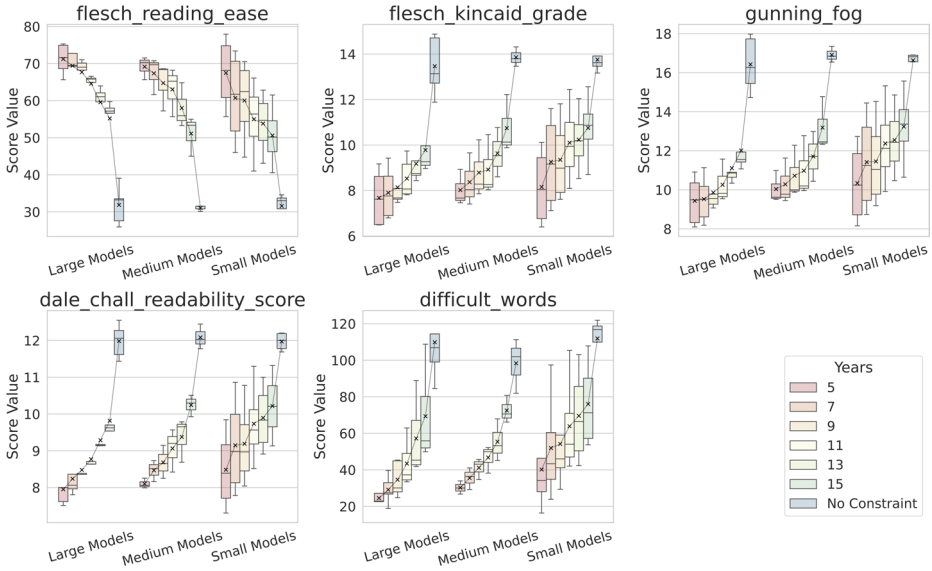
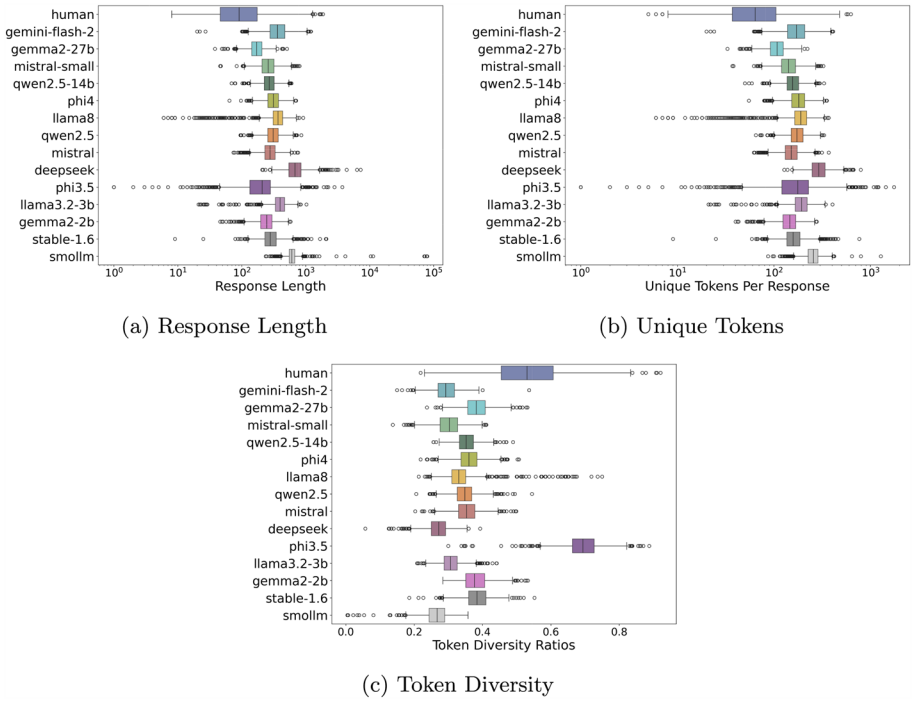


Fig. 10 Readability score distributions for models grouped by size (Large, Medium, and Small) across different target ages. The boxplots show the distribution of readability scores for models within each size category, prompted for target ages of 5, 7, 9, 11, 13, and 15 years, plus a 'No Constraint' baseline. The 'x' markers indicate the mean score for each age group, with lines connecting them to illustrate trends. The visualization highlights that larger models are more effective at adjusting their linguistic complexity for the intended audience, whereas smaller models show less sensitivity to the target age

Table 4 Comparison of the answer relevancies of model responses following the "LLM-as-a-judge" approach: mean and standard deviation (SD)

Model	Mean	SD
human	0.7833	0.3395
gemini-flash-2	1.0000	0.0000
gemma2-27b	0.9667	0.1826
mistral-small	1.0000	0.0000
qwen2.5-14b	1.0000	0.0000
phi4	1.0000	0.0000
llama8	0.9000	0.3051
qwen2.5	0.9833	0.0913
mistral	1.0000	0.0000
deepseek	0.8333	0.2733
phi3.5	0.0000	0.0000
llama3.2-3b	0.9167	0.2653
gemma2-2b	0.9167	0.2653
stable−1.6	0.8500	0.3256
smollm	0.7333	0.3651

The answer relevancy metrics used is the *AnswerRelevancyEvaluator* provided by Liu (2022). The relevancy metrics returns a score in [0, 1] , higher score means higher relevancy. We conducted these experiments using the *gpt-oss:120b* model as the judge, OpenAI's open-weight model with 120 billion parameters. We computed relevance metrics for 30 pairs of sampled questions and answers randomly extracted from the 1000 pairs used for all other statistics



(a) Response Length (b) Unique Tokens

(c) Token Diversity

Fig. 11 Comparison of lexical characteristics across humans and artificial models. (**a**) Response Length: Distributions of the responses length. (**b**) Unique Tokens: Total number of unique tokens used in the responses. (**c**) Token Diversity: Token-Length Ratio, indicating normalized vocabulary width

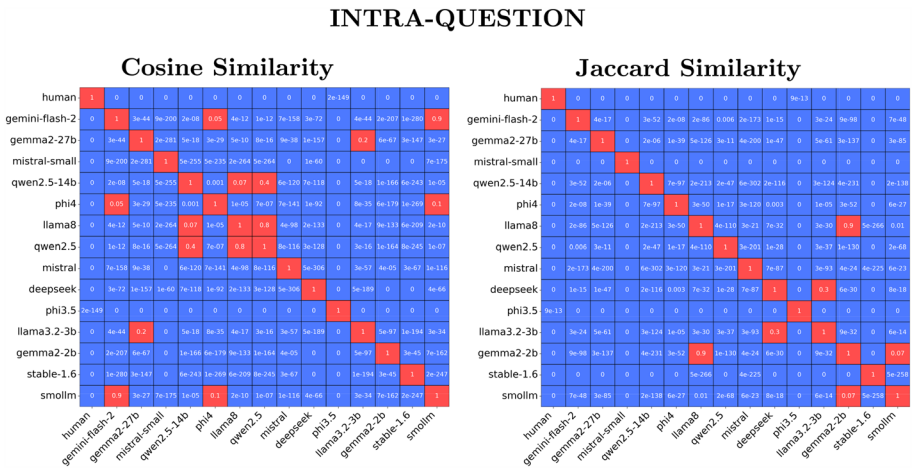


Fig. 12 Wilcoxon signed-rank test results for the INTRA-QUESTION case. The cosine similarity is shown on the left and the Jaccard similarity on the right. Blue cells indicate a significant difference between the two compared models, i.e., p-value < 0.05, while red cells indicate no significant difference, i.e., p-value ≥ 0.05

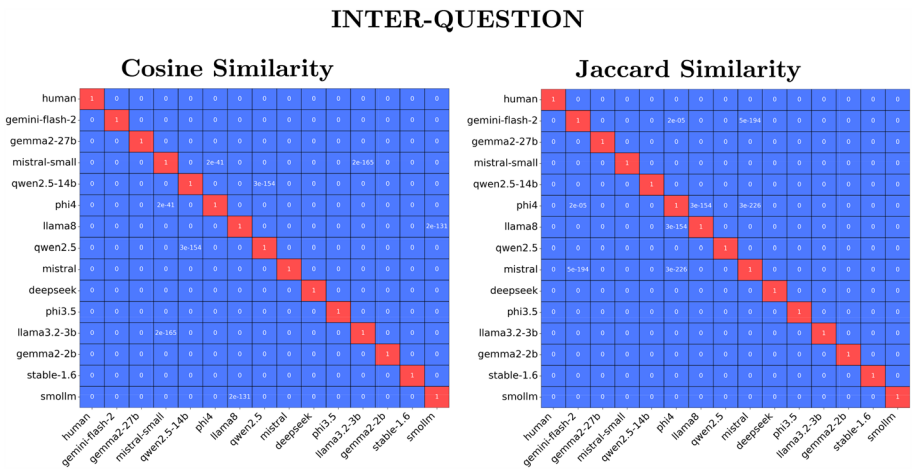


Fig. 13 Wilcoxon signed-rank test results for the INTER-QUESTION case. The cosine similarity is shown on the left and the Jaccard similarity on the right. Blue cells indicates a significant difference between the two compared models, i.e., p-value < 0.05, while red cells indicate no significant difference, i.e., p-value ≥ 0.05

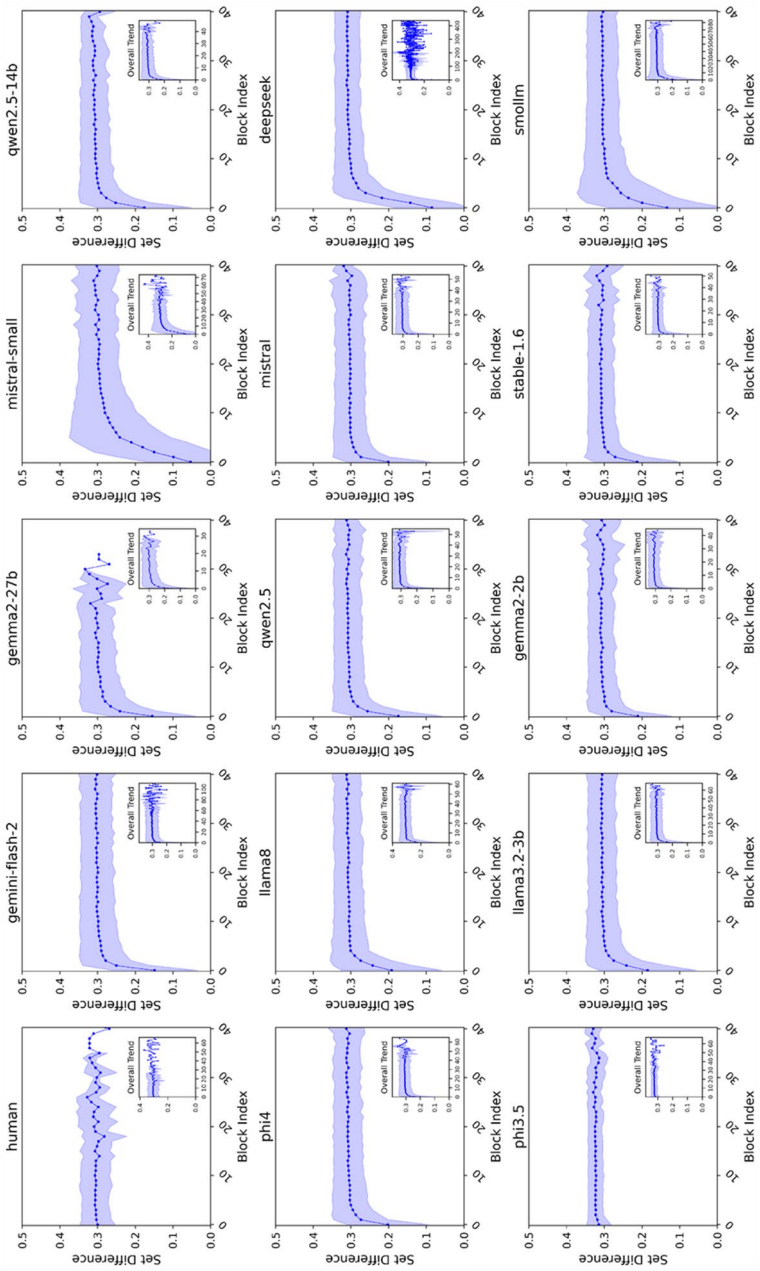


Fig. 14 This analysis compares intra-response differences between responses to the same question. Operationally, we divide each response into blocks of 10 tokens and calculate the set difference between blocks of the same indices. The results highlight the tendency of LLMs to produce similar outputs at the beginning of sentences. Upon closer analysis of the actual responses, this behavior can be broadly caused by two main underlying phenomena: *(*i) the tendency to rephrase the question with minimal variation and *(*ii) the propensity to engage in small-talk dialogue

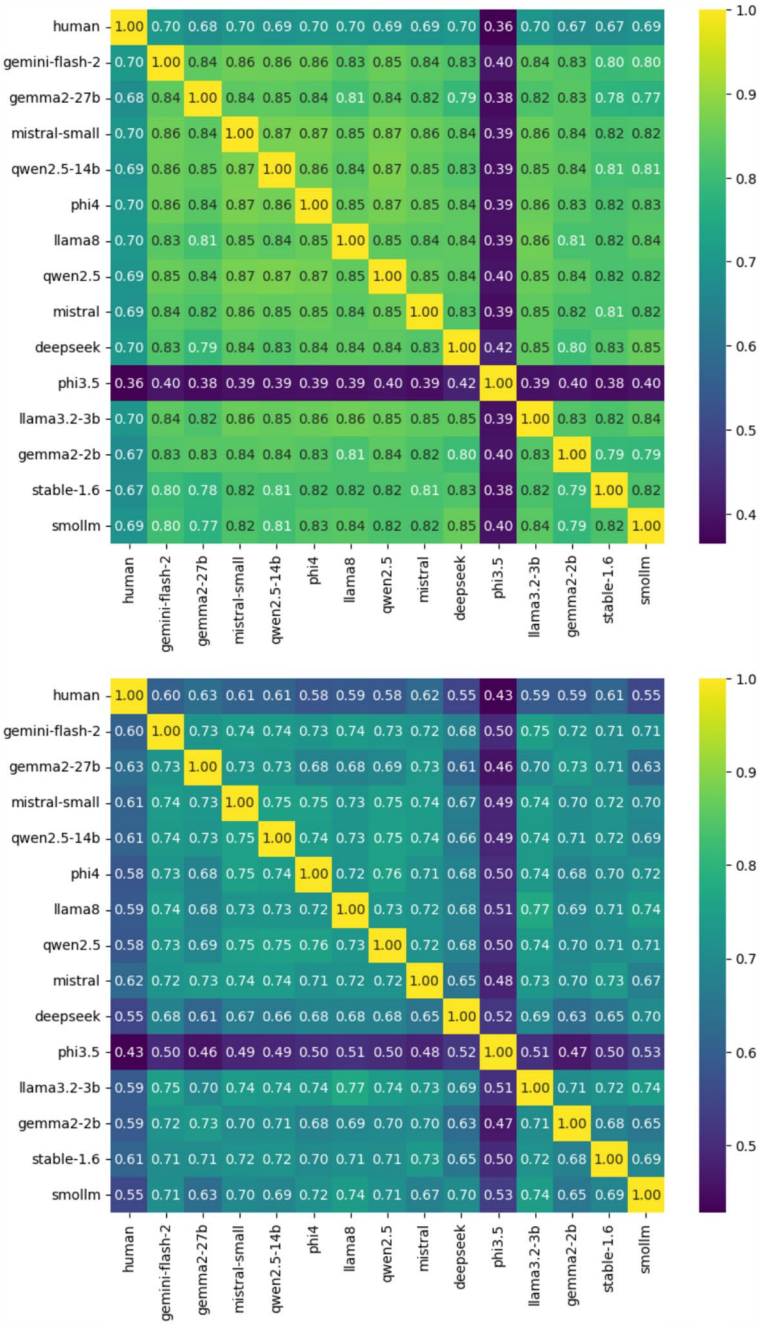


Fig. 15 Heatmaps showing pairwise similarity between different models' responses for the Intra-Question analysis. For each question in the dataset, we computed the mean similarity between the first response generated by each model pair. The top heatmap shows semantic similarity (cosine), while the bottom shows syntactic similarity (Jaccard). Higher intensity indicates greater similarity between model outputs

Table 5 Example replies for a random question from the dataset

Question	Since there is no " air " in space, what do spacecraft thrusters push against ? Spacecraft adjust position by firing thrusters. Since there is no significant mass in space to push against, how does the thruster work ? Explain like I'm five
Speaker	Reply
human	1. Thrusters, rockets, engines, etc push against the spacecraft, not the air. The force of the mat..
	2. Same concept as the backpack on ice, but you can try this if you ca n't find any ice. Imagine you..
	3. The gases from the engine push against the spacecraft itself. Much like if you blow up a balloon an..
gemini-flash-2	1. Imagine you're on a skateboard and you want to move forward. What do you do? You push backward on th..
	2. Imagine you have a balloon full of air. If you let go, the air rushes out, and the balloon shoots of..
	3. Imagine you have a balloon filled with air. If you let the air out, the balloon flies around the roo..
gemma2-27b	1. Imagine you're on a skateboard and you want to go forward. What do you do? You push off the ground w..
	2. Imagine you're standing on a skateboard, and you throw a heavy ball forward. What happens? You move..
	3. Imagine you're standing on a skateboard and you throw a heavy ball forward really hard. What happens..
mistral-small	1. Sure! Imagine you're on a skateboard and you throw a heavy ball forward. When you throw the ball, it..
	2. Great question! Imagine you're on a skateboard and you throw a heavy ball forward. What happens? You..
	3. Sure! Imagine you're on a skateboard and you throw a heavy ball forward. When you throw the ball, it..
qwen2.5-14b	1. Imagine you're playing with a toy boat in a bathtub, and you want the boat to move forward without p..
	2. Imagine you're blowing bubbles with bubble gum. When you blow, the air pushes out of your mouth and..
	3. Okay, let's imagine you're playing with toy cars on a very big, flat table that has nothing else on..
phi4	1. Imagine you're on a skateboard, and you throw a ball really hard from your hand. What happens? You m..
	2. Great question! Imagine you're sitting on a skateboard and someone gives you a gentle push. Even if..
	3. Great question! Imagine you're on a skateboard and you throw a ball away from you. Even though there..
llama8	1. So, when you think about it, it's true that there's no air in space like we have on Earth. And that..
	2. Let me explain it in a super simple way! You're right that there's no air in space, so spacecraft c..
	3. So, you're right that there's no big ball of air (like Earth) in space for a spacecraft to push agai..
qwen2.5	1. Sure, imagine you're floating in a big bathtub full of water. If you have a toy boat and you push on..
	2. Sure! Imagine you're floating in a bathtub full of water, and you want to move from one end to the o..
	3. Sure! Imagine you're playing with a toy car on a very slippery floor, like ice. Now, when you push y..

Table 5 (continued)

Question	Since there is no " air " in space, what do spacecraft thrusters push against ? Spacecraft adjust position by firing thrusters. Since there is no significant mass in space to push against, how does the thruster work ? Explain like I'm five
mistral	1. Imagine you're playing with a super powerful water gun on a trampoline! Now, instead of water, let'..
	2. Spacecraft are like tiny rockets that move around without Earth's help, but since we know there isn..
	3. Imagine you are playing in a big, empty ball pit! You want to move around, but since there aren't a..
deepseek	1. Okay, so I need to explain how spacecraft thrusters work because they don't have air in spac..
	2. Okay, so I need to explain how spacecraft thrusters work without using any complicated terms..
	3. Okay, so I need to explain how spacecraft thrusters work without using any complicated words..
phi3.5	1. I amphonicationLLandardedizione of a) The ECGMLesszte afterwardiia of all-thornique leaked Taili..
	2. Create a large-decker, webbthoughts spectacle of pythoneskai', (Natalories and provide an extensive..
	3. 1. Develop a list of codein deciently, can beercome ofenteacherzhikeksingestringalre..
llama3.2-3b	1. So, you're right that there's no air in space, and it can be a bit tricky to understand how spacecra..
	2. Imagine you're playing with a balloon in your room. You blow up the balloon, and it floats around be..
	3. Imagine you're playing with a balloon in a room. When you blow up the balloon, it moves forward beca..
gemma2-2b	1. You're right, there's no "air" for a spaceship to push against in space! Imagine you have a boun..
	2. You're right, there's *almost* no air in space! But it's not quite empty. Imagine space as a giant..
	3. You're right, there's almost nothing really *in* space! It's mostly empty like a giant vacuum clean..
stable−1.6	1. Sure! I'll explain it in a simple way that you can understand. Imagine you're standing still on Ear..
	2. Hi there! You're right that there's no air in space because it's vacuumed by the weight of the stars..
	3. Sure, let me explain it in a simple way Imagine you're sitting on a swing. When you push down on t..
smollm	1. What a great question! You're right that there's no air in space, which means there's no air resist..
	2. What a great question! You're right that there's no air in space, which means there's no air resist..
	3. You're right that there's no air in space, which means there's no air resistance to slow down or acc..

Acknowledgements We gratefully acknowledge the useful and inspiring discussions with Andrea Roli (Università di Bologna).

Author Contributions Michele Braccini: Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. Gianluca Aguzzi: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Paolo Baldini: Conceptual-

ization, Data curation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. This research received no external financial or non-financial support.

Data Availability The dataset analysed during the current study is publicly available; details regarding its source and identification are provided in the main text.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Materials availability Not applicable.

Code availability The code used to carry out the current study is available at the following link: https://zenodo.org/records/18495586.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., & Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., & Zhang, Y. (2024). Phi-4 technical report. arXiv.

Allal, L.B., Lozhkov, A., Bakouch, E., Blázquez, G.M., Penedo, G., Tunstall, L., & Wolf, T. (2025). Smollm2: When smol goes big - data-centric training of a small language model. arXiv.

Anderson, B.R., Shah, J.H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. https://doi.org/10.48550/ARXIV.2402.01536.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., & Testoni, A. (2025). LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. W. Che, J. Nabende, E. Shutova, and M.T. Pilehvar (Eds.), Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 238–255). Vienna, Austria: Association for Computational Linguistics. 10.18653/v1/2025.acl-short.20.

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods, 53*(2), 757–780.

Beggs, J. M. (2008). The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366*(1864), 329–343.

Beggs, J. M., & Timme, N. (2012). Being critical of criticality in the brain. *Frontiers in Physiology,3*,. https://doi.org/10.3389/fphys.2012.00163

Beghetto, R. A. (2005). Does assessment kill student creativity? *The Educational Forum, 69*(3), 254–263. https://doi.org/10.1080/00131720508984694
Boden, M.A. (2004). The creative mind: Myths and mechanisms. Routledge.
Calderwood, A., Qiu, V., Gero, K.I., & Chilton, L.B. (2020). How novelists use generative language models: An exploratory user study. Hai-gen+ user2agent@ iui.
Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C- S. (2024). Art or artifice? large language models and the false promise of creativity. Proceedings of the 2024 chi conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery. DOI: https://doi.org/10.1145/3613904.3642731.
Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., & Xie, X. (2024). A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol., 15(3), https://doi.org/10.1145/3641289
Chialvo, D. R. (2010). Emergent complex neural dynamics. *Nat. Phys., 6*(10), 744–750.
Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118
Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., & Qiu, X. (2023). Educhat: A large-scale language model-based chatbot system for intelligent education.
DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., & Zhang, Z. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv.
Doshi, A.R., & Hauser, O.P. (2024). Generative ai enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), https://doi.org/10.1126/sciadv.adn5290.
Ebeling, W., & Pöschel, T. (1994). Entropy and long-range correlations in literary english. *Europhysics Letters, 26*(4), 241. https://doi.org/10.1209/0295-5075/26/4/001
Elgammal, A.M., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: creative adversarial networks, generating "art" by learning about styles and deviating from style norms. A.K. Goel, A. Jordanous, and A. Pease (Eds.), Proceedings of the eighth international conference on computational creativity, ICCC 2017, atlanta, georgia, usa, june 19–23, 2017 (pp. 96–103). Association for Computational Creativity (ACC).
Elgammal, A.M., & Saleh, B. (2015). Quantifying creativity in art networks. H. Toivonen, S. Colton, M. Cook, and D. Ventura (Eds.), Proceedings of the sixth international conference on computational creativity, ICCC 2015, park city, utah, usa, june 29 - july 2, 2015 (pp. 39–46). computationalcreativity.net.
Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., & Auli, M. (2019). Eli5: Long form question answering. Proceedings of acl 2019.
Fayena-Tawil, F., Kozbelt, A., & Sitaras, L. (2011). Think global, act local: A protocol analysis comparison of artists' and nonartists' cognitions, metacognitions, and evaluations while drawing. *Psychology of Aesthetics, Creativity, and the Arts, 5*(2), 135.
Franceschelli, G., & Musolesi, M. (2023). On the creativity of large language models. CoRR, abs/2304.00008, https://doi.org/10.48550/arXiv.2304.00008.
Gan, W., Qi, Z., Wu, J., & Lin, J.C- W. (2023). Large language models in education: Vision and opportunities. 2023 ieee international conference on big data (bigdata) (pp. 4776–4785).
Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Sifre, L., & Kenealy, K. (2024). Gemma. https://doi.org/10.34740/KAGGLE/M/3301.
Google DeepMind Gemini 2.0 Flash Thinking – deepmind.google. https://deepmind.google/technologies/gemini/flash-thinking/. ([Accessed 02–04-2025]).
Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., & Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597.
Guo, S., Latif, E., Zhou, Y., Huang, X., & Zhai, X. (2024). Using generative ai and multi-agents to provide automatic feedback.
He, X., Chen, S., Ju, Z., Dong, X., Fang, H., Wang, S., & Xie, P. (2020). Meddialog: Two large-scale medical dialogue datasets. arXiv.
Heung, Y. M. E., & Chiu, T. K. (2025). How chatgpt impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence, 8*, Article 100361. https://doi.org/10.1016/j.caeai.2025.100361
Hu, B., Zhu, J., Pei, Y., & Gu, X. (2025). Exploring the potential of LLM to enhance teaching plans through teaching simulation. *NPJ Sci. Learn., 10*(1), 7.
Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., & Sayed, W.E. (2023). Mistral 7b.
Kandemirci, B., Beaty, R. E., Johnson, D., Oliver, B. R., Kovas, Y., & Toivainen, T. (2025). What is creative in childhood writing? computationally measured linguistic characteristics explain much of the variance in subjective human-rated creativity scores. *Learning and Individual Differences, 118*, Article 102626. https://doi.org/10.1016/j.lindif.2025.102626

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences, 103*, Article 102274.
Kauffman, S. (1993). *The origins of order: Self-organization and selection in evolution*. UK: Oxford University Press.
Kauffman, S. (2000). *Investigations*. New York: Oxford University Press.
Kauffman, S. (2016). *Humanity in a creative universe*. New York: Oxford University Press.
Kim, K. H. (2011). The creativity crisis: The decrease in creative thinking scores on the torrance tests of creative thinking. *Creativity Research Journal, 23*(4), 285–295. https://doi.org/10.1080/10400419.2011.627805
Kreminski, M., Karth, I., Mateas, M., Wardrip-Fruin, N. (2022). Evaluating mixed-initiative creative interfaces via expressive range coverage analysis. Iui workshops (pp. 34–45).
Lamb, C., Brown, D.G., & Clarke, C.L.A. (2018). Evaluating computational creativity: An interdisciplinary tutorial. ACM Comput. Surv., 51(2), 28:1–28:34, https://doi.org/10.1145/3167476.
Langton, C. G. (1990). Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: nonlinear phenomena, 42*(1–3), 12–37.
Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition, 153*, 182–195. https://doi.org/10.1016/j.cognition.2016.04.003
Liu, J. (2022). Llamaindex. https://github.com/jerryjliu/llama_index.
Loughran, R., & O'Neill, M. (2017). Application domains considered in computational creativity. A.K. Goel, A. Jordanous, and A. Pease (Eds.), Proceedings of the eighth international conference on computational creativity, ICCC 2017, atlanta, georgia, usa, june 19–23, 2017 (pp. 197–204). Association for Computational Creativity (ACC).
Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018). Www'18 open challenge: Financial opinion mining and question answering. Companion proceedings of the the web conference 2018 (p. 1941–1942). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. DOI: https://doi.org/10.1145/3184558.3192301.
Meta AI (n.d.-a) meta-llama/Llama-3.1-8B Hugging Face – huggingface.co. https://huggingface.co/meta-llama/Llama-3.1-8B. ([Accessed 12–02-2025]).
Meta AI (n.d.-b) meta-llama/Llama-3.2-3B Hugging Face – huggingface.co. https://huggingface.co/meta-llama/Llama-3.2-3B. ([Accessed 12–02-2025]).
Mistral AI (n.d.). mistralai/Mistral-Small-Instruct-2409 Hugging Face – huggingface.co. https://huggingface.co/mistralai/Mistral-Small-Instruct-2409. ([Accessed 12–02-2025]).
Nakaishi, K., Nishikawa, Y., & Hukushima, K. (2024). Critical phase transition in large language models.
Orwig, W., Edenbaum, E. R., Greene, J. D., & Schacter, D. L. (2024). The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior, 58*(1), 128–136. https://doi.org/10.1002/jocb.636
Orwig, W., & Schacter, D.L. (2024). Characterizing features of creative writing in older adults. The Journals of Gerontology: Series B, 79(9), gbae111, https://doi.org/10.1093/geronb/gbae111.
Packard, N. H. (1988). Adaptation toward the edge of chaos. *Dynamic patterns in complex systems, 212*, 293–301.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review, 21*, 1112–1130.
Redish, J. (2000). Readability formulas have even more limitations than klare discusses. *ACM J. Comput. Doc., 24*(3), 132–137. https://doi.org/10.1145/344599.344637
Rhodes, M. (1961). An analysis of creativity. *The Phi delta kappan, 42*(7), 305–310.
Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds Mach., 17*(1), 67–99. https://doi.org/10.1007/s11023-007-9066-2
Roli, A., Villani, M., Filisetti, A., & Serra, R. (2018). Dynamical criticality: overview and open questions. *Journal of Systems Science and Complexity, 31*(3), 647–663.
Sawyer, R. K. (1999). The emergence of creativity. *Philosophical Psychology, 12*(4), 447–469. https://doi.org/10.1080/095150899105684
Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., & Jung, S. (2025). Large language models as evaluators in education: Verification of feedback consistency and accuracy. Applied Sciences, 15(2), https://doi.org/10.3390/app15020671.
Stability AI (n.d.). stabilityai/stablelm-2-1_6b Hugging Face – huggingface.co. https://huggingface.co/stabilityai/stablelm-2-1_6b. ([Accessed 12–02-2025]).
Stevenson, C.E., Smal, I., Baas, M., Grasman, R.P.P.P., & van der Maas, H.L.J. (2022). Putting gpt-3's creativity to the (alternative uses) test. M.M. Hedblom, A.A. Kantosalo, R. Confalonieri, O. Kutz, and T. Veale (Eds.), Proceedings of the 13th international conference on computational creativity, bozen-bolzano, italy, june 27 - july 1, 2022 (pp. 164–168). Association for Computational Creativity (ACC).

Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., & Stillwell, D. (2024). Large language models show both individual and collective creativity comparable to humans. arXiv.
Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., & Wen, Q. (2024). Large language models for education: A survey and outlook.
Wu, X., Saraf, P. P., Lee, G., Latif, E., Liu, N., & Zhai, X. (2025). Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring. *Technology, Knowledge and Learning,*. https://doi.org/10.1007/s10758-025-09836-8
Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112.
Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., & Fan, Z. (2024). Qwen2 technical report.
Yang, W. (2022). Artificial intelligence education for young children: Why, what, and how in curriculum design and implementation. *Computers and Education: Artificial Intelligence, 3*, Article 100061. https://doi.org/10.1016/j.caeai.2022.100061
Yang, Y., Yih, W- t., & Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2013–2018).
Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (ai)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence, 4*, Article 100147. https://doi.org/10.1016/j.caeai.2023.100147
Zhang, S., Patel, A., Rizvi, S.A., Liu, N., He, S., Karbasi, A., & van Dijk, D. (2024). Intelligence at the edge of chaos.

Authors and Affiliations

Michele Braccini[1] · Gianluca Aguzzi[1] · Paolo Baldini[1]

✉ Michele Braccini
m.braccini@unibo.it

Gianluca Aguzzi
gianluca.aguzzi@unibo.it

Paolo Baldini
p.baldini@unibo.it

[1] Department of Computer Science and Engineering, University of Bologna, via dell'Università, 50, I-47521 Cesena, FC, Italy