*big data and cognitive computing*

MDPI

*Article*

# Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students

**Katherine Abramski [1,†], Salvatore Citraro [2,†], Luigi Lombardi [3], Giulio Rossetti [2,†] and Massimo Stella [3,*,†]**

[1] Department of Computer Science, University of Pisa, 56127 Pisa, Italy; katherine.abramski@phd.unipi.it
[2] Institute of Information Science and Technologies—National Research Council, 56124 Pisa, Italy; salvatore.citraro@isti.cnr.it (S.C.); giulio.rossetti@isti.cnr.it (G.R.)
[3] Department of Psychology and Cognitive Science, University of Trento, 38122 Trento, Italy; luigi.lombardi@unitn.it
* Correspondence: massimo.stella-1@unitn.it or mass.stella@unitn.it
† These authors contributed equally to this work.

**Abstract:** Large Language Models (LLMs) are becoming increasingly integrated into our lives. Hence, it is important to understand the biases present in their outputs in order to avoid perpetuating harmful stereotypes, which originate in our own flawed ways of thinking. This challenge requires developing new benchmarks and methods for quantifying affective and semantic bias, keeping in mind that LLMs act as psycho-social mirrors that reflect the views and tendencies that are prevalent in society. One such tendency that has harmful negative effects is the global phenomenon of anxiety toward math and STEM subjects. In this study, we introduce a novel application of network science and cognitive psychology to understand biases towards math and STEM fields in LLMs from ChatGPT, such as GPT-3, GPT-3.5, and GPT-4. Specifically, we use behavioral forma mentis networks (BFMNs) to understand how these LLMs frame math and STEM disciplines in relation to other concepts. We use data obtained by probing the three LLMs in a language generation task that has previously been applied to humans. Our findings indicate that LLMs have negative perceptions of math and STEM fields, associating math with negative concepts in 6 cases out of 10. We observe significant differences across OpenAI's models: newer versions (i.e., GPT-4) produce 5× semantically richer, more emotionally polarized perceptions with fewer negative associations compared to older versions and $N = 159$ high-school students. These findings suggest that advances in the architecture of LLMs may lead to increasingly less biased models that could even perhaps someday aid in reducing harmful stereotypes in society rather than perpetuating them.

**Keywords:** cognitive networks; large language models; math anxiety



**Citation:** Abramski, K.; Citraro, S.; Lombardi, L.; Rossetti, G.; Stella, M. Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. *Big Data Cogn. Comput.* **2023**, *7*, 124. https://doi.org/10.3390/bdcc7030124

Academic Editor: Min Chen

Received: 22 May 2023
Revised: 19 June 2023
Accepted: 21 June 2023
Published: 27 June 2023



## 1. Introduction

The introduction of Large Language Models (LLMs) has taken the world by storm, and society's reaction has been anything but unanimous, ranging from humorous amusement to catastrophic fear. Among the most prominent LLMs are OpenAI's GPT-3, GPT-3.5, and GPT-4 (ordered oldest to newest). GPT-3 and GPT-4 are powerful and flexible models that can be fine-tuned to perform a wide variety of natural language processing tasks, while GPT-3.5 turbo is a variant of the other two, specifically designed to perform well in conversational contexts. All three belong to the family of generative pre-trained transformer (GPT) models [1] that are trained on massive amounts of textual data to learn patterns and relationships in text [2,3]. Their power and versatility for accomplishing a range of tasks with incredible human-like finesse have led to a boom in their popularity in society and among researchers across disciplines.

As LLMs secure their role in our lives as useful tools for everyday tasks such as composing emails, writing essays, debugging code, and answering questions, the need to

understand the behavior and risks of these models is ever more important [4–7]. There has been a spike in research dedicated to this topic, surrounded by a debate about the nature of the capabilities of LLMs [8]. Some researchers have suggested that the impressive performance of LLMs on difficult reasoning tasks is indicative of an early version of general artificial intelligence [9]. Many others argue that LLMs exhibit nothing resembling true understanding because they lack a grasp of meaning [10], arguing that they perform well but for the wrong reasons [8]. In fact, much of the success of LLMs at human-like reasoning tasks can be attributed to spurious correlations rather than actual reasoning capabilities [11].

Despite opposing views regarding the nature of intelligence exhibited by LLMs, a relatively undisputed topic is the issue of bias. Bias, in the context of LLMs, has recently been studied as the presence of misrepresentations and distortions of reality that result in favouring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions [12]. While these biases can be influenced by many factors, they largely originate from biases in the massive text corpora on which the models are trained. This can be due to certain groups or ideas being underrepresented in the training data or to implicit biases present in the training data themselves. Thus, the output produced by LLMs inevitably reflects stereotypes and inequalities prevalent in society. This is problematic since exposure through interaction with LLMs could lead to perpetuating existing stereotypes and even the creation of new ones [12,13].

As LLMs become more integrated into our lives, it is even more important to investigate the biases produced by them. This includes understanding our own human biases as well, since LLMs act as "psycho-social mirrors" [14] that reflect human features of personality as well as societal views and tendencies. Thus, it is important to investigate the individual cognitive sphere in conjunction with LLM behavior to understand how our individual and societal tendencies are diffused into the knowledge possessed by artificial intelligence agents. A very natural yet negative human phenomena is affective bias [15], the tendency to prioritize the processing of emotionally negative events compared to positive ones [16]. An example of affective bias is attributing negative attitudes to neutral concepts, such as the attribution of negative perceptions to the neutral concept *math*. These types of biases and stereotypes are inherited by LLMs, adopting perceptions of neutral concepts that deviate significantly from neutrality as a result of our own biased perceptions. It should be the goal of researchers working on developing LLMs to understand such nuanced biases in humans to ensure that LLMs adopt neutral unbiased views of concepts or phenomena that have been historically stigmatized or misrepresented. In doing so, regular widespread interaction with LLMs might actually contribute to a reduction in the harmful biases held by humans.

In this work, we investigate biases produced by LLMs, specifically GPT-3, GPT-3.5, and GPT-4, regarding their perception of academic disciplines, particularly math, science, and other STEM fields. In many societies, these disciplines have a reputation for being difficult [17]. Math in particular, which is arguably the language of science, has been known to cause a great deal of anxiety in many people. This anxiety is a global phenomenon [17,18], and it is deeply rooted, beginning in childhood and persisting throughout adulthood. Unpleasant feelings about math may already begin to develop as early as first grade [19]. Children pick up on the anxieties of their teachers and parents [20], similar to how LLMs absorb biases from training data. Unfortunately, negative perceptions of math have become so commonplace that it is not unusual to hear people identify themselves as not "math people". While this kind of self-categorization may seem harmless, math anxiety can actually have severe individual and societal consequences [17,21–23]. Math anxiety may cause individuals to avoid situations in which math is involved, ultimately having a negative impact on performance. This avoidance tendency may cause bright and capable students to avoid math-intensive classes, determining the course of their academic and professional career [23]. This scales to the societal level. Math anxiety may deter a large portion of the workforce from pursuing careers in STEM, which are in high demand, and

since math anxiety is more prevalent in females as a result of societal stereotypes [24], it may contribute greatly to the gender gap in STEM fields.

Just as children are likely to mirror the math anxiety expressed by their teachers or parents [25], LLMs are "psycho-social mirrors" [8,14], which reflect the tone of the language that we use to talk about math. Thus, we expect to find negative attitudes towards math in large language models. It is critical to investigate the nature of this bias, in order to identify ways to overcome it as AI architectures become more advanced. Crucially, quantitative techniques measuring bias in large language models can provide pivotal ways for better understanding of how such LLMs work and to reduce their negative societal impact when producing text read by massive human audiences. This is particularly impactful for fighting the spread of distorted mindsets in education [26].

To accomplish this, we applied behavioral forma mentis networks (BFMNs) as a method of investigation. BFMNs are a type of cognitive network model that capture how concepts are perceived by individuals or groups by building a network of conceptual associations [27]. This framework, which arises from cognitive psychology coupled with tools from network science, can also be applied to probe LLMs to reveal how they frame concepts related to math, science, and STEM. In this study, we investigated perceptions of these disciplines in three LLMs: GPT-3, GPT-3.5, and GPT-4. A comparison of these models allows us to gain a temporal perspective about how these biases may evolve as subsequent versions of these LLMs are released.

The rest of the paper is organized as follows. In Section 2, we provide a review of recent research dedicated to investigating bias in language models, discussing benchmarks and methods for conducting psychological investigations of LLMs. In Section 3, we describe the framework of BFMNs, and we provide details about data collection, analysis, and visualization. In Section 4, we summarize the results of our investigation of bias towards academic disciplines present in the output from GPT-3, CGPT-3.5, and GPT-4, and in Section 5, we discuss the implications of our findings.

## 2. Review of Recent Literature

Bias has been a significant obstacle to the distributed approach to semantic representation from early on. Since the introduction of word embeddings such as word2vec [28], researchers have been aware that the advantageous operations provided by these models, such as using vector differences to represent semantic relations, are likely to express undesired biases. For example, sexist and racist word analogies such as *"father" is to "doctor" as "mother" is to "nurse"* [29] and *black is to criminal as Caucasian is to police* [30] produced by word embeddings demonstrate how language contains biases that reflect adverse societal stereotypes. Unfortunately, these types of biases are present in tools that we use every day. For example, Google Translate has been found to overrepresent males when translating from gender-neutral languages to English, especially in male-dominated areas such as STEM fields, perpetuating existing gender imbalances [31].

Cutting-edge LLMs such as GPT-3, GPT-3.5, and GPT-4 are not immune to these types of dangers, and the facility of LLMs to simulate human-like language-related competencies, including GPT-3.5's tremendous ability in question-answering and storytelling, makes it necessary to investigate the behavior of LLMs. This has led to the development of new methods and benchmarks for investigating bias that shed light on the variety of demographic and cultural stereotypes and misrepresentations present in the output of language models [12,32].

Gender, racial, and religious stereotypes are among the most widely investigated biases. These biases can be detected in several ways, often by prompting the language model to generate language and then evaluating the output in several ways. One approach involves using Association Tests [13,32–34], which may be performed at different levels of discourse. For example, at the word level, the strength of the association of two words such as *sister* and *science* can be measured [13], providing a simple and intuitive way to measure bias in word embeddings. At the sentence level, the model may be prompted to complete a

sentence such as *girls tend to be more* _____ *than boys*, or to make assumptions following a given context such as *He is an Arab from the Middle East* [32].

Similar approaches have been applied to investigate different types of bias in various LLMs, from BERT and RoBERTa to GPT-3 and GPT-3.5. Persistent anti-Muslim bias has been detected by probing GPT-3 in various ways, including prompt completion, analogical reasoning, and story generation [35]. Topic modeling and sentiment analysis techniques have been used to find gender stereotypes in narratives generated by GPT-3 [36]. Sentiment scores and measurements of "regard" towards a demographic have been applied to assess stereotypes related to gender and sexual orientation in output produced by GPT-2 [37].

While some biases are easier to spot, others are more nuanced [38] and hidden deeply in the architecture of LLMs but also in their training corpus, e.g., training an LLM on students' texts complaining about math might produced a biased model unless additional filtering techniques were implemented externally. Tools from cognitive psychology may be better suited for detecting the subtler dangers of language models where performance-based methods fall short [4,6,7]. For example, one may ask whether a chatbot such as ChatGPT can manifest dangerous psychological traits or personalities when asked if it agrees or disagrees with statements such as *I am not interested in other people's problems* or *I hate being the center of attention* [39]. Such psychological investigations can measure the extent to which LLMs inherently manifest negative personalities and dark connotations, including Machiavellianism and narcissism [39]. Such investigations are an example of the emerging field of "machine psychology" [7], which applies tools from cognitive psychology to investigate the behavior of machines as if they were human participants in psychological experiments. The goal of this new field is to investigate the emergent capabilities of language models where traditional NLP benchmarks are insufficient.

## 3. Methods

Given that our method of investigation can be applied to both humans and LLMs, our approach using behavioral forma mentis networks (BFMNs) can be considered a type of "machine psychology". Combining knowledge structure and affective patterns, forma mentis networks identify how concepts are associated and perceived by individuals or populations. Here, we build BFMNs out of free association data and valence estimates produced by OpenAI's large language models: GPT-3, GPT-3.5, and GPT-4.

BFMNs represent ways of thinking as a cognitive network of interconnected nodes/concepts. Connections/links represent conceptual similarities or relationships. In BFMNs, links indicate memory recall patterns between concepts, which, in this case, are obtained through a free association game. In this cognitive task, an individual is presented with a cue word and asked to generate immediate responses to it, "free" from any detailed correspondence (responses need not be synonyms with the cue word). These free associations represent memory recall patterns, which can be represented as a network. For example, reading *math* may make one think of *number*, so the link (*math*, *number*) is established. In continued free association tasks [40], up to three responses to a given cue can be recorded. Responses are not linked to each other; instead, they are connected only to the cue word. This maximizes the explanatory power that cognitive networks have in terms of explaining variance across a variety of language-processing tasks related to human memory (see [40]). Importantly, BFMNs are feature-rich networks, in that their network structure is enriched by node-level features expressing the valence of each concept, i.e., how positively or negatively a given concept is perceived by an individual or group.

Rather than building BFMNs from responses provided by humans, as carried out in previous works [27,41,42], in this study, BFMNs were constructed out of responses from textual interactions with language models. The same methodology was applied for GPT-3, GPT-3.5, and GPT-4. The resulting networks thus represented how each LLM associates and perceives key concepts related to math, science, and STEM fields based on their responses to the language generation task.

### *3.1. Data Collection: Free Associations and Valence Norms*

As a language generation task, we implemented a continued free association game [40], providing each of the three language models with the following prompt, substituting different cue words:

**Instruction 1.** *Write a list of 3 words that come to your mind when you think of CUE_WORD and rate each word on a scale from 1 (very negative) to 5 (very positive) according to the sentiment the word inspires in you.*

For each prompt, the language model responded by providing 3 textual responses coupled with 3 related numerical responses (valence scores) between 1 and 5. Punctuation and blank spaces were manually removed. In addition to valence scores corresponding to the responses, we also asked each language model to provide a single valence score (independently evaluated) from 1 to 5 for each of the cue words. The language model failed to comply with the instructions only 5% of the time, producing repetitions of the cue word in the response. Those instances were discarded and did not count as repetitions.

In a similar study performed on high-school students [27], there were 159 participants, each providing around 3 responses to each cue word. Therefore, in this study, for comparison purposes, we repeated the above instructions to obtain at least 159 responses for each cue word, matching the number of students who took part in the human study. For GPT-3, we selected the DaVinci model with a temperature of $T = 0.7$, which is the default setting. We used the "vanilla" version of ChatGPT, that is, the default setting to simulate a "neutral" point of view when asking a prompt to the model, without any specific impersonation. Iterations were automated in Python through the API service provided by OpenAI, and the generated text was downloaded and processed in Mathematica. Therefore, we obtained three datasets, one for each of the language models tested, with sample sizes comparable to that of the human dataset from [27]. This enabled interesting comparisons between the recollection patterns of language models and high-school students.

To investigate attitudes towards math, science, and STEM subjects, we tested ten different cue words, corresponding to the same ten key concepts tested in the study with high-school students [27]: *math*, *physics*, *science*, *teacher*, *scientist*, *school*, *biology*, *art*, *chemistry*, and *STEM*. Therefore, the above instructions can be read by substituting *CUE_WORD* with any of these ten key concepts (throughout this paper, we use the terms *key concept* and *cue word* interchangeably).

For each key concept and its associated responses, valence scores (1 through 5) were converted into valence labels (*negative*, *positive*, or *neutral*) using the Kruskall–Wallis non-parametric test (see Section 3.2.1 for details). Thus, valence could be considered categorically rather than numerically.

### *3.2. Network Building and Semantic Frame Reconstruction*

Behavioral forma mentis networks (BFMNs) were constructed such that nodes represented lexical items and edges indicated free associations between words. Following the first part of Instruction 1, we built BFMNs as cognitive networks which simulated human memory recall patterns by linking the cue words to their associative responses. Given the selected cue words and the sets of three responses, our goal was to retrieve a network structure mapping how concepts were connected in the recall process, facilitated by the above instructions (see also [27]).

First of all, associative responses were converted to lowercase letters and checked automatically for common spelling mistakes. The automatic spell checkers used here were the ones implemented in Wolfram's Mathematica 11.3 (manufactured by Wolfram Research, Champaign, IL, USA). Secondly, different word forms were stemmed to reduce the occurrence of multiple word variants that convey the same concept. For stemming words, we used the WordStem command as implemented in Mathematica 11.3.

In the literature about semantic networks, there exist several ways to connect cue words to their associative responses [40,43,44]. We chose to connect each cue word to all

three of its responses, since this method has been shown to provide more heterogeneity in associative responses [44] and has been used in previous works with forma mentis networks [26,27,42]. Moreover, this approach to network construction has been shown to improve the accuracy of many language-related prediction tasks (such as associative strength prediction) compared to other strategies, e.g., connecting the cue word to the first response only [44]. We also considered idiosyncratic associations, i.e., associations provided only one time, which were visually represented as narrower edges compared to non-idiosyncratic associations.

Using the valence labels for the key concepts and associated responses, we enriched the BFMNs, representing them as feature-rich cognitive networks [45] in which information about the sentiment of associative responses could be used to describe the properties of the cue word [27]. As in previous works, we leveraged the notion of a node's neighborhood, consisting of the set of adjacent nodes to a target node: in this case, the neighborhoods of a cue word were the sets of all the associative responses generated by the participants (the language models or humans) responding to the same set of instructions. Inspired by the famous quote *You shall know a word by the company it keeps* [46], which is also the foundation of the distributional semantic hypothesis [47], we could obtain a better understanding of the valence attributed to the cue word by considering the valences of its neighboring associates.

#### 3.2.1. Statistical Analysis of Word Valence

For all key concepts and associated responses, in order to convert numerical valence scores (1 through 5) into categorical valence labels (*negative*, *positive*, or *neutral*) we used a non-parametric statistical test. For each LLM, all valence scores provided for all key concepts and responses were aggregated together. A Kruskall–Wallis test was used to assess whether the scores attributed to concept $w_i$ had a lower, compatible, or higher median valence compared to the entire distribution of valence scores. Non-parametric testing was used because the distribution of valence scores $\bigcup_j w_j$ was mostly skewed with a heavy left tail across all models (Pearson's skewness coefficient $s_s = 3(mean_s - median_s)/\sigma = 1.39$ for students' data and $s_r > 1$ for each language model). Given the relatively small sample size (fixed in order to make suitable comparisons between large language models and humans), and inspired by previous works [27], we fixed a significance level $\alpha = 0.1$, motivated by the aim of detecting more deviations from neutrality despite the contained sample size. Therefore, valence labels were assigned as follows: *negative*—lower median valence score than the rest of the sample; *positive*—higher median valence score than the rest of the sample; *neutral*—same median valence as the rest of the sample.

#### 3.2.2. Data Visualization, Emotional Analysis, and Network Neighborhood Measurements

In our network visualizations, we focused on reproducing the neighborhood of a given target concept, i.e., the associates corresponding to *math*. We rendered valence through colors: positive words were rendered in cyan, negative words in red, and neutral words in black. Idiosyncratic links were rendered with narrower edges compared to associated responses provided more than once. To better highlight clusters of associates, we used a hierarchical edge-bundling layout for network visualization. Because of space issues and to avoid overlap between node labels, we also used a star-graph layout. Both visualizations provide insights into the network structure of associates surrounding a key concept.

In this manuscript, we also used visualizations inspired by the circumplex model of affect [48], which maps individual concepts as points in 2D dimensional space with valence and arousal. According to semantic frame theory [49] and distributional semantics in psycholinguistics [50], each network neighborhood represents a semantic frame indicating ways in which a given concept is associated with others. Hence, understanding the distributions of valence and arousal scores attributed to associates in a given neighborhood provides crucial insights to better understand how key concepts are perceived by a LLM or by a group of individuals [27,51]. For instance, in order to better understand the emotional content of the BFMN neighborhood surrounding *math*, we can plot the 2D density plot

for valence–arousal scores attributed to all words in the neighborhood, and then observe where associate words tend to cluster within the circumplex. We based these investigations on valence–arousal scores obtained by the National Research Canada Valence–Arousal–Dominance lexicon [52].

Last but not least, we compared network neighborhoods, also called semantic frames, across large language models and humans. We measured the following aspects of a frame for each key concept *K* across LLMs and high-school students: (1) semantic frame size, i.e., the number of unique associates in the semantic frame; (2) estimated valence, i.e., the arithmetic mean of the valence scores attributed to *K*; (3) estimated frame valence, i.e., the mode of the valence labels attributed to the associates of *K*; (4) the fractions of positive/neutral/negative words present in the frame; (5) the fraction of non-emotional words present in the frame, i.e., the fraction of words that did not elicit any emotion (according to an emotion–word associative thesaurus [53]) and could, thus, be considered as neutral domain–knowledge or technical associates to a key concept; and (6) the fraction of positive/negative/neutral non-emotional words present in the frame.

## 4. Results

This section outlines the key results we achieved in our interactions with large language models. To begin, we focus on the results from GPT-3 and GPT-3.5. We start with an overall analysis of the valence patterns corresponding to each key concept for both LLMs. We then continue with a detailed analysis of the semantic frames surrounding each key concept, including an investigation of the content and valence of the associates, adopting an approach that uses a circumplex model of affect. Finally, we compare the results from GPT-3 and GPT-3.5 with results from GPT-4 to gain a better understanding of how LLMs are evolving as subsequent versions are released.

### *4.1. Semantic Frames of STEM Concepts Produced by GPT-3 and GPT-3.5*

As discussed in the previous section, the LLMs were prompted to assign valence scores to all cue words and associated responses, and those valence scores were then converted to valence labels: *negative*, *positive*, or *neutral*. Figure 1 reports the fraction of negative (red), positive (cyan), and neutral (black) associated responses comprising the semantic frames of all 10 cue words provided to GPT-3.5 and GPT-3. Cue words are reported at the bottom of each bar chart and colored according to the valence scores produced by each LLM. Notice that this coloring was independent of the valence polarity of the cue word's associates. The most frequent valence label in a semantic frame represents a connotation, also called a valence aura in [27], not to be confused with a valence label. A valence label depends only on the scores attributed to that specific concept, while a valence aura/connotation depends on the valence labels of all the associates of that concept. Hence, whereas valence labels are applied only to individual concepts in isolation, valence connotations include information about network structure and, thus, constitute additional information about how a concept was perceived (see Section 3.2).

In terms of valence labels, as reported in Figure 1 (top), GPT-3 did not identify any cue words as positive. The concepts *math*, *teacher*, and *school* were all identified as negative concepts. Furthermore, *math* and *school* were surrounded predominantly by other negative concepts, i.e., they had negative semantic frames [27]. The semantic frames provided by GPT-3 for *physics*, *chemistry*, and *teacher* were highly polarized in terms of containing similar proportions of positive and negative associates, while *art* and *scientist* were associated mostly with neutral jargon (68% and 52%, respectively). These patterns indicate a non-negligible amount of negative associations provided for *math* and other academic concepts.

As reported in Figure 1 (bottom), GPT-3.5 identified four of the ten cue words as positive concepts, and it provided considerably fewer negative semantic frames compared to GPT-3 overall. While both LLMs perceived *math* negatively, the semantic frame of *math* produced by GPT-3.5 contained 10% fewer negative associates compared to that produced by GPT-3. Valence polarization was present in the semantic frames of *physics*, *science*,

and *chemistry* for GPT-3.5. Notably, the number of negative associates corresponding to the concept *STEM* provided by GPT-3.5 was nearly half of those provided by GPT-3 (24% vs. 15%).

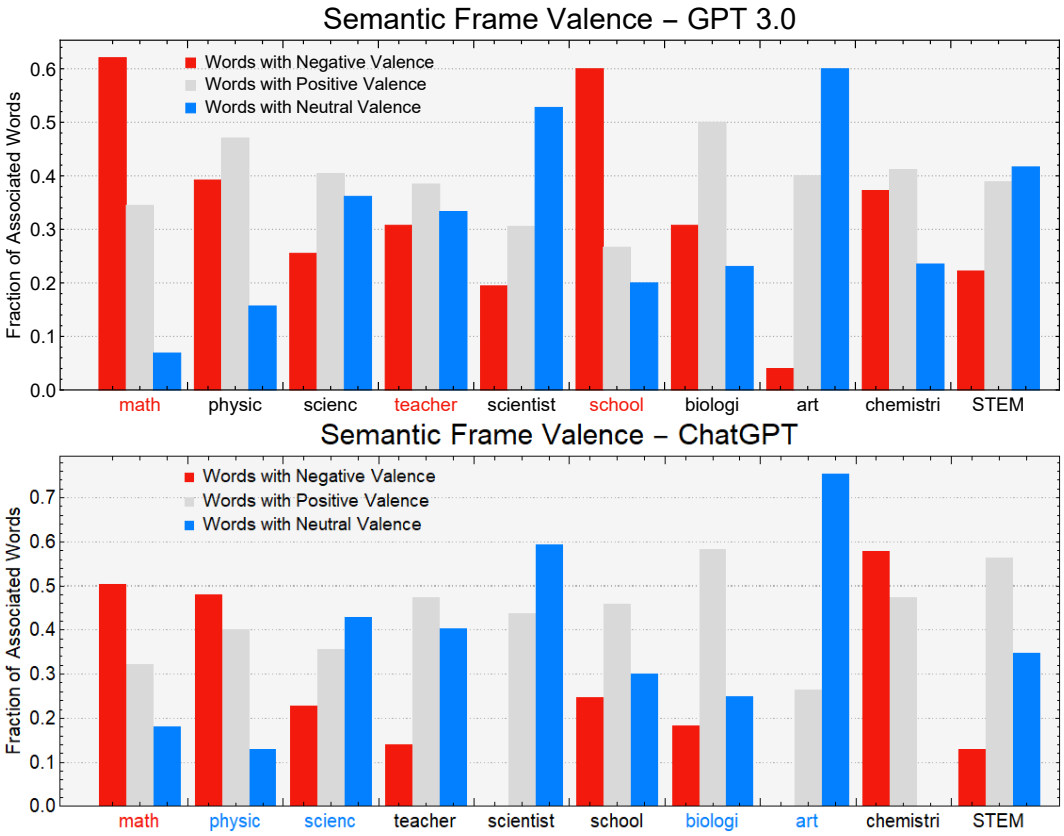


**Figure 1.** Fractions of positive, negative, and neutral associated words populating the semantic frames of each cue word in GPT-3 (**top**) and GPT-3.5 (**bottom**).

The above findings provide evidence that both GPT-3 and GPT-3.5 not only perceived *math* as a negative concept, but also framed it negatively. There is also evidence of polarized semantic frames surrounding several concepts. Our findings warranted further investigation, so we proceeded by investigating the semantic content of associations, aiming to understand how they would be emotionally interpreted by humans.

#### 4.1.1. LLMs Perceive Math Much More Negatively Than Science

As discussed earlier, the words in the semantic frame of a key concept provide important contextual information about how that key concept is perceived. Figure 2 visualizes the semantic frames for *science* (top) and *math* (bottom) as produced by GPT-3 (left) and GPT-3.5 (right).

GPT-3 framed *science* mostly in neutral (*data*, *hypothesis*, *method*) and positive (*curious*, *discover*, *knowledge*) terms, although there was a non-negligible fraction of negative associates. Some of these negative associates related to objects of investigation in science (e.g., *bacteria*, *chemicals*) while another cluster of negative associates described science as *hard*, *boring*, and *complicated*. Noticeably, *physics* also appeared in this cluster.

GPT-3.5 framed *science* in noticeably more positive terms. It is also easy to see that GPT-3.5 provided a richer, larger set of associates compared to GPT-3, which might be a consequence of the more advanced level of sophistication achieved by GPT-3.5 compared to its predecessor. While the semantic frame of *science* provided by GPT-3.5 was overwhelmingly positive, there were some negative associations that mostly related to theoretical and mathematical aspects of science. In contrast to GPT-3, the terms *complicated* and *boring* were not present in the semantic frame produced by GPT-3.5.

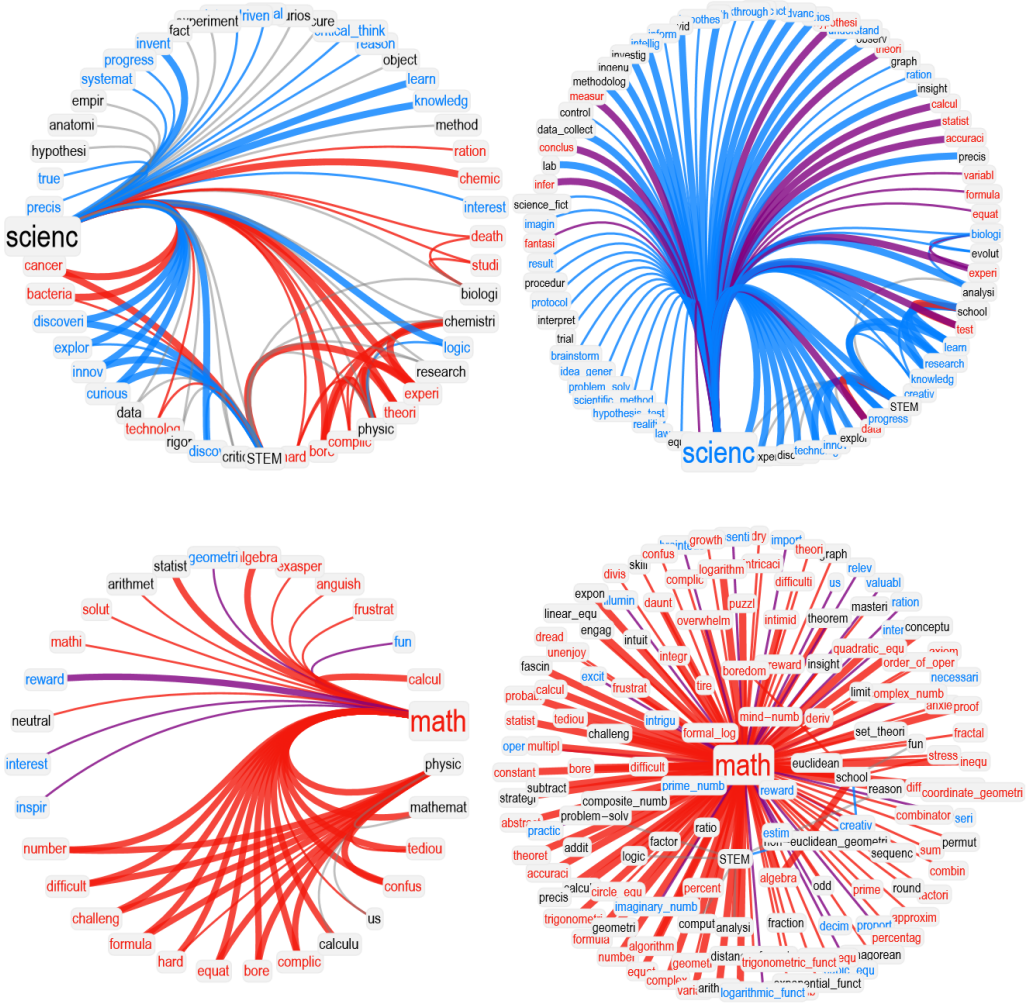


**Figure 2.** Semantic frames for *science* (**top**) and *math* (**bottom**) as produced by GPT-3 (**left**) and GPT-3.5 (**right**). Words in red (cyan) were rated as negative (positive). Words in black were rated as neutral. The cue word is displayed in a larger font size. Links between two negative terms are shown in red, while links between two positive terms are shown in cyan. Links between positive and negative words are shown in purple, indicating conflicting associations.

The concept *math* was perceived and framed overwhelmingly negatively by both LLMs. As with *science*, GPT-3.5 produced almost twice as many associates for *math* compared to GPT-3, another indication of the more advanced competencies exhibited by GPT-3.5.

GPT-3 framed *math* as a *boring*, *difficult*, *tedious*, *frustrating*, and *exasperating* concept. Theoretical tools used in math, such as *equation* and *formula*, were also perceived negatively. Such overwhelmingly negative perceptions were echoed by the associates provided by GPT-3.5, which identified similar negative aspects of math, describing it as *stressful*, *complicated*, *overwhelming*, and *dreadful*.

Compared to GPT-3, GPT-3.5 provided a considerably larger amount of associates related to domain knowledge for *math*, reflecting a more advanced knowledge of mathematical tools, including *exponentials*, *fractions*, *trigonometrics*, *percentages*, and *equations*, among others. Most associates from domain knowledge were perceived negatively, bolstering the overall negative connotation attributed by GPT-3.5 to *math* in its semantic frame.

The mixture of negative descriptive associations and negatively perceived domain knowledge terms provided by GPT-3.5 strongly echoes the negative semantic frame of *math* provided by high-school students identified in the previous work [27]. Our findings here provide strong evidence that GPT-3.5 provides a rather complex but strongly negative

perception of math, which is consistent with some negative perceptions possessed by some student populations.

The above results depend on the valence scores provided by LLMs. To further assess the presence and extent of negative emotions in semantic frames, we also used external scores, namely, arousal scores based on human judgment (from [52], see Methods) for assessing the emotional connotation of every associate. Figure 3 reports the 2D distributions of valence and arousal scores for all concepts in the semantic frames of *science* and *math* as produced by GPT-3 and GPT-3.5. A more intense yellow color indicates a concentration of associations within the same region of the circumplex model of affect [54], where the dimensions of valence ($x$-axis) and arousal ($y$-axis) map different emotional states. Notice that these models identify how concepts in semantic frames would be emotionally perceived by humans, thus providing a different perspective compared to the valence scores provided by language models discussed so far.

The associates in the semantic frame of *science* produced by both GPT-3.5 and GPT-3, Figure 3 (top right) and (top left), respectively, are concentrated mostly in the lower right quadrant, which corresponds to emotions of serenity and tranquility, i.e., positive valence and low arousal. This indicates that both the circumplex model and the forma mentis neighborhood portrayed *science* as a concept inspiring calm. More negative associations persisted in GPT-3 for *science*, as indicated by a cluster of concepts in the upper left quadrant, corresponding to anxiety and alertness, i.e., negative valence and higher arousal. This pattern, which is absent in GPT-3.5, corresponds with the negative associations outlined in the above semantic frame analysis.

The distribution of concepts in the circumplex of affect is considerably different for *math*. In Figure 3 (bottom left), GPT-3 features only a few concepts with positive valence scores and negligible arousal, falling outside the neutral range, a configuration considerably different from the one corresponding to *science*. Several concepts fall in the upper-left quadrant, confirming the anxious perception of *math* provided by GPT-3, which was discussed in the semantic frame analysis above.

GPT-3.5 framed *math* in a considerably more emotionally polarized way (see Figure 3, bottom right), compared to GPT-3. The associations provided by GPT-3.5 are distributed along a direction that spans the lower-right and the upper-left quadrants, combining emotions of serenity and tranquility with alertness, anxiety, and alarm. This emotional polarization further underlines the complexity of GPT-3.5, which is a language model capable of providing semantically richer and more emotionally polarized semantic frames for *math* compared to GPT-3. Furthermore, both GPT-3 and GPT-3.5 frame *math* in more negative terms compared to *science*.

#### 4.1.2. GPT-3 Perceives School and Teachers Much More Negatively Than GPT-3.5

Figure 4 reports the semantic frames for *school* for GPT-3 (top left) and GPT-3.5 (top right), together with their circumplex models (bottom).

GPT-3 perceived *school* as a negative concept and framed it with mostly negative and positive concepts. The language model associated school with positive jargon about learning (e.g., *learn*, *knowledge*, *education*) but also with negative jargon about tests, boredom, and dullness (*dull*, *unenjoyable*, *bored*). This dichotomy was confirmed also by the circumplex model, where most concepts fell in the lower-right quadrant (expressing serenity) and in the lower-left quadrant (expressing boredom). GPT-3 thus framed *school* as a partly boring, partly serene concept, crucial for learning.

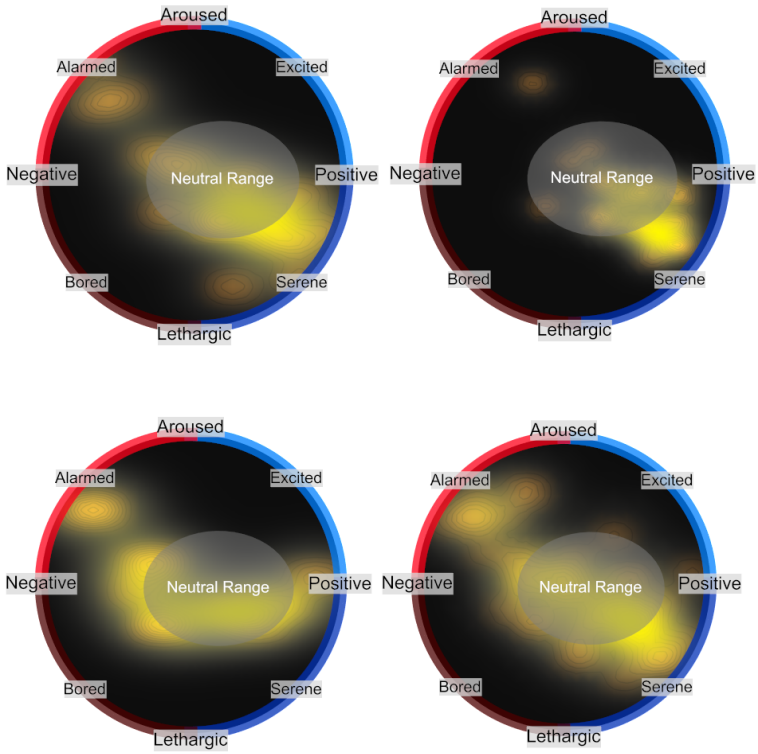


**Figure 3.** Circumplex model for the semantic frames of *science* (**top**) and *math* (**bottom**), as produced by GPT-3 (**left**) and GPT-3.5 (**right**).

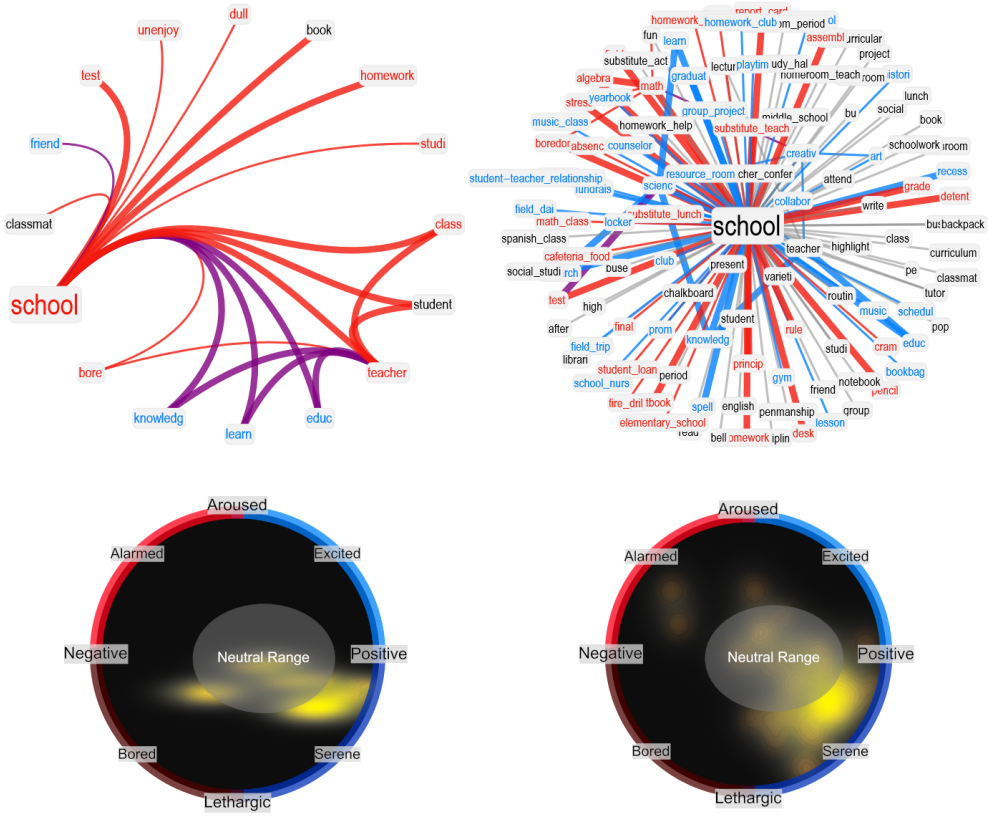


**Figure 4.** Semantic frames (**top**) and circumplex models (**bottom**) for *school*, as produced by GPT-3 (**left**) and GPT-3.5 (**right**); color patterns are the same as in Figure 2.

GPT-3.5 provided a much richer semantic frame for *school* compared to GPT-3, mixing both positive (*student–teacher relationship*, *knowledge*, *education*) and negative associates (*cafeteria food*, *detention*, *algebra*, *lunch*). Interestingly, the language model associated *school* with *algebra* and perceived the latter as a negative concept. Negative perceptions of *algebra* represent one of several indicators of math anxiety, as captured by the psychometric scale detecting math anxiety developed by Hunt and colleagues [55]. Interestingly, the circumplex model for the semantic frame of *school* does not reflect strong negative patterns; rather, most concepts in the model concentrate in the lower-right quadrant, corresponding to emotions of serenity and tranquility. This dichotomy indicates that most of the negative associations reported in the semantic frame are due to the specific perceptions produced by GPT-3.5 and cannot be reflected or reproduced by how a large population of humans would perceive those same concepts. For example, GPT-3.5 might perceive *algebra* as a negative concept while the NRC lexicon does not attribute negative valence scores to *algebra*. Observing this difference further underlines the power of behavioral forma mentis networks in terms of adapting to the specific perceptions portrayed by specific individuals or groups.

Figure 5 reports the semantic frames for *teacher* for GPT-3 (top left) and GPT-3.5 (top right), together with their circumplex models (bottom). The semantic frames for *teacher* differ significantly between GPT-3 and GPT-3.5. GPT-3 identified "teacher" with a negative valence and surrounded it mostly with other negative concepts, e.g., *authoritarian*, *demanding*, *know-it-all*, *boredom*. These negative associations co-existed with positive ones, mentioning aspects related to knowledge transmission (e.g., *education*, *wisdom*, *wise*) and mentoring (e.g., *dedicate*, *caring*, *mentor*). This emotional polarity is also confirmed by the circumplex model.

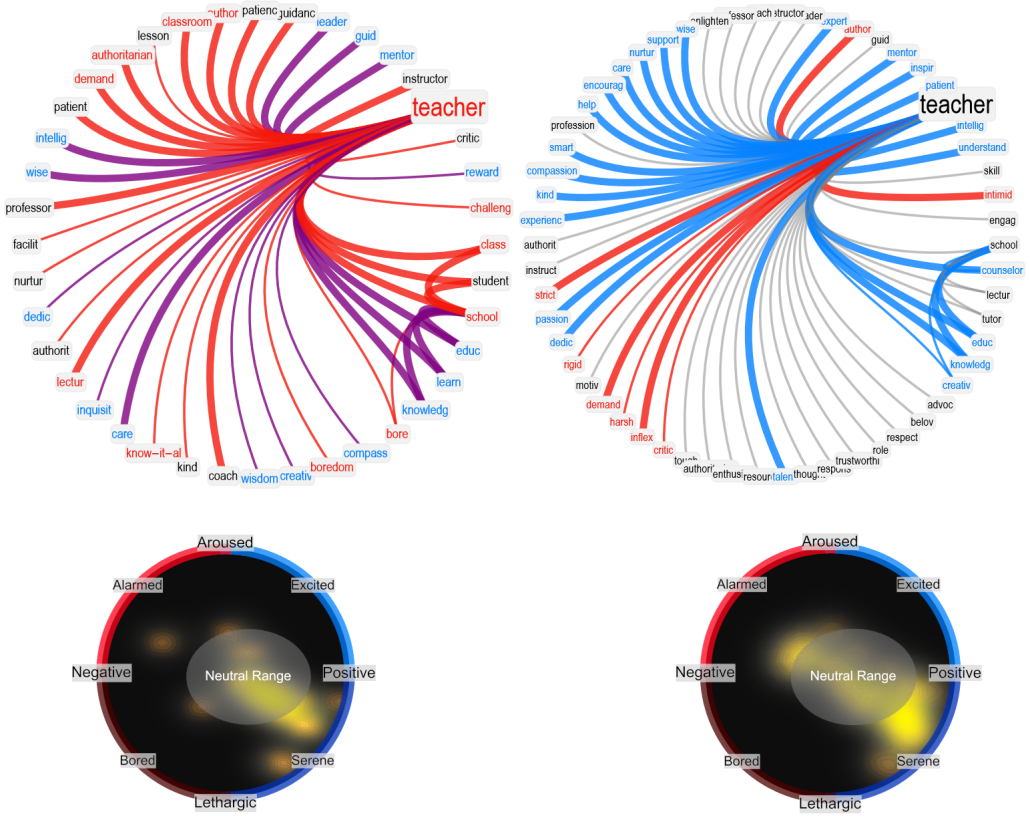


**Figure 5.** Semantic frames (**top**) and circumplex models (**bottom**) for *teacher* as produced by GPT-3 (**left**) and GPT-3.5 (**right**). color patterns are the same as in Figure 2.

Compared to GPT-3, GPT-3.5 put a stronger focus on the positive aspects of *teacher*, underlining their ability to *nurture*, *encourage*, *care*, and *support* in their roles. These aspects provide a perception of teachers as leaders and mentors, such that positive aspects dominate the entire semantic frame. However, negative perceptions are still present, i.e., associations between *teacher* and *strict*, *demand*, *harsh*, and *criticize*. This dualistic positive/negative perception of teachers is also confirmed by the circumplex model, where concepts concentrate in the lower-right and upper-left quadrants, corresponding to emotions of calmness and alertness, respectively.

#### 4.1.3. Other Perceptions: Physics, Chemistry, STEM, Art, and Scientist

Figure 6 portrays the semantic frames for *chemistry* (top), *STEM* (middle), and *physics* (bottom).

GPT-3 identified all three concepts as neutral but surrounded *physics* with a mostly negative frame, whereas semantic frames for both *chemistry* and *STEM* were polarized. Semantic network analysis revealed that the negative associates of *chemistry* mostly relate to aspects of reagents and acids (*poison*, *pungent*, *acrid*). Similarly, the negative associates of *STEM* were mostly related to aspects in the health sciences where STEM can bring substantial improvements to well-being. In this way, the negativity found in the semantic frames for *chemistry* and *STEM* can be explained in terms of negative elements or challenges studied in these disciplines. This pattern is strikingly different from the negative associates found in the semantic frame of *physics*, which mentions the concepts *hard*, *difficult*, and *complicated*. These associates are not elements studied in physics, but rather negative perceptions, which were found also in the semantic frames of *math*.

GPT-3.5 associated *chemistry* with general-level experimental and theoretical elements, mostly perceived as neutral or negative. This indicates another difference in how the two language models portray the same concept. Interestingly, GPT-3.5 produced a positive semantic frame for *STEM*, establishing links with concepts such as *technology*, *innovation*, and *progress*. This finding is analogous to high-school students holding negative perceptions of math and physics while holding positive attitudes towards science [27,41,42]. GPT-3.5 framed *STEM* in positive terms but also linked it with negatively perceived terms such as *math* and *mathematical*. This dichotomy suggests that GPT-3.5 reflects distorted perceptions in which math is perceived negatively but still associated with positive perceptions of STEM and research.

Notably, GPT-3.5 perceived *physics* as a positive entity, framed within a highly polarized semantic frame, rich with positive and negative concepts. This is in contrast with GPT-3.5's negative perception of *math*. This pattern differs from what was found in previous studies [27,41,42], where high-school students framed both *math* and *physics* in negative terms.

The overwhelmingly positive semantic frames of *art* and *scientist* produced by both GPT-3 and GPT-3.5, as shown in Figure 7, are in stark contrast to the frame of *math*, demonstrating the vast range of perceptions about academic disciplines exhibited by these LLMs. Interestingly, GPT-3 provided associations related to the *mad scientist* stereotype [56], which were found also in the perception that high-school students had of *scientist* in [42].

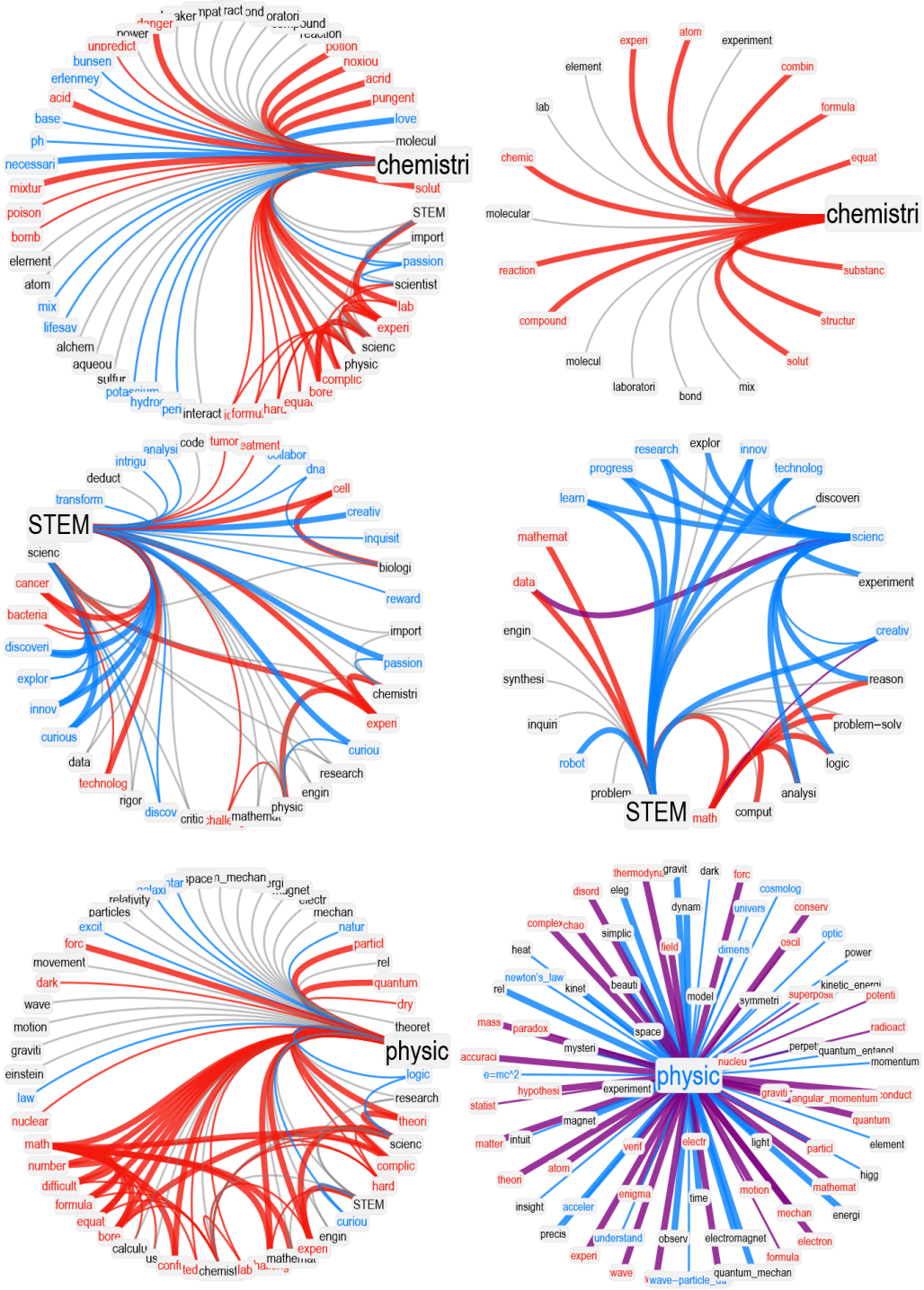


**Figure 6.** Semantic frames for *chemistry* (**top**), *STEM* (**middle**) and *physics* (**bottom**), as produced by GPT-3 (**left**) and GPT-3.5 (**right**); color patterns are the same as in Figure 2.

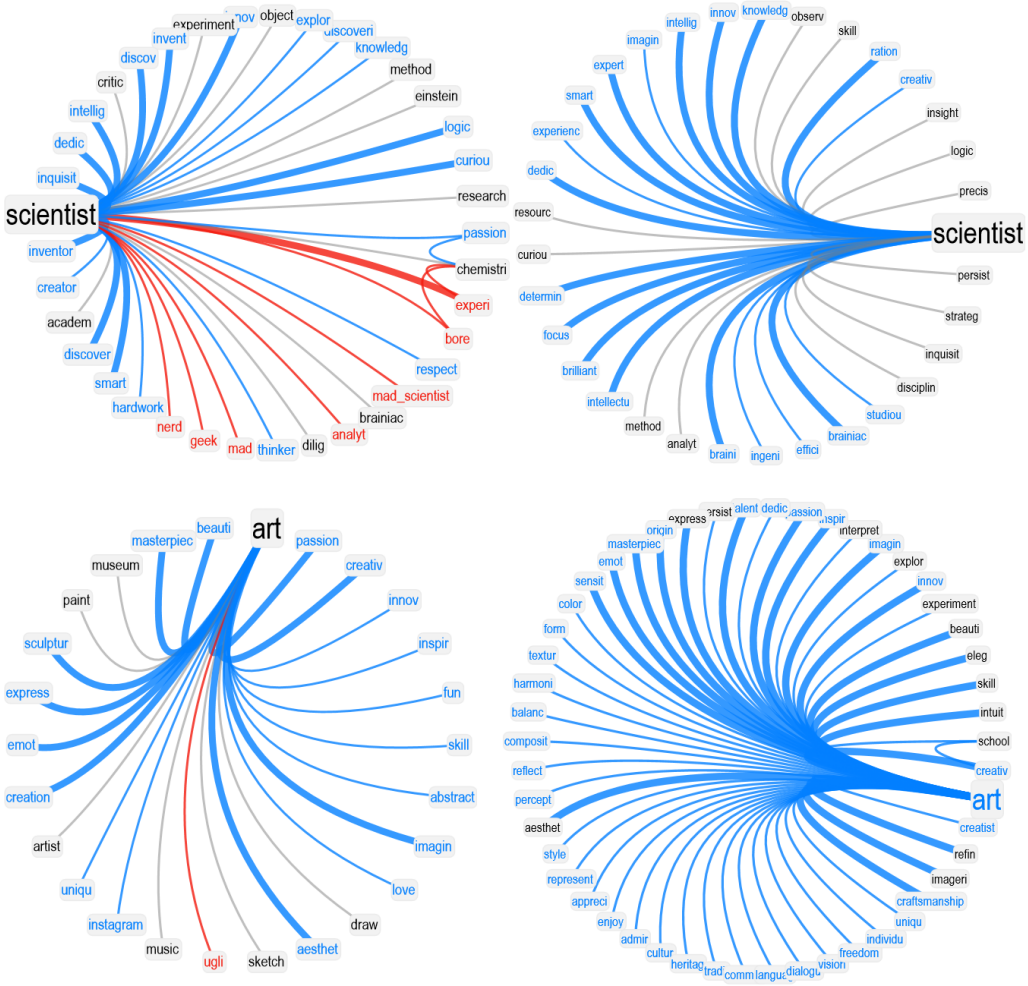


**Figure 7.** Semantic frames for *scientist* (**top**) and *art* (**bottom**), as produced by GPT-3 (**left**) and GPT-3.5 (**right**); color patterns are the same as in Figure 2.

### *4.2. Comparison with GPT-4*

At the time of writing this manuscript, GPT-4 was an unreleased product made available to paying customers by OpenAI. We sampled GPT-4 a few days before it became available in Italy. Here, we present the results of our experiments as they relate to the in-depth results for GPT-3 and GPT-3.5 discussed in the above sections.

Figure 8 reports the semantic frames of *math*, *physics*, and *school* produced by GPT-4. Significantly notable is that these semantic frames are far more positive compared to those produced by GPT-3 and GPT-3.5, especially for *math* and *physics*. Both *math* and *physics* were perceived neutrally by GPT-4 but associated mostly with positive concepts. Although negative associations constituted less than 15% of the semantic frames, stereotypical associations related to math anxiety persisted. Even in the positive associations provided by GPT-4, *math* was associated with *frustrating*, *anxiety*, *fearful*, *intimidating*, *confusing*, and *struggle*. These negative associations were not found in the semantic frame of *physics*, whose negative associates were related to domain knowledge (e.g., *chaos*, *nuclear*). This semantic frame analysis thus implies that negative perceptions towards *math* by GPT-4 have been reduced compared to its earlier versions, but negative stereotypical perceptions persist, reflecting the psychological phenomena of math anxiety that is pervasive in society.

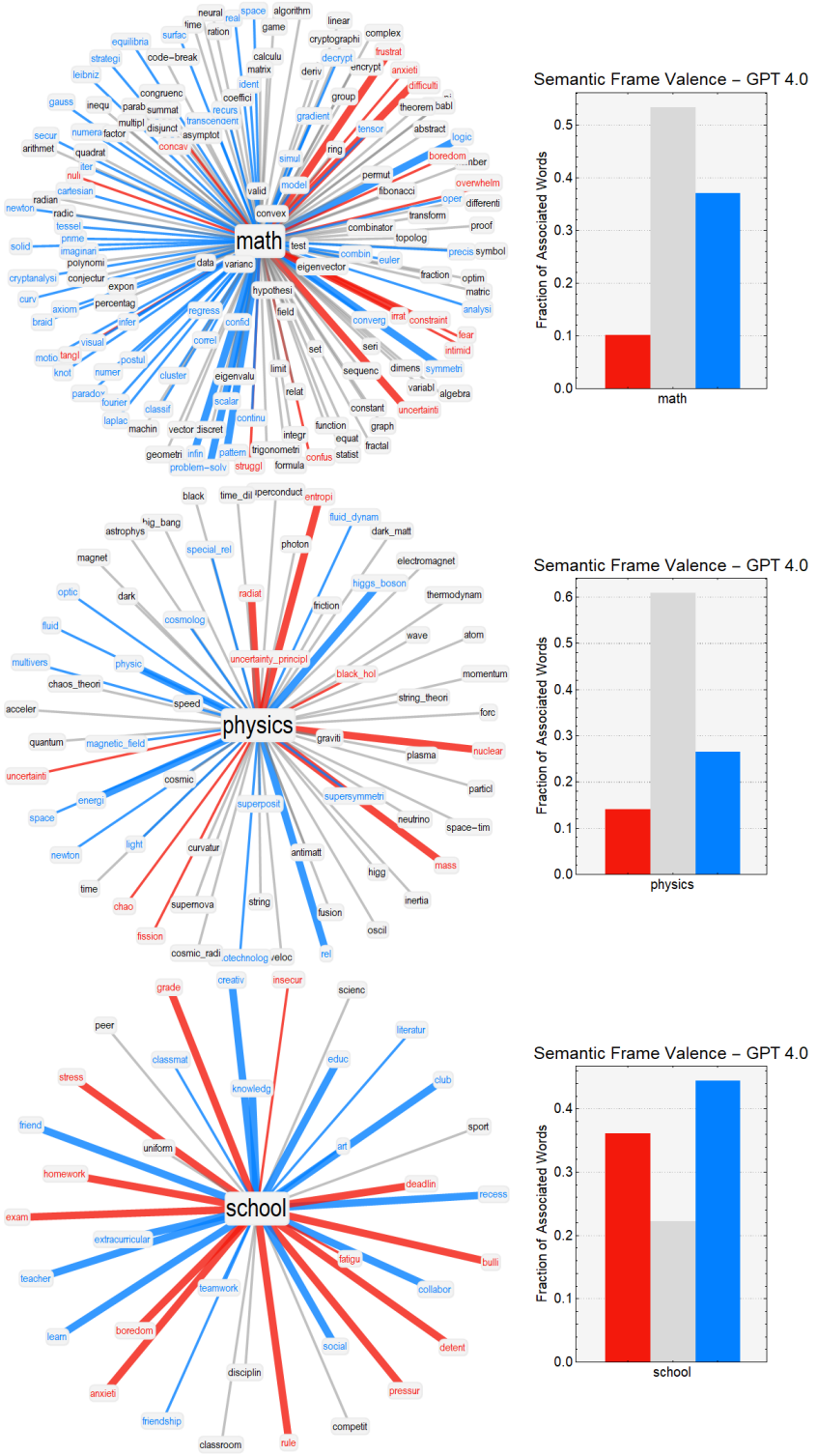


**Figure 8.** Semantic frames for *math* (**top**), *physics* (**middle**), and *school* (**bottom**), as produced by GPT-4. The frequency of negative, positive, and neutral words in each frame are coded in frequency histograms next to each semantic frame. Color patterns are the same as in Figure 2.

The semantic frame of *school* produced by GPT-4 was less dominated by negative valence than the one produced by GPT-3, but was similarly polarized as the one produced by GPT-3.5. Negative associations related *school* with *frustrating*, *exam*, *anxiety*, and *boredom*, indicating a persistent negative perception of school settings with negative emotions and test anxiety. Interestingly, GPT-4 associated *school* with *bullying*, an association that was

absent in results from previous language models. With GPT-4 being trained on a larger amount of web data, this association might reflect the growing sensitivity to bullying in school, as discussed in online forums.

Overall, our semantic frame analysis shows that negative perceptions of math and physics exhibited by GPT-3 and GPT-3.5 have been reduced in GPT-4. Nonetheless, harmful stereotypical perceptions about math related to frustration and anxiety persist. This is of great concern, considering the increasing use of LLMs by students. Exposure to negative implicit biases towards math in LLMs poses the risk of exacerbating the harmful math anxiety that has negative consequences on both individuals and our society.

*A Focus on the Evolution of Math Perceptions in LLMs*

Table 1 reports the different measurements outlined in Section 3 relative to the semantic frames of *math* as obtained from GPT-3, GPT-3.5, GPT-4, and high-school students (data obtained from [27]). Interestingly, GPT models produced larger math-focused semantic frames of increasing semantic richness (i.e., number of unique associates provided for *math*) in subsequent generations. This indicates a progressively richer framing/connotation of math corresponding to an increase in complexity and parameters in an LLM. Even though the sample size of responses was the same for LLMs and high-school students, GPT-3.5 and GPT-4 produced larger semantic frames compared to high-school students, i.e., more variety of responses (first row).

**Table 1.** Reference values for the semantic frame of *math* as reported by GPT 3, GPT-3.5, GPT 4, and high-school students.

| Measure/Model | GPT-3 | GPT-3.5 | GPT-4 | High School Students |
|---|---|---|---|---|
| Semantic Frame Size | 30 | 134 | 149 | 48 |
| Estimated Valence | $1.8 \pm 0.1$ (Negative) | $2.0 \pm 0.1$ (Negative) | $3.3 \pm 0.2$ (Neutral) | $1.8 \pm 0.3$ (Negative) |
| Estimated Frame Valence | Negative | Negative | Positive | Negative |
| Positive/Neutral/Negative % in Frame | 0.06/0.33/0.61 | 0.18/0.32/0.50 | 0.37/0.53/0.10 | 0.10/0.44/0.46 |
| Non-Emotional Words in Frame | 0.37 | 0.56 | 0.74 | 0.84 |
| Non-Emotional W. Positive/Neutral/Negative % in Fr. | 0.18/0.37/0.45 | 0.12/0.33/0.54 | 0.39/0.51/0.10 | 0.07/0.40/0.43 |

Interestingly, as summarized in Table 1 (second row), GPT-3, GPT-3.5, and high-school students all assigned a negative valence label to *math*, while GPT-4 assigned a neutral valence label. These negative and neutral perceptions are consistent with the respective negative and positive semantic frames (third/fourth row). Hence, the negative perceptions of *math* by older LLMs are in line with those of high-school students who are influenced by math anxiety. On the other hand, the newer language model (GPT-4) shows some improvement in this area, identifying *math* as a neutral concept linked with several positively perceived concepts. This pattern indicates an intriguing evolution of GPT-4 compared to its predecessors in terms of overcoming negative attitudes toward math.

Table 1 also considers non-emotional words (fifth row), i.e., words that were featured in the semantic frame of *math* but were not featured in the National Research Canada Emotion Lexicon [53]. These non-emotional words are interesting because they provide a way to gauge the number of domain knowledge associates, i.e., associations of *math* related to its foundational elements, instruments, and tools. Interestingly, high-school students provided the highest percentage of non-emotional words (84%). In LLMs, the percentage of non-emotional words increased with newer versions, demonstrated by a growing tendency for GPT-4 (74%) to produce domain–knowledge associations compared to GPT-3 (37%) and GPT-3.5 (56%), thus approaching the amount of domain–knowledge produced by high-school students (84%).

Regarding the perceptions of these non-emotional concepts (sixth row), high-school students (43%), GPT-3 (45%), and GPT-3.5 (54%) all tended to perceive them negatively, indicating an unpleasant feeling about mathematical methods and instruments. In contrast, GPT-4 identified these non-emotional concepts as mostly neutral (39%) and positive (51%).

This finding highlights GPT-4's more positive attitude towards math compared to older and simpler language models.

## 5. Discussion

Our findings provide compelling evidence that large language models, including GPT 3, GPT-3.5 and even GPT-4, frame academic concepts such as math, school, and teachers with strongly negative associations. These deviations from neutrality were quantified within the quantitative framework of behavioral forma mentis networks [27,41], i.e., cognitive networks representing continued free association data enriched with valence scores. In the absence of impersonation, GPT-3 and GPT-3.5 in particular provided negative connotations for *math*, perceiving it as a boring and frustrating discipline, and providing no positive associations with complex real-world applications. Unlike STEM experts, who linked creativity and real-world applications to *math* (as found in previous work [27]), LLMs framed *math* as detached from scientific advancements and real-world understanding. This pattern was identified in two different approaches, one leveraging semantic frame analysis [26] and another using the circumplex model of affect [48], powered through psychological data. Our analyses identified concerning deviations from neutrality in how GPT-3.5 and GPT-3 framed *math*, highlighting negative stereotypical associations as expressed through negative emotional jargon, even in the latest GPT-4 model.

Exposure to these stereotypical associations and negative attitudes/framings could have serious repercussions. As discussed in Section 1, LLMs act as psycho-social mirrors, reflecting the biases and attitudes embedded in the language used for training LLMs [3,8,14]. These models are complex enough to capture and mirror such human biases and negative attitudes in ways we do not yet fully understand [15]. This lack of transparency translates into a relative difficulty in tracking the outcome of inquiries to LLMs: Are the framings provided by these artificial agents prevalent in the text produced by them? More importantly, could subtle and consistent exposure to such negative associations have a negative impact on some users? This represents an important research direction for future investigations of LLMs, particularly regarding the worsening of math anxiety. Social interactions with LLMs may, thus, exacerbate already existing stereotypes or insecurities about mathematical topics among students and even parents, analogous to the unconscious diffusion of math anxiety through parent–child interactions, as identified by recent psychological investigations [57]. Negative associations of math and other concepts may be very subtle, e.g., LLMs might produce text framing math in ways that confirm students' pre-existing negative attitudes [21,22]. They may also bolster subliminal messages that math is hard for some specific groups, influencing their academic performance through a phenomenon known in social psychology as stereotype threat (cf. [25]). Such negative attitudes can have harmful effects on learning technical skills in mathematics and statistics, as evidenced by previous studies [17,23] that found a negative association between math anxiety levels and learning performance in math and related courses.

Notably, compared to GPT-3.5, GPT-3 provided more negative associations and fewer positive associations for STEM disciplines such as *math* and *physics*, but also for *school* and *teacher*. In all these cases, the semantic frames produced by GPT-3.5 featured more unique associations compared to GPT-3, leading to semantically richer neighborhoods (e.g., the semantic frame of *math* featured associations with several aspects of domain knowledge in GPT-3.5 but not in GPT-3). Hence, richer and more complex semantic representations for GPT-3.5 might depend on the more advanced level of sophistication achieved by its architecture, at least when compared to its predecessor GPT-3. This observation is further supported if we consider the performance provided by GPT-4, which was associated with more domain–knowledge concepts compared to previous LLMs. Noticeably, not only was the semantic frame for *math* richer in GPT-4 compared to semantic frames from other LLMs, but GPT-4 also overcame negative math attitudes by displaying more neutral and positive associations for that category. This makes the overall valence connotation for *math* in GPT-4 much closer to the positive levels observed among STEM experts and very

different from the overwhelmingly negative, displeasing attitudes observed in high-school students [27]. In general, in GPT-4, the negative connotations for *math*, *physics*, and *school* that were present in GPT-3.5 and GPT-3 seemed to be drastically diminished, probably due to a combination of effects, e.g., a set of richer and more complex training resources selected by human intervention during the training phase to minimize bias, or a more sophisticated model parameterization, in which human intervention might filter our biases [1]. Either phenomenon would consequently cause GPT-4 to have weaker manifestations of the biases encoded in previous instances of the model, i.e., GPT-4 might be mirroring different bias levels when compared to GPT-3 and GPT-3.5. This reduction in bias could also be related to the use of reinforcement learning with human feedback (RLHF) fine-tuning that GPT-4 authors claim could reduce undesirable/overly cautious responses when unsafe/safe inputs are given by users [1], thus leading to improved neutrality in GPT-4 responses even when prompts are not neutral. This lets us know more directly that the need for appropriate behavior in LLM outputs is central to the interests of the authors of GPT-4 regarding expressions of neutrality and objectivity. Intriguingly, there might also be a third phenomenon at play: the increased model complexity of GPT-4 might either make the model more "aware" of negative biases, or change the way it "relates" to math itself, leading to bias reduction in either case. Spreading awareness about math anxiety is a key first step to reducing it, mainly because acknowledging its potential psycho-social impacts could reduce the spread of negative attitudes towards math among peers, teachers, and family members [25]. Recent psychological investigations of math anxiety among humans found reduced levels of math anxiety in students with stronger self–math overlap [58], i.e., a psychological construct expressing the extent to which an individual integrates math into their sense of self. Analogously to humans, GPT-4 might thus have an increased awareness of the biases related to math anxiety or a stronger self–math overlap, which would both explain the reduced levels of math-related biases observed in its semantic frames. Alas, in absence of more detailed information about the training material, filtering process, and architecture, we cannot narrow down the specific mechanisms for explaining the patterns observed here, but rather call for future research investigating these aspects in more detail.

In summary, the application of behavioral forma mentis networks to LLMs confirms the benefits of adopting a cognitive psychology approach for evaluating how large language models perceive and frame math and STEM-related concepts. In this respect, our contribution aligns with the goals of machine psychology [7], which aim to discover emergent capabilities of LLMs that cannot be detected by most traditional natural language processing benchmarks. In particular, because of the sophisticated ability of LLMs to elaborate and engage in open-domain conversations [1], a structured cognitive investigation of behavioral patterns shown by LLMs appears to be natural and necessary. However, some caution should be taken when analogizing LLMs to participants in psychology experiments and then using the corresponding experimental paradigms to measure relevant emerging properties of LLMs.

Firstly, in cognitive psychology, there must be an adequate match between a given implemented task measuring a target process and the cognitive theory or model used to explain that process [59,60]. For instance, past works have established a quantitative and theory-driven link between continued free association tasks—deriving free associations between concepts—and models based on such data whose network structure could explain aspects of conceptual recall from semantic memory [40,44] or even higher-level phenomena such as problem solving [43]. For instance, according to the spreading activation model established by pioneering work of psychologists Collins and Loftus [61], providing an individual with a cue word activates a cognitive process acting on a network representation, such that concepts are nodes linked together via conceptual associations. The activation of the node representing that given cue word facilitates a process such that activation signals start spreading iteratively through the network, diffusing or concentrating over other related nodes/concepts. Retrieval is then guided by stronger levels of activation which accumulate over other nodes (e.g., the cue *book* leading to the retrieval of *letter*). This

spreading activation model has been extensively tested in cognitive psychology and it represents one among many potentially suitable models for interpreting free association data and their psychological nature within human beings [62,63]. However, in LLMs, this link between cognitive theory and experimental paradigms is mostly absent. Researchers do not yet know whether LLMs are able to approximate human semantic memory or any of its mechanisms [8], mainly because LLMs are trained on massive amounts of textual data [1] in ways that differ greatly from the usual ways in which humans acquire language [64] and its emotional/cognitive components [60]. Furthermore, another difference is that LLMs usually combine text sources from multiple authors and can thus end up reflecting multimodal-type populations [12], making it extremely difficult to compare LLMs against the workings of a prototypical cognitive model at the level of an individual. In other words, there is a problematic connotation for LLMs as "artificial persona": these models can produce language in ways that appear similar to those of humans but "learn" language in a way that is much different from humans [60].

Consequently, forma mentis networks in LLMs might not represent semantic frames [41,42] in ways that are analogous to how humans organize their semantic memory. This limitation strongly hampers the cognitive interpretation of semantic frames between human-generated and LLM-generated data. In fact, the main focus of this study is not to compare LLM-generated data with human-generated data, rather, the focus is on quantifying the attitudes expressed across several LLMs, and comparing how different implementations of the same overall cognitive architecture, i.e., transformer networks, represent and associate the same sets of stimuli according to the same initial prompt.

A consequence of the limited cognitive interpretation of LLM-generated data lies in the presence of an interplay between semantic and emotional aspects of memory. In humans, recent psychological studies have highlighted an interplay between retrieval processes in the categorical organization of episodic memory and the activation of related concepts in semantic memory [16,65,66]. This translates into an interplay that emotions—potentially coming from past positive, neutral, or negative episodic memories [16]—might have in guiding or influencing retrieval (rather than encoding) of semantic knowledge [65,66]. Past works using behavioral forma mentis networks have shown that students and STEM experts attribute rather different affective connotations to the same concepts, particularly physics and mathematics [27,42]. Such differences could be interpreted in terms of episodic memories attributing different emotional connotations to the outputs of the recall processes activated by the continued free association task in BFMNs (see also [25]). However, such an interpretation would not hold for LLMs, given their opaque structure and the uncertainty in the "cognitive" phenomena which regulate their concept retrieval [8]. To the best of our knowledge, no explanation of how LLMs work has yet to leverage cognitive models of human memory, mainly because of the intrinsically different ways in which humans and LLMs function. We raise this cautionary point as an encouragement for the psychology and cognitive science communities to provide novel theoretical models that surpass the mere description of optimization processes and search in training data [1], to develop frameworks that take into account the cognitive aspects of language for training data. Given that GPT-4 and its predecessors use vast amounts of human data, interpreting the cognitive structure of LLMs might lead to substantial advancements in understanding how human social cognitions are structured [31].

Can we ever expect future LLMs to be completely free from biases, stereotypical perceptions, and negative attitudes? Probably not. We found that GPT-4 produced fewer negative associations for *math* compared to previous LLMs, so there is evidence of reduced biases. However, it is unlikely, and perhaps even undesirable, that future LLMs will be completely free from biases, at least when considering their training. According to [12**?** ], biases in LLMs can foster efficient algorithmic decision-making, especially when dealing with complex, unstable, and uncertain real-world environments. Furthermore, biases in the training data of LLMs can greatly boost the efficiency of learning algorithms [12]. Unlike artificial systems, however, real people may produce biases because of three fundamental

limitations of human cognition [68]: limited time, limited computation power, and limited communication. Limited time magnifies the effect of limited computation power, and limited communication makes it harder to draw upon more computational resources, which may ultimately lead to biased behaviors. Cognitive science thus entails a kind of *bias paradox*, where the two systems (artificial LLMs and human cognitive systems) apparently manifest a similar behavior (including eventual observable biases) as a result of structurally and functionally different architectures. In this way, the negative attitudes found here within LLMs should be taken with a grain of salt when compared to the negative perceptions mapped in humans in previous works [27,42,51]. Despite different psychological roots [64], the biases found here have much in common, considering the negative perceptions currently flowing online that depict math and other STEM concepts as boring, dry, and frustrating [22,23]. Overcoming these stereotypical perceptions will require large-scale policy decisions. Focused efforts should concentrate on reducing negative biases within LLMs, whose sphere of influence reaches an ever-increasing audience. Whenever possible, explainable AI methods can provide methods to reduce the bias in LLMs. For instance, they have also been used to explain a model trained to differentiate between texts generated by humans and ChatGPT, demonstrating that ChatGPT generates texts that are more polite and generic, impersonal, and without expressing feelings [69]. Together with forma mentis networks, or with a suitable combination, such methods could be useful towards the construction of frameworks able to discover and reduce bias in LLMs. Reducing the amount of bias present in LLMs after training is a feasible way to promote ethical interactions between humans and LLMs without perpetuating subtle negative perceptions of math and other neutral concepts.

Lastly, regarding limitations of our work, we would like to point out that, because of its structure, GPT systems are commercial products whose validity can be investigated by researchers but cannot be fully reproduced by everyone. For instance, the GPT-3 system is not available to the public via the old interface or API system, and there is no guarantee that the mini-versions released to the public correspond to the model made available by OpenAI almost one year ago. The same is true also for GPT-4, which is being continuously updated even while being available for Pro users. These remain limitations of ours and all other studies using OpenAI systems.

*Conclusions*

In this work, we showed how the cognitive framework of behavioral forma mentis networks (BFMNs) can produce quantitative insights about the ways in which large language models portray specific concepts. Despite several limits to the cognitive interpretation of this approach, which is rooted in psychological theories about the nature of semantic and lexical retrieval processes in humans, BFMNs represent a powerful framework for highlighting key associations that are likely promoted by many LLMs. Here, we found that different LLMs can greatly vary in the amount and type of negative, stereotypical, and biased associations they produce, indicating that machine psychology approaches such as BFMNs can contribute to understanding differences in the structure of knowledge promoted across various large language models.

**Author Contributions:** Conceptualization, M.S.; Methodology, K.A., S.C. and M.S.; Validation, L.L., G.R. and M.S.; Formal analysis, M.S.; Investigation, K.A., S.C., L.L., G.R. and M.S.; Resources, M.S.; Data curation, M.S.; Writing—original draft, K.A., S.C., L.L. and M.S.; Visualization, M.S.; Supervision, G.R. and M.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Acknowledgments:** We acknowledge Clara Rastelli and Nicola De Pisapia for insightful feedback on very early stages of this project.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
2. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
3. e Souza, B.C.; Silva, F.N.; de Arruda, H.F.; da Silva, G.D.; Costa, L.D.F.; Amancio, D.R. Text characterization based on recurrence networks. *Inf. Sci.* **2023**, *641*, 119124. [CrossRef]
4. Binz, M.; Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2218523120. [CrossRef]
5. Shiffrin, R.; Mitchell, M. Probing the psychology of AI models. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2300963120. [CrossRef]
6. Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A.A.M.; Abid, A.; Fisch, A.; Brown, A.R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv* **2022**, arXiv:2206.04615.
7. Hagendorff, T. Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv* **2023**, arXiv:2303.13988.
8. Mitchell, M.; Krakauer, D.C. The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2215907120. [CrossRef]
9. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* **2023**, arXiv:2303.12712.
10. Bender, E.M.; Koller, A. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5185–5198.
11. Niven, T.; Kao, H.Y. Probing neural network comprehension of natural language arguments. *arXiv* **2019**, arXiv:1907.07355.
12. Ferrara, E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *arXiv* **2023**, arXiv:2304.03738.
13. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [CrossRef] [PubMed]
14. Sasson, G.; Kenett, Y.N. A Mirror to Human Question Asking: Analyzing the Akinator Online Question Game. *Big Data Cogn. Comput.* **2023**, *7*, 26. [CrossRef]
15. Anoop, K.; Gangan, M.P.; Deepak, P.; Lajish, V. Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias. In *Responsible Data Science: Select Proceedings of ICDSE 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 13–45.
16. Pulcu, E.; Browning, M. Affective bias as a rational response to the statistics of rewards and punishments. *eLife* **2017**, *6*, e27879. [CrossRef] [PubMed]
17. Foley, A.E.; Herts, J.B.; Borgonovi, F.; Guerriero, S.; Levine, S.C.; Beilock, S.L. The math anxiety-performance link: A global phenomenon. *Curr. Dir. Psychol. Sci.* **2017**, *26*, 52–58. [CrossRef]
18. Luttenberger, S.; Wimmer, S.; Paechter, M. Spotlight on math anxiety. *Psychol. Res. Behav. Manag.* **2018**, *11*, 311–322. [CrossRef]
19. Maloney, E.A.; Beilock, S.L. Math anxiety: Who has it, why it develops, and how to guard against it. *Trends Cogn. Sci.* **2012**, *16*, 404–406. [CrossRef]
20. Ramirez, G.; Hooper, S.Y.; Kersting, N.B.; Ferguson, R.; Yeager, D. Teacher math anxiety relates to adolescent students' math achievement. *AERA Open* **2018**, *4*, 2332858418756052. [CrossRef]
21. Ashcraft, M.H. Math anxiety: Personal, educational, and cognitive consequences. *Curr. Dir. Psychol. Sci.* **2002**, *11*, 181–185. [CrossRef]
22. Ashcraft, M.H.; Ridley, K.S. Math anxiety and its cognitive consequences. In *Handbook of Mathematical Cognition*; Taylor & Francis Group: Abingdon, UK, 2005; pp. 315–327.
23. Daker, R.J.; Gattas, S.U.; Sokolowski, H.M.; Green, A.E.; Lyons, I.M. First-year students' math anxiety predicts STEM avoidance and underperformance throughout university, independently of math ability. *NPJ Sci. Learn.* **2021**, *6*, 17. [CrossRef]
24. Hembree, R. The nature, effects, and relief of mathematics anxiety. *J. Res. Math. Educ.* **1990**, *21*, 33–46. [CrossRef]
25. Stella, M. Network psychometrics and cognitive network science open new ways for understanding math anxiety as a complex system. *J. Complex Netw.* **2022**, *10*, cnac012. [CrossRef]
26. Stella, M.; Kapuza, A.; Cramer, C.; Uzzo, S. Mapping computational thinking mindsets between educational levels with cognitive network science. *J. Complex Netw.* **2021**, *9*, cnab020. [CrossRef]
27. Stella, M.; De Nigris, S.; Aloric, A.; Siew, C.S. Forma mentis networks quantify crucial differences in STEM perception between students and experts. *PLoS ONE* **2019**, *14*, e0222870. [CrossRef] [PubMed]
28. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*.[CrossRef]
29. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. [CrossRef]
30. Manzini, T.; Lim, Y.C.; Tsvetkov, Y.; Black, A.W. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv* **2019**, arXiv:1904.04047.

31. Prates, M.O.; Avelar, P.H.; Lamb, L.C. Assessing gender bias in machine translation: A case study with google translate. *Neural Comput. Appl.* **2020**, *32*, 6363–6381. [CrossRef]
32. Nadeem, M.; Bethke, A.; Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv* **2020**, arXiv:2004.09456.
33. Greenwald, A.G.; McGhee, D.E.; Schwartz, J.L. Measuring individual differences in implicit cognition: The implicit association test. *J. Personal. Soc. Psychol.* **1998**, *74*, 1464. [CrossRef]
34. Kurita, K.; Vyas, N.; Pareek, A.; Black, A.W.; Tsvetkov, Y. Measuring bias in contextualized word representations. *arXiv* **2019**, arXiv:1906.07337.
35. Abid, A.; Farooqi, M.; Zou, J. Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; pp. 298–306.
36. Lucy, L.; Bamman, D. Gender and representation bias in GPT-3 generated stories. In *Third Workshop on Narrative Understanding*; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 48–55.
37. Sheng, E.; Chang, K.W.; Natarajan, P.; Peng, N. The woman worked as a babysitter: On biases in language generation. *arXiv* **2019**, arXiv:1909.01326.
38. Magee, L.; Ghahremanlou, L.; Soldatic, K.; Robertson, S. Intersectional bias in causal language models. *arXiv* **2021**, arXiv:2107.07691.
39. Li, X.; Li, Y.; Liu, L.; Bing, L.; Joty, S. Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective. *arXiv* **2022**, arXiv:2212.10529.
40. De Deyne, S.; Navarro, D.J.; Storms, G. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav. Res. Methods* **2013**, *45*, 480–498. [CrossRef]
41. Stella, M.; Zaytseva, A. Forma mentis networks map how nursing and engineering students enhance their mindsets about innovation and health during professional growth. *PeerJ Comput. Sci.* **2020**, *6*, e255. [CrossRef]
42. Stella, M. Forma mentis networks reconstruct how Italian high schoolers and international STEM experts perceive teachers, students, scientists, and school. *Educ. Sci.* **2020**, *10*, 17. [CrossRef]
43. Luchini, S.; Kenett, Y.N.; Zeitlen, D.C.; Christensen, A.P.; Ellis, D.M.; Brewer, G.A.; Beaty, R.E. Convergent thinking and insight problem solving relate to semantic memory network structure. *Think. Ski. Creat.* **2023**, *48*, 101277. [CrossRef]
44. De Deyne, S.; Navarro, D.J.; Perfors, A.; Brysbaert, M.; Storms, G. The "Small World of Words" English word association norms for over 12,000 cue words. *Behav. Res. Methods* **2019**, *51*, 987–1006. [CrossRef]
45. Citraro, S.; Vitevitch, M.S.; Stella, M.; Rossetti, G. Feature-rich multiplex lexical networks reveal mental strategies of early language learning. *Sci. Rep.* **2023**, *13*, 1474. [CrossRef]
46. Firth, J.R. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*; Blackwell: Oxford, UK, 1957.
47. Lenci, A. Distributional models of word meaning. *Annu. Rev. Linguist.* **2018**, *4*, 151–171. [CrossRef]
48. Posner, J.; Russell, J.A.; Peterson, B.S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **2005**, *17*, 715–734. [CrossRef] [PubMed]
49. Fillmore, C.J.; Baker, C.F. Frame semantics for text understanding. In Proceedings of the WordNet and Other Lexical Resources Workshop, NAACL, Pittsburgh, PA, USA, 3–4 June 2001; Volume 6.
50. Malandrakis, N.; Potamianos, A.; Iosif, E.; Narayanan, S. Distributional semantic models for affective text analysis. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2379–2392. [CrossRef]
51. Poquet, O.; Stella, M. Reviewing Theoretical and Generalizable Text Network Analysis: Forma Mentis Networks in Cognitive Science. *Proc. ISSN* **2022**, *1613*, 0073.
52. Mohammad, S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 174–184.
53. Mohammad, S.M.; Turney, P.D. Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
54. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
55. Hunt, T.E.; Clark-Carter, D.; Sheffield, D. The development and part validation of a UK scale for mathematics anxiety. *J. Psychoeduc. Assess.* **2011**, *29*, 455–466. [CrossRef]
56. Toumey, C.P. The moral character of mad scientists: A cultural critique of science. *Sci. Technol. Hum. Values* **1992**, *17*, 411–437. [CrossRef]
57. Soni, A.; Kumari, S. The role of parental math anxiety and math attitude in their children's math achievement. *Int. J. Sci. Math. Educ.* **2017**, *15*, 331–347. [CrossRef]
58. Necka, E.A.; Sokolowski, H.M.; Lyons, I.M. The role of self-math overlap in understanding math anxiety and the relation between math anxiety and performance. *Front. Psychol.* **2015**, *6*, 1543. [CrossRef]
59. Ashby, F.G.; Perrin, N.A. Toward a unified theory of similarity and recognition. *Psychol. Rev.* **1988**, *95*, 124. [CrossRef]
60. Aitchison, J. *Words in the Mind: An Introduction to the Mental Lexicon*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
61. Collins, A.M.; Loftus, E.F. A spreading-activation theory of semantic processing. *Psychol. Rev.* **1975**, *82*, 407. [CrossRef]
62. Hills, T.T.; Todd, P.M.; Goldstone, R.L. Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychol. Sci.* **2008**, *19*, 802–808. [CrossRef] [PubMed]

63. Siew, C.S. spreadr: An R package to simulate spreading activation in a network. *Behav. Res. Methods* **2019**, *51*, 910–929. [CrossRef] [PubMed]
64. Demetriou, A.; Spanoudis, G.; Makris, N.; Golino, H.; Kazi, S. Developmental reconstruction of cognitive ability: Interactions between executive, cognizance, and reasoning processes in childhood. *Cogn. Dev.* **2021**, *60*, 101124. [CrossRef]
65. Weidemann, C.T.; Kragel, J.E.; Lega, B.C.; Worrell, G.A.; Sperling, M.R.; Sharan, A.D.; Jobst, B.C.; Khadjevand, F.; Davis, K.A.; Wanda, P.A.; et al. Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *J. Exp. Psychol. Gen.* **2019**, *148*, 1. [CrossRef]
66. De Brigard, F.; Umanath, S.; Irish, M. Rethinking the distinction between episodic and semantic memory: Insights from the past, present, and future. *Mem. Cogn.* **2022**, *50*, 459–463. [CrossRef]
67. Hagendorff, T.; Fabi, S. Why we need biased AI: How including cognitive biases can enhance AI systems. *J. Exp. Theor. Artif. Intell.* **2023**, 1–14. [CrossRef]
68. Griffiths, T.L. Understanding Human Intelligence through Human Limitations. *Trends Cogn. Sci.* **2020**, *24*, 873–883. [CrossRef]
69. Mitrović, S.; Andreoletti, D.; Ayoub, O. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv* **2023**, arXiv:2301.13852.