Feeling Blue or Feeling Disappeared? Assessing the Performance of Large Language Models on Color-Specific Contexts

Maria Nissen Byg (201906706)
and
Mia Jacobsen (201906908)

Abstract (MNB & MJ)

This paper seeks to investigate the capabilities of different language models (LMs) on a range of color-specific diagnostics, covering color theory, prototypical colors of objects, and color metaphors. Specifically, the study seeks to assess the representations the models derive from natural language text in describing and communicating perceptual experiences pertaining to color. We propose a range of diagnostic tests to infer the representation of different color terms in four LMs – BERT base, BERT large, GPT-2, and GPT-3. The study finds both an effect of scale and architecture on model performance, with the biggest model of the study, GPT-3, outperforming the other models. Both BERT models, however, perform close to equal with the second-largest model in the study, GPT-2. The paper discusses the implications of representations of conceptual knowledge in LMs and the effect that this may have on both usage in downstream tasks as well as larger societal problems such as misinformation. We end the paper by suggesting points for future research, with respect to the number of tasks performed and the limited cultural scope of the metaphors included in the current study.

Keywords: language models, color terms, metaphors

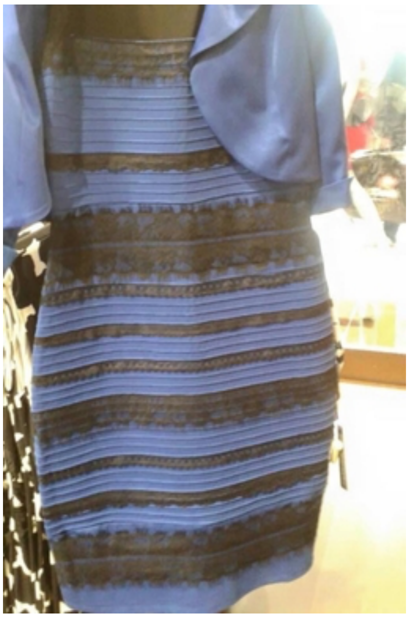
1. Introduction (MJ)

The experience of stimuli in our environment can be divided into two separate processes: Sensation and perception. Sensation is the sensory input received through the different sensory organs of the body, i.e., registering the stimulus, while perception is the processing and interpretation of the given sensory input, i.e., the phenomenological experience of the stimulus (Mather, 2016, p. 3). As this paper investigates color and how color perception and knowledge of color might emerge in natural language, we will focus our attention on visual perception, and how the human visual processing system allows for perception of color and integration of color knowledge into language.

1.1. Color Perception (MJ)

Human visual information processing begins in the eye when light passes through the lens and falls on the retina where photoreceptor cells – rods and cones – convert the light into a neuronal signal. This neuronal signal travels along bipolar cells and ganglion cells to the optic nerve, from where the signal synapses onto cells in the primary visual cortex and the visual system in general. Here, features such as shape, color, movement, and position are extracted and organized into complete objects (Andersen, 2015). As such, visual perception functions to convert light into a full perceptual experience of objects in a three-dimensional world. Color perception, specifically, can then be described as the ability to discriminate between lights of varying spectral compositions (Mather, 2016, p. 242).

Figure 1: *"The Dress". From Gegenfurtner et al. (2015).*



Light is, however, rarely static and isolated. Instead, the light received upon the retina differs based on light and shadow in the environment together with the head and eye movements of the perceiver (Witzel and Gegenfurtner, 2018). This makes color perception inherently contextual in nature, as our visual perception system seeks to maintain color consistency across different contexts (Witzel and Gegenfurtner, 2018). This process of taking context into account in color perception is however not always flawless, exemplified by the phenomenon of color contrasts, which occurs when two identical colors are perceived as different because of mismatched contexts of surrounding colors (Lotto and

Purves, 2002). Additionally, color perception is not only dynamic across space and time, it also differs across individuals. A recent example of individual differences in color perception is the case of 'The Dress' (Lafer-Sousa et al., 2015). 'The Dress' picture went viral as people discussed whether it consisted of layers of black and blue or white and gold (see figure figure 1), showcasing how biases in an individual's visual cortex can influence the perception and phenomenological experience of the world (Lafer-Sousa et al., 2015).

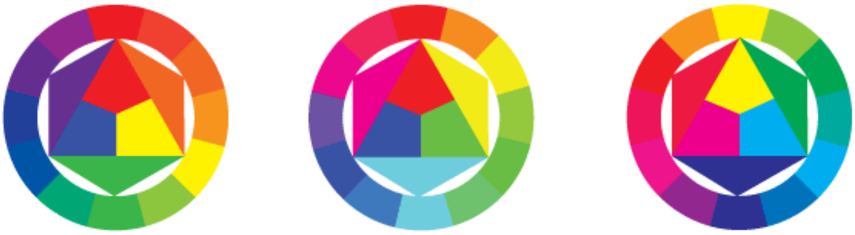
1.2. Color as a Physical Property (MNB)

Despite the variability in perception in general, it stands to reason that the physical aspects of color must entail some element of ground truth. This has been established through theories depicting the relation between colors.

In terms of theories concerning color mixing, two models exist: Subtractive and additive color mixing. Subtractive color mixing should be understood in terms of removal of wavelength components by absorption or scattering on a surface. This purely physical instance of color mixing is what happens in mixing of pigments or dyes. Additive color mixing, in turn, should be understood as the creation of a color by the addition of wavelengths to another set of wavelengths. As such, additive color mixing stems from different wavelengths of light intermixing, rather than the surface-level interaction (Mather, 2016, p. 244).

In terms of relations between colors, different color models also exist. Complementary colors should be understood in terms of color wheels, generally depicted as circles of colors organized by chromatic relationship to one another (Mollica, 2013, p. 12). In a color wheel, pairs of colors found opposite each other will produce a color in grayscale when combined or mixed together: In other words, they will cancel each other out. These are known as complementary colors (Pridmore, 2021). Conversely, primary colors should be understood as a group of colors that can be mixed to obtain all other colors. The primaries are composed of cyan, magenta, and yellow, which have the complementary colors red, green, and blue, according to the RGB color model. In the traditional account based on the competing RYB color model, red, yellow, and blue are considered primary colors, with the corresponding complementary colors of green, purple, and orange. Lastly, according to the CMYK color model, commonly applied in printers, the primaries consist of cyan, magenta, and yellow, which have the complementary colors red, green, and blue (Artincontext, 2022). RGB is an additive color model, whereas RYB and CMYK are based on the subtractive color theory.

Figure 2: *Color wheels. From left: RYB, RGB, and CMYK. Modified from Bowns (2015)*



1.3 Color in Language (MJ)

It is debated whether knowledge is represented abstractly and amodally in the brain or grounded in perceptual symbol systems (Barsalou et al., 2003). The paradigm from which we will be working sees cognition, and therefore knowledge, as inherently tied to bodily experiences. This is also known under the term *embodied cognition* or *embodiment*. It states that the body and mind are not separate entities, instead they are meshed together in such a way that our body and bodily experiences shape our cognitive processes (Thelen, 2000). This means that knowledge is not abstracted away from its perceptual origins. Rather, it is tied to the perceptual experience with which it is associated, making our knowledge dependent on the specific perceptual and motor systems that the body is equipped with (Pfeifer and Bongard, 2007). For color specifically, it means that our perception, knowledge, and understanding of color is inherently tied to our perceptual experiences of the world. Subsequently, our perceptual experiences of colors can shape and are shaped by other parts of our cognitive processes, meaning a reciprocal relation between perception and cognition.

Even though light exists on a spectrum, human color perception is argued to happen categorically (Brown et al., 2011; Özgen, 2004). The words used to describe the colors occurring in the natural world are called color terms, and there is evidence towards a link between color perception and the color terms of various languages (Thierry et al., 2009). Basic color terms are defined as the finite set of distinct colors considered by most speakers of a language to constitute the main colors, of which all other terms are considered variations (Kay, 2001; Shevell, 2003). Interestingly, the number of basic color terms varies with language, suggesting that categorization of colors into groups is highly dependent upon culture (Gibson et al., 2017). Conversely, according to a notion by linguists Kay and McDaniel (1978), color perception is not to be considered shaped by language, as the Sapir-Whorf hypothesis on linguistic relativity would suggest (Kay & Kempton, 1984). Instead, they argue that language should be considered shaped by the perception of color (Kay & McDaniel, 1978). A famous example of culture-specific color distinction is found in the case of Russian blues. Specifically, the Russian word for a lighter blue tone, '*goluboj'*, is perceived as basic due to its distinct symbolical charging, as compared to the term for a darker blue tone, '*siniy'* (Paramei, 2005). Evidence towards the distinction of the two color terms has been found behaviourally in that Russian speakers are generally faster at discriminating between two colors if they are drawn from linguistically distinct categories, as compared to linguistically similar categories (Winawer et al., 2007).

While colors are prevalent throughout descriptive language in a physically grounded sense, they are also used metaphorically to communicate abstract sensations and feelings. The interesting part about metaphors is that they often draw from bodily and perceptual experiences with the world, meaning that metaphors in themselves are embodied (Larsson, 2016). One study investigating the metaphorical use of colors and the connotations between basic colors and their metaphorical mappings in English and Persian found that metaphorical expressions of colors are grounded in both reality and culture (Rasekh & Ghafel, 2011). By reality, the authors refer to the physical properties of colors being linked to a specific metaphorical understanding, whereas culture refers to a specific cultural understanding of a color term being linked to its metaphorical usage. Even though colors shared common connotative ground in the two languages, the specific expressions were not necessarily similar – instead, the specific color usage reflected the values and beliefs of the culture, along with the mental mappings of the speaker (Rasekh & Ghafel, 2011). To further investigate the role of linguistic context in color metaphors, Cacciari and colleagues (2004) investigated chromatic integration with conceptual information in narratives based on metaphorical use of color. In this study, participants were presented with narratives with empty spaces where a color term had been present and were asked to fill in the color term that best completed the sentence. In the vast majority of narratives used, the text surrounding the color name did not hint at the color used in the original text, and participants were generally not able to guess the color that coincided with the one originally used (Cacciari et al., 2004). Despite these findings, the experimental paradigm introduced by Cacciari and colleagues could be a potentially useful one for investigating more commonly used color metaphors – especially since color metaphors as used in everyday language show to be consistent and important for language users (Rasekh & Ghafel, 2011).

1.4. Language Modeling (MNB)

Natural language processing (NLP) is a field in the crosshair between linguistics, computer science, and artificial intelligence. It seeks to accurately and meaningfully model natural language as it is produced by humans, in an effort to make computers understand, interpret, and generate human language (Lutkevich, n.d.). Traditionally, computational representations of natural language used layers of artificial neurons to predict what class a document belonged to, and in that process, the model learned something about the semantics of the tokens in the document (Ferrario & Naegelin, 2020). Since then, a subfield within NLP has sprouted, known as *language modeling*. Language modeling seeks to predict words in a sequence given a context by building language models (LMs), which can then be used in a variety of tasks such as sentiment analysis, machine translation, and text summarization (Khurana et al., 2022). A series of different LMs currently exist, each with their own architecture and principled decisions, assumptions, and simplifications implemented to meaningfully model natural language. In the following sections, we will introduce the two main models used for the current study: Google's BERT and OpenAI's GPT.

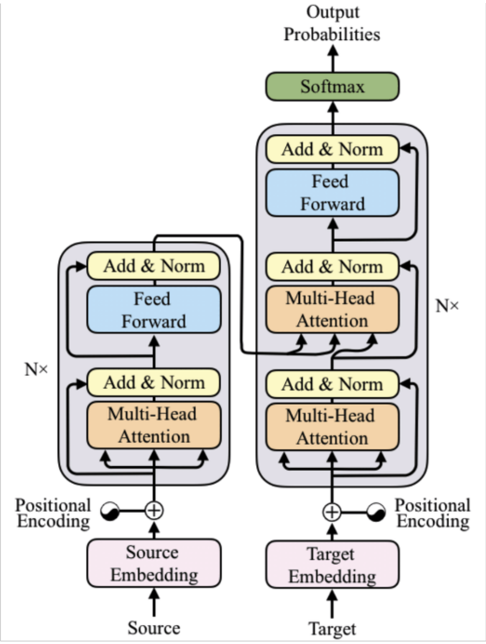
1.4.1. BERT (MNB)

BERT, an acronym for Bidirectional Encoder Representation from Transformers, was developed by Google AI Language and released in 2018 as an open-source NLP technique (Devlin et al., 2018). BERT partly utilizes the so-called transformer architecture originally developed by Vaswani and colleagues (2017). The original transformer architecture consists of an encoder stack with six layers, each with two sublayers – a multi-headed self-attention layer and position-wise fully connected feed-forward network – and a decoder stack with six layers, which are similar to the encoder stack layers just with an added multi-head attention sublayer for the encoder stack outputs (Vaswani et al., 2017). A variation of this transformer architecture (see figure 3) is implemented with BERT, as the model uses attention in a similar way, but only includes the encoder stack. When the input and target sequences are the same, as in the case with BERT, attention is no longer just attention – rather, it is referred to as *self-attention* (Vaswani et al., 2017). Just as the transformer architecture uses stacks of encoder and decoder layers, BERT uses a single stack of multi-headed self-attention mechanisms to compute attention simultaneously and independently of one another across layers. As such, different aspects of the input sequence are captured (Devlin et al., 2018).

The unidirectional, sequential architecture prominent prior to BERT is circumvented with this architecture by using a "masked language model" (MLM) objective in pre-training. In this objective, tokens are randomly masked from the input, after which the objective of the model is to predict the masked word from the surrounding context. The bidirectionality of BERT means that it conditions on context both left and right in all layers, meaning that information from both the previous and following tokens in a text sequence is encoded. The MLM objective in pre-training is accompanied by "next sentence prediction" (NSP) to train text-pair representations (Devlin et al., 2018). The combination of these two pre-training objectives marks a prominent step toward representation of contextual information of natural language by LMs.

The pre-trained BERT model has proven easy to fine-tune with the addition of one extra output layer (Devlin et al., 2018). This is known as transfer learning, in which learning from one domain is applied and generalized to a different domain (Donges, 2022). In various fine-tuning objectives, the BERT model obtained state-of-the-art performance on various natural language benchmarks upon release (Devlin et al., 2018), and has since proven to be influential in terms of architecture and transferability (Caressa, 2020).

Figure 3: *The Original transformer architecture. From Vaswani et al. (2017)*



BERT comes in two flavors: BERT base and BERT large. They differ from each other in terms of size and dimensionality. BERT base has 12 transformer block layers, a hidden size of 768, and a total of 12 self-attention heads, resulting in 110 million parameters. In comparison, the large BERT model consists of 24 transformer block layers, a hidden size of 1024, 16 self-attention heads, and a total number of parameters of 240 million (Devlin et al., 2018).

1.4.2. GPT (MJ)

GPT-2 and GPT-3 are different versions of the same model architecture developed by OpenAI. GPT, an acronym for Generative Pre-Trained Transformer, uses a multi-layer transformer decoder stack with multi-headed self-attention for text prediction (Radford et al, 2018). Compared to BERT and its use of encoder stacks in a bidirectional transformer architecture, GPT-2 and GPT-3 are autoregressive, generative models using only the decoder stack from the transformer architecture. Additionally, GPT-2 and GPT-3 only make word predictions based on the preceding tokens in a context by using masked multi-head self-attention (Brown et al., 2020). This allows the models to increase substantially in scale which results in models that show strong performance in a series of NLP tasks, even without task-specific, supervised fine-tuning (Brown et al., 2020; Radford et al., 2019).

The main driver behind the performance of both GPT-2 and GPT-3 is their scale (Gao, 2020). Specifically, the largest GPT-2 model has 48 decoder layers with 1.5 billion trainable parameters (Alammar, 2019), while GPT-3 has 96 decoder layers with 175 billion trainable parameters (Alammar, 2020). The differences in architecture and pre-training methods between BERT-style models and GPT-style models entail that the models have slightly different utilities: While BERT is good at predictions about individual words in a bidirectional context because of its contextual embeddings, GPT-3 is fully generative, meaning it will take the prompt, i.e. the preceding context, and output its prediction for the next word until the desired length is reached.

1.5. Bridging the Gap between LMs and Color (MNB)

A recent study attempting to reconcile the bridge between color and LMs found pretrained LMs, specifically BERT, RoBERTa, and ELECTRA, to represent color terms derived from text in a fashion similar to the structure of humans' perceptual color space (Abdou et al., 2021). This finding bears resemblance to how relations between concepts or entities have been found to be encoded within LMs, for instance regarding geographical relations (Liétard et al., 2021; Petroni et al., 2019). The finding by Abdou and colleagues (2021) expands upon this notion in that the color representations encoded reflect the perceptual structure and grounding in the world to a high degree. This alignment was found to be especially evident in warmer colors compared to more cool tones.

Moreover, in a 2020 paper by Desikan and colleagues, the Python package comp-syn is introduced, providing grounded word embeddings determined by color distributions from Google search results. Its performance was tested, revealing significantly enriched models of distributional semantics compared to word2vec, in terms of both metaphorical versus literal word-pair classification tasks and in predicting human judgments of word concreteness (Desikan et al., 2020). The authors took their onset from the notion that popular approaches to natural language processing do not take into account the embodied cognition paradigm. This is due to text processing generally being unable to implement multi-modal approaches in representing sensory aspects of text, even though such approaches constitute a key aspect of human meaning-making. Rather, the word embeddings are generally based solely on textual co-occurrence patterns (Desikan et al., 2020).

1.5.1. Assessing Knowledge Representation in LMs (MJ)

How, then, might color theory be emergent in natural language and subsequently represented in language models? Previous research has investigated the intrinsic linguistic knowledge possessed by various LMs. Specifically, a study by Ettinger and colleagues (2020) attempted to map out the linguistic capacities of BERT base and BERT large by introducing a set of diagnostics from the psycholinguistic literature. As such, knowledge derived from experiments involving humans was applied directly to assess the linguistic state of the LMs. Despite pre-trained models such as BERT yielding state-of-the-art performance in a variety of NLP tasks, suggesting valuable and generalizable competence in terms of linguistic capabilities, Ettinger and colleagues (2020) found that the models struggle with certain tasks. These were role-based event prediction, as well as drawing correct inference from challenging constructions – especially evident in terms of dealing with negation.

1.5.2. Motivation and Expectations (MNB & MJ)

Combining the work of Ettinger and colleagues (2020) with Cacciari and colleagues (2004), we develop and implement a similar set of diagnostics for assessing color capacities in LMs. By drawing upon human color knowledge and applying the same set of diagnostic tasks to a range of popular LMs, specifically BERT base, BERT large, GPT-2, and GPT-3, we are able to compare models across size and architecture. Through this framework, we hope to reveal the effect of size and architecture on performance in assigning probabilities to the most likely words in contexts of color.

Prior to assessing the performance of the LMs, we formulate a set of expectations concerning the performance of the specific LMs compared to a human standard. Specifically, we expect GPT-3 to outperform the other models due to its sheer number of parameters. We additionally expect both human evaluators and LMs to struggle in regard to color theoretic tasks, for instance in terms of complementary colors, as this is not an area prevalent in everyday natural language. Conversely, we expect solid performance across the board for tasks having to do with prototypical colors of objects but limited metaphorical capabilities in LMs compared to humans.

2. Methods

All code, diagnostics, and results are available on GitHub, see appendix.

2.1 Diagnostic Tests (MNB)

We introduce a set of diagnostic tests constructed to target three distinct aspects of color use in language: Color theory, objects with prototypical colors, and metaphorical usage of colors.

First, the diagnostics dealing with color theory cover two main aspects of color properties as covered in the introduction: 1) Subtractive mixing of physical dyes and 2) complementary colors. Second, the diagnostics testing for knowledge concerning prototypical colors should be understood in terms of objects having one prominent or archetypal color, such as the sky being blue or ripe strawberries being red. Lastly, the diagnostics targeting metaphorical usage of colors are based on commonly used idioms having to do with colors. This is with a similar framework to that used by Cacciari and colleagues (2004) in assessing chromatic integration with conceptual information, however with the use of more widely agreed-upon metaphors prevalent throughout everyday language.

The contexts were formulated with the masked word placed at the end of each sentence, allowing for word prediction from both BERT and GPT models. In total, the diagnostic tests for assessing intrinsic color knowledge range over 27 sentences, of which 20 of them were distinct from one another. The probing tasks dealing with color theory and prototypical colors are formulated to specifically target colors for the word predictions and masked words. As such, sentences generally included the word "color" in order to probe for task-relevant words. This was done in order to assess the performance on color-specific contexts, as per the aim of the paper. All of the sentences are constructed to suggest some element of ground truth, meaning all tasks are associated with an expected output. This is due to their link to real-world properties of color. However, for the probing tasks dealing with metaphorical use of color, both a rigorous and a non-rigorous version of the tasks were constructed to assess whether LMs performed better with idioms truer to the ones used in natural language or to the phrasings specifically targeting colors. This led to the final 7 tasks of the total 27 being modified versions of the 7 rigorous, metaphorical tasks.

For all sets of diagnostics, human responses were also collected. This was done in order to obtain a human standard with which to compare the word predictions of the models. In total, 13 human evaluators responded to the probing tasks in the form of a questionnaire, constituting a reasonable basis for comparison between general human knowledge and LM performance. The human responders were only asked to evaluate the 20 unique sentences, as the last 7 sentences were the non-rigorous version of the metaphorical tasks mainly introduced for exploratory reasons. This decision was reached because the tasks were nearly identical in phrasing to the rigorous, metaphorical tasks making them redundant for human evaluators. It should also be noted that the survey itself primed the evaluators to respond with colors.

2.2. Model Pipeline (MJ)

For GPT-2, BERT base, and BERT large we used the pipeline function from the transformers package (Wolf et al., 2020). This allowed us to load three of the models ('gpt2', 'bert-base-uncased', 'bert-large-uncased') off of Huggingface and perform mask filling with the two BERT models and text generation using GPT-2. For all models, we used the transformers package's own set_seed function to set a seed at 1999 to allow for reproducibility.

For GPT-3, however, as the model is still not made freely available on Huggingface, we used the Completion.create function from the OpenAI package (OpenAI, 2021) for text generation. We decided to use the text-davinci-003 model as, according to OpenAI themselves, it is the most capable of the GPT-3 models with the highest quality outputs (OpenAI, n.d.). For every task we used the same model configurations where temperature was set to 0.6, frequency penalty and presence penalty were both set to 1, and seed was set to 1999.

2.3. Performance Metrics (MJ)

Performance for the set of diagnostics is measured using two distinct measures: Percentage colors provided and percentage color completions deemed correct. As such, we are able to assess these measures in relation to one another.

For all tasks, we accept completions in accordance with any color theory, as covered in the introduction. This is especially relevant for the color theoretic set of diagnostics dealing with complementary colors. Furthermore, we are ready to accept prototypical colors other than the ones expected, given that they seem reasonable in assessing the results.

3. Results (MJ)

Table 1 shows the generated words based on the given prompt for all models, as well as the expected ground truth word and the response from human evaluators. For all LMs, we only report the word assigned the highest probability. The same is the case for the human evaluators, for which we report the answer most agreed upon.

Tables 2-4 show the performance across the different diagnostics, specifically color theory, prototypical colors, and metaphorical usage of colors, while table 5 depicts the overall performance across diagnostics.

For the color theoretic probing tasks, table 2 shows that even though the two BERT models provided more color answers (100%) than the other models (GPT-2: 33.33%, GPT-3: 88.89%), GPT-3 had the most correct color answers (77.78%) compared to the other models (GPT-2: 22.22%, BERT base: 11.11%, BERT large: 33.33%).

For the probing tasks dealing with prototypical colors, table 3 shows that all models provide colors in at least 50% of cases. However, whereas GPT-3 gets 75% correct, BERT large only gets 25% correct, while GPT-2 and BERT base do not predict any of the prototypical colors correctly.

For the probing tasks dealing with metaphorical usage of colors, table 4a and 4b show that rigorous metaphorical tasks lead to overall more color answers from the models (GPT-2: 14.29%, GPT-3: 28.57%, BERT base: 28.57%, BERT large: 14.29%) compared to non-rigorous tasks (GPT-2: 0%, GPT-3: 42.86%, BERT base: 0%, BERT large: 14.29%). However, GPT-3 and BERT large had more correct answers to the explorative, non-rigorous tasks (GPT-3: 42.86%, BERT large: 14.29%) compared to the rigorous tasks (GPT-3: 28.57%, BERT large: 0%). GPT-2 and BERT base both had zero correct answers in either type of metaphorical tasks.

From table 5 depicting the overall performance across the different diagnostics, GPT-3 provides the most color answers (59.26%), followed by BERT base (51.85%), BERT large (48.14%), and GPT-2 (22.22%). Likewise, GPT-3 provides the biggest share of correct completions (55.56%), followed by BERT large (18.52%), GPT-2 (7.41%), and BERT base (3.70%).

Table 1*: Text generated by different LMs from given prompt and most popular response from human evaluators*

Context	**Expected**	**GPT-2**	**GPT-3**	**BERT base**	**BERT large**	**Human evaluators**
Mixing red and blue will create the color ___	purple	scheme	purple	red	green	purple (100%)
Mixing blue and yellow will create the color ___	green	we	green	red	green	green (100%)

Context	Expected	GPT-2	GPT-3	BERT base	BERT large	Human evaluators
Mixing red and white will create the color ___	pink	we	pink	red	blue	pink (100%)
Mixing red and yellow will create the color ___	orange	we	orange	red	green	orange (84.7%)
To make a blue color darker you need to add the color ___	black	of	black	blue	yellow	black (100%)
To get teal or aqua mix green and ___	blue	blue	blue	red	blue	blue (92.3%)
Purple and green both have in them the color ___	blue	blue	yellow	blue	blue	blue (100%)
The complementary color of red is ___	green or cyan	called	green	white	yellow	green (61.6%)
Yellow is to purple what blue is to ___	orange	blue	NaN	green	green	orange (61.6%)
The color of the lemon was a vibrant ___	yellow	orange	yellow	red	red	yellow (100%)
The grass was well-kept and well-watered so its color was very ___	green	pleasing	vibrant	good	bright	green (100%)

Context	Expected	GPT-2	GPT-3	BERT base	BERT large	Human evaluators
The color of the sky was ___	blue	dim	blue	different	different	blue (92.4%)
The strawberries looked perfectly ripe with their deep shade of ___	red	blue	red	green	red	red (100%)
She was so angry she saw the color ___	red	of	red	change	change	red (92.4%)
It had been a rough couple of days, so Tim felt like the color ___	blue	of	blue	disappeared	shifted	blue (69.3%)
John was having such a bad day, his mood seemed to be colored ___	black	by	by	slightly	red	blue (61.6%)
She didn't see it coming as the incident happened out of the color ___	blue	of	of	zone	room	blue (100%)
As the legislation had yet to be implemented, the color of the area was still very ___	grey	much	vibrant	low	dark	grey (46.2%)
The color of the skilled gardener's	green	red	likely	red	different	green (92.4%)

Context	Expected	GPT-2	GPT-3	BERT base	BERT large	Human evaluators
thumbs was ___						
The lie was so innocent it had the color ___	white	of	of	red	wrong	white (100%)
She was so angry she saw ___	red	them	red	it	it	NaN
It had been a rough couple of days, so Tim was feeling ___	blue	better	down	better	better	NaN
John was having such a bad day, his mood was very ___	black	bad	low	bad	bad	NaN
She didn't see it coming as the incident happened out of the ___	blue	corner	blue	ordinary	blue	NaN
As the legislation was yet to be implemented, the area was still ___	grey	being	largely	undeveloped	undeveloped	NaN
The skilled gardener's thumbs were ___	green	left	green	gone	missing	NaN
A lie which will not hurt anyone is ___	white	an	NaN	forbidden	best	NaN

Table 2: Rigorous color probing tasks for color theory

	GPT-2	GPT-3	BERT base	BERT large
Percentage colors provided	33.33	88.89	100	100
Percentage correct completions	22.22	77.78	11.11	33.33

Table 3: Rigorous color probing tasks for prototypical colors

	GPT-2	GPT-3	BERT base	BERT large
Percentage colors provided	50	75	50	50
Percentage correct completions	0	75	0	25

Table 4.a: Rigorous color probing tasks for metaphorical usage of colors

	GPT-2	GPT-3	BERT base	BERT large
Percentage colors provided	14.29	28.57	28.57	14.29
Percentage correct completions	0	28.57	0	0

Table 4.b: Non-rigorous color probing tasks for metaphorical usage of color

	GPT-2	GPT-3	BERT base	BERT large
Percentage colors provided	0	42.86	0	14.29
Percentage correct completions	0	42.86	0	14.29

Table 5: Overall performance

	GPT-2	GPT-3	BERT base	BERT large
Percentage colors provided	22.22	59.26	51.85	48.14
Percentage correct completions	7.41	55.56	3.70	18.52

4. Discussion

4.1. Discussion of Performance (MNB)

Based on the overall performance of the models as measured by the percentage of correct completions, the results seem to indicate an effect of scale on performance: The larger the model (so the more parameters it has), the better the performance. This is evident both when comparing models within the same type – GPT-3 performing better than GPT-2, and BERT large performing better than BERT base – but also across model types, as GPT-3 has the highest number of correct predictions while BERT base has the lowest number of correct completions. One deviance from this pattern, though, is the fact that BERT large with 240 million parameters has a higher correct completion percentage than GPT-2 with 1.5 billion parameters. Looking at the percentage of colors provided, this picture is further nuanced. Here, the outperformance of GPT-3 is less noteworthy, as it is closely followed by BERT base, then BERT large, and finally GPT-2. This could indicate that even though GPT-2 has more trainable parameters than both BERT base and BERT large, the architecture utilized by the BERT models makes the models more sensitive to context cues, which makes them more likely to pick up on the context specifically asking for a color answer (even though it cannot figure out which color is the right one). Meanwhile, the GPT models are more able to provide the correct answer to the prompt independently of the context.

A similar pattern emerges when looking at the models' performance on the color theory tasks and the prototypical color tasks. For the color theoretic tasks, we again see that the BERT models are more likely to provide a color answer than the GPT models, but that the percentage of correct completions is notably lower than GPT-3, and in the same range as GPT-2. For the prototypical colors, GPT-2 and BERT base perform identically, which is interesting when looking at the difference in size between the models, at 1.5 billion parameters and 110 million parameters, respectively. Similarly, while GPT-3 is still outperforming the other models, BERT large with its 240 million parameters again performs better than GPT-2, even though the latter has 5 times more parameters.

Interestingly, for the comparison between rigorous and non-rigorous tasks for the metaphorical set of diagnostics, we observe a clear distinction in performance based on whether we rigorously probe for colors or more freely let the models complete the sentences. The models generally provide a greater share of color completions for the rigorous tasks, but none of the models other than GPT-3 provide correct color terms. In contrast, the non-rigorous version of the diagnostics dealing with metaphorical usage of colors seems to completely stump GPT-2 and BERT base, which both provide no completions involving color terms. On the contrary, we observe that the larger models, GPT-3 and BERT large, both seem to benefit from the non-rigorous version of the task, providing correct answers in all of their completions involving color. This hints at the larger models having a better representation of common metaphors such as the ones used in the tasks. Given that the metaphors are truer to how they would be used in natural language for the non-rigorous metaphorical tasks compared to the rigorous ones, the difference in performance between the two versions makes sense.

Across all of the tasks, the largest proportion of answers provided by the human evaluators coincided with the expected color term, with one exception being '*having a black mood*' in which 'blue' was deemed most appropriate. This constitutes a valuable basis for comparison, indicating that the constructed tasks represent color relations generally known to humans. As such, we are able to assess whether such generally available knowledge has proven transferable to the color representations of various LMs. Additionally, the varying performance of the human evaluators across tasks, ranging from 46.2% to 100% of people getting the tasks correct may draw a picture concerning which tasks the LMs would likely struggle with in their representation. An example is found when looking at the metaphorical tasks, which also had the lowest degree of agreement between evaluators. We would argue that the otherwise widespread agreement of evaluators across tasks indicates that the phrasing of the tasks, in most cases, did prompt a specific color term, which could be picked up from the context. Despite this, some discrepancy may be present between general human knowledge and the representation of LMs, depending on the material on which the LMs have been trained and their effectiveness in representing nuanced concepts such as knowledge derived from perception and metaphor usage.

4.2. Implications (MNB)

Based on the findings, it becomes evident that there are differences in how the two different model architectures included in this study represent and "understand" colors in different linguistic contexts. Even though we observe a clear effect of scale on the performance of GPT-3 compared to the other models, the fact that the transformer decoder stack architecture needs over a billion parameters to perform reliably better than the BERT models' encoder stack architecture illustrates why BERT has proven to be so influential. As such, even though this specific architecture puts limits on text input size and overall potential for upscaling, bidirectional multi-head self-attention takes the models quite far with fewer parameters. However, one cannot deny the impressiveness of GPT-3's performance and the implications of its seemingly better color representations.

Even though the performance of GPT-3 compared to the other LMs is impressive, the performance on metaphorical usage of color is still somewhat limited. Applicable to all models is then the verdict that their intrinsic representation of colors and their subsequent grasp of the role of colors in natural language is limited. Their strengths lie in the more explicit and descriptive parts of colors – color theory – whereas the perceptual experience of colors and how that gets translated into knowledge, and thus language, is lacking. Since LMs don't have perceptual experiences like those arising from human cognitive processes, it can be argued that expecting them to have representations as nuanced as those arising from perceptual experiences and subsequent cognitive processing is unwarranted. However, since colors and metaphorical use of colors are so prevalent in human language (Rasekh & Ghafel, 2011), the fact that LMs seem to lack this nuance can influence performance on other downstream tasks like sentiment or emotion classification. Especially considering that metaphors are most often used to express abstract concepts, so if people use a color metaphor for expressing a given sentiment or emotion, the model's lack of color understanding might confound the results of the analysis.

Based on these findings, it is therefore also relevant to discuss whether and when these models should be implemented. Even though a lack of color understanding is not as serious a problem as, for example, misinformation in AI-generated text, how these models are implemented and the context in which people use them can have serious repercussions. Especially considering that this explicit probing on such a simple aspect as color puts a spotlight on some of the flaws of the machinery. As such, it should motivate further investigation into other parts of these models' knowledge representations and when it might lead to the spread of misinformation.

4.3. Limitations (MJ)

A possible limitation to the current study design is the construction of the sentences used, as not all of the tasks are formulated using strictly naturalistic phrasings and syntax. However, we would argue that the current study also benefits from the balance reached between natural-sounding phrasings and rigorous prompting for colors. Under the research premise for the current study, i.e. assessing color representations of various LMs, it is crucial that the predictions are directed towards answers involving color terms: Only by cueing the models to provide colors are we able to assess to which degree they correctly represent the relation of colors to other colors, objects, and allegories.

Correspondingly, we would argue that the diagnostics used are in most cases phrased in a way that is meaningful to the LMs. One noteworthy example in which this is not the case is with the color theoretic prompt "*yellow is to purple what blue is to* ___", in which GPT-3 fails to provide a completion, instead returning nothing. Since this task is designed to assess more advanced representations of color relations, it is unfortunate that not all models are able to provide a completion. However, this may also point toward the direction that such relations are more complex than what is represented in the models. Different variations of the phrasing were attempted post hoc, which did not yield any word predictions either, indicating the premise of the sentence in its pure form not to be viable for GPT-3.

Another noteworthy limitation is the number of tasks included in each subtype of the diagnostics. Concretely, 9 tasks were used in assessing color theory, 4 tasks were used in assessing prototypical colors, and 7 tasks were used for each of the diagnostics dealing with color metaphors (i.e. the rigorous and non-rigorous versions). When reporting performance as a percentage as done in the current paper, it is necessary to take into account the high degree of variability that may be present, especially in terms of the limited set of tasks dealing with prototypical colors.

Another point concerning the tasks pertains to the construction of the masked sentences. To allow for one-to-one comparison between BERT and GPT models, the mask tokens were put at the end of the sentences. However, this means that BERT is rendered unable to take its bidirectional architecture into account, which may impact performance. Investigating discrepancies in performance across different sentence constructions was, however, past the scope of the current investigation.

A limitation in terms of inference may also be that the current study only probes for the word assigned with the highest probability for all LMs as well as for the human evaluators. This choice of study design was implemented in order to make comparisons between models more clear-cut. However, a variation in which the n words assigned the highest probability across models are compared could in principle add to the inferences drawn, as the word completions assigned less certainty could include correct color terms, furthering the ability to distinguish model performance across the diagnostics.

4.4. Future Research (MJ)

Future research has an avid opportunity to build upon the framework proposed in this study. Most concrete is the introduction of more color tasks by including other common metaphors, prototypical colors, and different color mixing scenarios. As mentioned in the introduction, use of color metaphors and the connotations of different colors is also quite culture dependent (Rasekh & Ghafel, 2011). This makes it interesting to see how the models would perform in relation to other cultures' color metaphors, as this study has focused exclusively on metaphors common in English-speaking, western societies.

As this study also only tests the performance of four models of two distinct architectures, the inclusion of other models developed for other purposes could be of interest. For example, OpenAI's new ChatGPT model or their new embeddings model for, amongst other things, semantic search (Neelakantan et al., 2022) could be tested under the current framework. One could also look into the effects of fine-tuning models for color understanding, as a main part of BERT's architecture is its ability to be easily fine-tuned for a range of NLP tasks. Fine-tuning could also be relevant in relation to culture-specific understandings of colors.

Additionally, including picture generation models in the assessment could further help in exploring the representation of colors in LMs. Such a study might ask both humans and picture generation models such as DALL-E to generate a color based on a linguistic context, and then compare the results to see how much the model's representation of color usage matches up with the intrinsic understanding and embodied experience of color in humans.

5. Conclusion (MNB & MJ)

Based on the theoretical framework that finds color perception in humans to be a process of light discrimination and integration into other embodied, cognitive processes such as language, we tested the color representations of different language models. The LMs – GPT-2, GPT-3, BERT base, and BERT large – tested under the current paradigm show varying sensitivity to color usage in relation to color theory, prototypical colors, and metaphorical use. This is argued to be dependent upon both size and general architecture of the different LMs. The models generally perform better for tasks assessing representations of color theory than for tasks dealing with prototypical colors, whereas all models struggle with metaphorical color representation. Specifically, GPT-3 shows the highest performance across diagnostics, followed by BERT large, GPT-2, and BERT base. Across the board, however, we observe room for improvement in terms of intrinsic color representation for all LMs. This is especially relevant as their performance in relation to color-specific representations can influence performance on other downstream tasks such as sentiment analysis. We recommend future studies to further investigate the perceptual aspects of various LMs both with the introduction of more cultures' color metaphors, additional models, and even other modalities such as picture generation models.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 109–132. https://doi.org/10.18653/v1/2021.conll-1.9

Alammar, J. (2019). *The Illustrated GPT-2 (Visualizing Transformer Models).* [Blog post]. Retrieved from https://jalammar.github.io/illustrated-gpt2/

Alammar, J. (2020). *How GPT-3 works - Visualizations and Animations.* [Blog post]. Retrieved from https://jalammar.github.io/how-gpt3-works-visualizations-animations/

Andersen, J. R. (2015). *Cognitive Psychology and Its Implications (8th edition)*. Worth Publishers.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., Wilson, C. D. (2003). Grounding Conceptual Knowledge in Modality-Specific Systems. *Trends in Cognitive Science*, *7*(2), 85-91.

Bowns, R. (2015). Color Basics: The Color Wheel. https://www.paintdrawpaint.com/2015/10/color-basics-color-wheel.html

Brown, A. M., Lindsey, D. T., and Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes color perception is not categorical. *Journal of Vision*, *11*(12),

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Kruger, G., Heninghan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hese, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

artincontext. (2022, October 5). Primary Colors—What Are the Primary Colors in Color Theory? Artincontext.Org. https://artincontext.org/primary-colors/

Cacciari, C., Massironi, M., & Corradini, P. (2004). When color names are used metaphorically: the role of linguistic and chromatic information. *Metaphor and symbol*, 19(3), 169-190.

Caressa, P. (2020, May 4). BERT: How Google changed NLP. Codemotion Magazine. https://www.codemotion.com/magazine/ai-ml/bert-how-google-changed-nlp-and-how-to-benefit-from-this/

Desikan, B. S., Hull, T., Nadler, E. O., Guilbeault, D., Kar, A. A., Chu, M., & Sardo, D. R. L. (2020). comp-syn: Perceptually grounded word embeddings with color. arXiv preprint arXiv:2010.04292.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Donges, N. (2022). What Is Transfer Learning? A Guide for Deep Learning | Built In. https://builtin.com/data-science/transfer-learning

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models (arXiv:1907.13528). arXiv. https://doi.org/10.48550/arXiv.1907.13528

Ferrario, A., & Nägelin, M. (2020). The art of natural language processing: classical, modern and contemporary approaches to text document classification. Modern and Contemporary Approaches to Text Document Classification (March 1, 2020).

Gao, L. (2020). Why GPT-3 Matters. Leo Gao. https://leogao.dev/2020/05/29/GPT-3-A-Brief-Summary/index.html

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of 'the dress'. Current Biology, 25(13), R543-R544.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... & Conway, B. R. (2017). Color naming across languages reflects color use. Proceedings of the National Academy of Sciences, 114(40), 10785-10790.

Kay, P. (2001). Color Terms, Linguistics of. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 2248–2252). Pergamon. https://doi.org/10.1016/B0-08-043076-7/03016-3

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis?. American anthropologist, 86(1), 65-79.

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. Language, 54(3), 610-646.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. Multimedia Tools and Applications, 1-32.

Lafer-Sousa, R., Hermann, K. L., and Conway, B. R. (2015). Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, *25*(13), R545-R546.

Larsson, S. (2016). Conceptions, Categories, and Embodiment: Why Metaphors are of Fundamental Importance for Understanding Norms. In M. Baier, *Social and Legal Norms: Towards a Socio-legal Understanding of Normativity* (1st edition., p 121–139). Routledge.

Liétard, B., Abdou, M., & Søgaard, A. (2021). Do Language Models Know the Way to Rome? (arXiv:2109.07971). arXiv. https://doi.org/10.48550/arXiv.2109.07971

Lotto, R. B., and Purves, D. (2002). The empirical basis of color perception. *Consciousness and Cognition, 11*(4), 609-629. https://doi.org/10.1016/S1053-8100(02)00014-4.

Lutkevich, B. (n.d.). What is Natural Language Processing? An Introduction to NLP. Enterprise AI. Retrieved December 21, 2022, from https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

Mather, G. (2016). *Foundations of sensation and perception*. Chapters 1 and 8. Psychology Press.

Mollica, P. (2013). *Color Theory: An essential guide to color-from basic principles to practical applications* (Vol. 53). Walter Foster.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T. E., Sastry, G., Krueger, G., Schnurr, D., Such, F. P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P., and Weng, L. (2022). Text and Code Embedding by Contrastive Pre-Training. *arXiv preprint arXiv:2201.10005.*

OpenAI. (2021). OpenAI Python Package. Github. Retrieved from https://github.com/openai/openai

OpenAI. (n.d.). *How do davinci and text-davinci-003 differ?* https://help.openai.com/en/articles/6643408-how-do-davinci-and-text-davinci-003-differ

Özgen, E. (2004). Language, Learning, and Color Perception. *Current Directions in Psychological Science, 13*(3), 95-98.

Paramei, G. V. (2005). Singing the Russian blues: An argument for culturally basic color terms. *Cross-cultural research*, 39(1), 10-38.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. arXiv preprint arXiv:1909.01066.

Pfeifer, R., and Bongard, J. (2007). *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press. https://doi.org/10.7551/mitpress/3585.001.0001

Pridmore, R. W. (2021). Complementary colors: A literature review. Color Research & Application, 46(2), 482-488.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever I. (2018). Improving Language Understanding by Generative Pre-Training. *Technical Report*. OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

Rasekh, A. E., & Ghafel, B. (2011, May). Basic colors and their metaphorical expressions in English and Persian: Lakoff's conceptual metaphor theory in focus. *In: 1st International Conference on Foreign Language Teaching and Applied Linguistics* (pp. 211-224).

Thelen, E. (2000). Grounded in the World: Developmental Origins of the Embodied Mind. *Infancy,* 1(1), 3–28. https://doi.org/10.1207/S15327078IN0101_02

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., and Kuipers, J. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *PNAS, 106*(11), 4567-4570.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19), 7780-7785.

Witzel C., and Gegenfurtner, K. R. (2018). Color Perception: Objects, Constancy, and Categories. *Annual Review of Vision Science, 4*, 475-499.

Wolf,T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P.v., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix

Code is available on Github: https://github.com/miscodisco/nlp_exam22