### AI and Writing: Do we speak the words of stochastic parrots?

Kenneth C. Arnold (Calvin University)



Ken Arnold (B.S., Cornell; M.S., MIT; Ph.D., Harvard) is an assistant professor of computer science and data science at Calvin University. His research focuses on understanding how AI-augmented interactive systems, like the predictive text interfaces found in smartphone keyboards and email apps, can shape human communication and creativity. His work is driven by a commitment to augment, not replace, human intelligence and caring. He is also interested in exploring how AI can augment the work of educators.

#### Abstract

Generative Artificial Intelligence systems have recently and publicly become increasingly comparable to the capabilities of many humans in some domains, such as text and image generation. These systems are rapidly being incorporated into human contexts, for example as phone keyboards and email apps offer us AI-generated suggestions of words and even complete phrases to enter with a single tap or swipe. But these suggestions don't just reduce physical effort; they reduce cognitive effort: we can accept a suggested phrase before even thinking about what to say. In this article, I first describe how mimicry pervades the design and implementation of generative AI systems and expound on some of the implications thereof. I then critique how these systems are designed and implemented in light of our God-given identity, a biblical perspective on mimicry, and a reflection on virtue. I finally outline several "swords-to-plowshares" approaches that re-form generative AI technologies to support human thinking in writing tasks: *structural editing tools, writer-directed questions* to support writers without speaking for them, and *visualizations* to enable writers to reflect on their work.

### 1 Introduction

AI systems for generating and suggesting content have become commonplace. Phone keyboards suggest the words and even phrases we might write next. Search engines suggest queries. Other systems generate images, music, or code based on prompts.<sup>1</sup> Recent systems have reached unprecedented levels of quality in their outputs, as Figures 2 and 1 illustrate. As this article goes to press, news outlets are reporting "The End of High-School English"<sup>2</sup> in response to the surprising level of quality of text generated by a language model (in this case, ChatGPT by OpenAI).

Despite the superficial novelty of the content generated by these systems, generative AI systems mimic their training data: they recombine elements of the examples they've seen, and even that recombination is according to tropes of how elements are typically combined. That is, mimicry is deeply woven into the design of interactive AI systems, both in the underlying models (and the

<sup>&</sup>lt;sup>1</sup>Generative AI can be distinguished from the pervasive recommendations provided on social media sites, video sharing sites, shopping sites, etc., since those systems recommend *existing* content that is (typically) created by other people. These recommendations also have enormous influence on what we buy, who we socialize with, and even what we believe. But the *generative* technology I discuss here generates new content, rather than selecting from among content that other people have already created.

<sup>&</sup>lt;sup>2</sup>https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/ 672412/, see also https://www.theatlantic.com/technology/archive/2022/12/ chatgpt-ai-writing-college-student-essays/672371/

associated mathematics, algorithms, and data) and in the design of people's interactions with those models. Recent scholarship has identified ways in which these so-called "stochastic parrots" can be harmful [2]; in this article I extend that critique to the interactive interfaces built on those systems and how we use those systems in creating the texts we use to communicate with each other. I argue that, although mimicry is useful in some cases, we must think deeply about how to both *build* and *use* these systems in ways that align more deeply with our God-given identity as co-creators who are called to love and serve God and our neighbors with thankfulness, humility, and integrity. In the following sections, I:

- Summarize how interactive generative AI systems work, in terms of data and computation,
- Draw implications about these systems from a Christian world view, and
- Propose some ideas for repurposing the underlying technology redemptively.

## 2 How interactive generative AI systems work

We begin with a systems-level overview of generative AI. The two main takeaways of this section are: (1) AI systems are mathematical not magical, and (2) a bias towards mimicry is deeply embedded into the entire life cycle of these systems. Informally, the generative modeling approach has a "groove" that it is following, the bias that the approach embeds.

We limit our scope here to what is technically called *self-supervised learning*, meaning that the system is trained to perform a prediction task that is based only on unlabeled examples. We will focus here on language modeling using Transformer-based large language models (sometimes abbreviated LLMs), but most of the comments will also apply to other language tasks and to image generation tasks using Generative Adversarial Networks (GANs), diffusion models such as DALL-E and Stable Diffusion, and others. Some models deviate slightly from this recipe by tuning the models based on human feedback; this sort of tuning, called Reinforcement Learning from Human Feedback, allows selecting for desirable behaviors among the set of behaviors that the model has learned.[3]; this was the approach used to make the headline-grabbing ChatGPT.<sup>3</sup>

<sup>3</sup>https://openai.com/blog/chatgpt/

**Input**: I was supposed to start writing the paper at 5:00 PM. But then I started playing with this cool new language model for 10 minutes. 10 minutes later, it's suddenly 9:30 PM!

**Model Output**: This joke is about how time flies when you're having fun. The person was supposed to start writing a paper at 5:00 PM, but they started playing with a language model instead. They thought they were only playing with it for 10 minutes, but it was actually 4.5 hours.

Figure 1: The PaLM language model (a type of generative AI system) explains a joke, after being given only 2 examples of joke explanations. Example from the research paper published by Google [1]; see there for more details.

#### 2.1 Collect Data

Building a generative AI system starts with **collecting data**. The beginning is typically a very large collection of text. For example, a model recently described by Google was trained on hundreds of billions of words drawn from publicly available text, such as web pages, books, Wikipedia articles, news articles, source code for computer programs, and conversations on social media platforms [1].

#### 2.2 Train a Model of the Data Distribution

The dataset is then typically used to train a statistical model of the distribution of training data. That is, the model is a function that can compute a probability for any possible text.<sup>4</sup> *Intuitively*, a "good" model is one that assigns larger probabilities to texts that are more likely to occur in its training data. The model's function has many parameters (e.g., 540 billion, for Google's PaLM model [1]) that are used in that computation. For practical reasons, the probability computation is typically performed sequentially through a text, one token (word, punctuation symbol, or sub-word unit) at a time, thus computing the probability as a product of conditional probabilities:

$$P(text) = P(word1)P(word2|word1)P(word3|word1,word2)...$$

The model can then be trained by an iterative process: give the model a text up until a certain word, ask it to compute the discrete probability distribution of all possible words that come next, then modify the parameters of the model so that it assigns a slightly greater probability to the word that did come next. In other words, the model becomes *less surprised* by the sequences of words that occur in its training documents. *Formally*, the model is trained using stochastic gradient descent to minimize the average cross-entropy loss across every token (word, subword, punctuation, etc.) in the training data, one small batch at a time; see [4] for details.

#### 2.3 Fine-Tune a Foundation Model

However, few developers train models from scratch. Instead, they download a large pre-trained "foundation" model, then *fine-tune* that model for their specific domain or task, if necessary. Most pre-trained models are released by large organizations such as Google, Microsoft, or NVIDIA, but some have been trained by academic and independent research groups.<sup>5</sup> Fine-tuning can be done either by further training the model on new data or by crafting ways that the existing model can be queried in ways that encourage certain types of behavior (such as "prefix tuning" or "prompt engineering").<sup>6</sup> For example, the joke explanation of Figure 1 was created by prepending to the given joke two examples of a joke followed by an explanation, and appending "Explanation:" to the prompt; this *prompt engineering* caused the model to generate words that would fit the pattern, i.e., an explanation of the preceding joke. Even though joke explanation was not explicitly anticipated in its data collection or training process, the corpus of text that the model was trained on presumably had sufficient examples of jokes followed by explanations that the model learned to utilize this pattern.

 $<sup>^{4}</sup>$ Encoder-only models like BERT work slightly differently; they only need to assign a probability to *parts* of a text; they perform a "fill-in-the-blanks" task.

<sup>&</sup>lt;sup>5</sup>Examples include https://bigscience.huggingface.co/ and https://www.eleuther.ai/.

 $<sup>^{6}</sup>$ Federated Learning can be used to incorporate private data while limiting the degree to which the learned model can memorize private data.

### 2.4 Query the Model Interactively

Interactive interfaces developed on top of these mathematical models tend to align with the task that the model is trained to do. Since the underlying mathematical models are based on modeling the distribution of existing examples, the interfaces built on them tend to offer samples from that distribution. For example, for systems powered by text models trained to predict the next word, the output presented by the system is typically a plausible continuation of a given text. These continuations may be presented as individual words (for smartphone keyboards), next phrases (for systems like Google's Smart Compose), or larger units like paragraphs. (Some models are instead trained as *translators* or *summarizers*, sometimes called *seq2seq* or sequence-to-sequence models; although it may seem fundamentally different, this setting can almost always be mapped to a next-word prediction task with a suitable prefix.)

One of the ways that generative systems can support people is as a creative partner, in a setting known as human-AI co-creativity [5, 6, 7]. These systems can support people in various ways. Digital artists can iteratively explore and refine creative concepts by exploring different prompts and partial images they give to systems like DALL-E or Stable Diffusion (see example in Figure 2).<sup>7</sup> Creative writers can collaborate with large language models like GPT-3 to explore new ideas, such as names for characters in a story [8, 9].

Generative systems can also support people through facilitating repetitive or predictable tasks. For example, systems can generate fluent emails given a rough summary of what to communicate, generate blog posts about desired topics, or summarize long articles.<sup>8</sup>

## 3 Implications of How Generative AIs Work

### 3.1 Since they're based on data...

Since generative AI systems are entirely based on data, the availability of training data powers the behavior of the system. Thus, the system needs no explicit instructions about any of the specific tasks it is to perform, in contrast to decades of manual effort spent towards manually engineering domain-specific features or laboriously encoding expert knowledge.

However, the data also *limits* the behavior of the system. For example, data, at best, reflects the world as it *is*, which implicitly allows systems to stereotype on the basis of categories like race, gender, or religion. And data is rarely a faithful representation of the world as it is; datasets include biases of selection, representation, aggregation, and evaluation,<sup>9</sup> which lead to biases in system behavior. For example, language models may associate Islam with terrorism and Black people with crime.<sup>10</sup> When system behavior is dictated by data, system designers have limited ability to embed in the system any notion of the way the world *should* be, such as discouraging such stereotypes.

Since current approaches require enormous datasets, organizations with more data of sufficient quality can train better models. For example, Google's sibling DeepMind trained a high-performing

<sup>&</sup>lt;sup>7</sup>See, for example, the DALL-E Prompt Book, https://dallery.gallery/the-dalle-2-prompt-book/

<sup>&</sup>lt;sup>8</sup>Products utilizing these approaches include Flowrite and Jarvis.

 $<sup>^9\</sup>mathrm{Many}$  systematic overviews are available for the interested reader, for example, the fast.ai Data Ethics online course

<sup>&</sup>lt;sup>10</sup>O'Sullivan, Liz; Dickerson, John P. (August 7, 2020). "Here are a few ways GPT-3 can go wrong"

model on 1.7 trillion tokens, including a large collection of "books," presumably from Google's large collection of scans. An open-source dataset called The Pile [10], in contrast, contains a smaller collection of books, which will necessarily limit the capabilities of systems trained on it.

### 3.2 Since they're trained to minimize surprise on that data...

For the mathematically inclined, it should be refreshing and even encouraging to see that the systems that generate seemingly impressive results are, at their core, college-level calculus (the chain rule), linear algebra (matrix multiplication), probability (conditioning on discrete events), and numerical methods (gradient descent) applied to large amounts of data. The functions employed have complex behavior but are composed of simple primitives; for example, the so-called Self-Attention module at the core of the popular Transformer model uses little more than matrix multiplication, element-wise exponentiation, and dividing a vector by the sum of its elements.

The word-by-word ("autoregressive") nature of language models, which may be key to their success, aligns with how God made the world: step by step. He could have spoken the world into being in a single act, but instead he constructed it in methodical sequence, one aspect at a time. Human co-creators likewise generally construct step by step, refining partial ideas.

Since the model's performance is measured on averages, the model can perform badly in uncommon situations without its creators noticing. Thus, parameter settings that cause the model to work well in common situations may have unintentional adverse effects [11].

### 3.3 Since many systems leverage big pretrained models...

The availability of large foundation models has presented many new opportunities by allowing practitioners to obtain good performance without extensive training data, but researchers have urged caution in using them because any models based on a foundation model inherit its flaws, which in most cases are poorly understood [12]. The use of foundation models also contributes to *homogenization* of AI system behavior, making many different systems all prone to exhibit similar flaws and biases. It also further disconnects those who use the model from the original authors of the work (text, images, etc.) that the algorithm was trained on, making it more difficult to practice *thankfulness* to the authors of the work.

### 3.4 Since they usually show people generated samples...

Efficient text entry can be desirable: it takes some of the tedium out of entering text. Predictive text interfaces are one way of doing this, but contemporary interaction techniques only work for the most highly predictable text, so it tends to nudge writers towards entering such text. Decoder-based approaches such as gesture typing [13], auto correction [14] and speech recognition post-editing work for a much broader range of predictability, but fail abruptly for the least predictable text (such as names), forcing users to make awkward transitions to different interaction techniques. New interfaces for text entry are regularly presented at conferences on human-computer interaction, but creative new interaction techniques are still needed in text entry.

Since interactive systems tend to offer samples from a distribution of existing examples, they tend to encourage users of those systems to create things that are likewise comparable to existing examples.

The "groove" for the technology becomes a "groove" for the users: it will be easiest to say things that have been said, to continue in the tracks forged by others.

For example, my own studies of writers found that word and phrase predictions, though typically presented as simply ways to enter text more easily, can nudge writers to write content that conforms to the expectations of the system along aspects of word choice, level of detail, and even sentiment [15, 16, 17].

In many cases it's good for our creations to be grounded in what has already been created. The more unusual or unexpected the words of a message are, the more difficult it is for others to understand. Even God's communication with humanity has worked this way, using words, analogies, and even patterns of documents (like covenants) that were culturally familiar to the original audience. Although Jesus's words had unique content and authority, he spoke about himself by referring to the Law and Prophets, a foundation of understanding that many of his listeners shared.

But imitation of past examples can become excessive. Predictive text suggestions can lead us to communicate platitudes or even seemingly thoughtful statements without thinking: consider a system suggesting "I hope you have a wonderful weekend" at the end of a message, or "How adorable!" in response to a friend sharing a baby picture. Our words don't just communicate information; they communicate human relationship; acknowledging the Image of the speaking God in the other. Generative AI systems have become, alongside social media and other developments, more ways that we are able to mimic the external signs of thoughtfulness and care without having exhibited those virtues.

When aiming to maximize use, mimicry looks like success (more use of the suggestions). So although systems like Google's Smart Compose might have stated aims to increase user satisfaction, some of the easiest metrics (such as number of suggestions accepted or number of words inserted) could be maximized by finding how to nudge people into following the system's lead more.

### 3.5 Is mimicry good?

We have seen that mimicry is woven throughout the design and development of generative AI systems. How should we evaluate mimicry? A consideration of the Biblical narrative shows that we are encouraged to practice mimicry, but with discernment.

Imitation is not the primary thrust of the Biblical narrative; more focus is given to accurate perception of and action toward God and others. Nevertheless, the Biblical narrative presents humans as mimics from the very beginning. The first one to speak in Genesis is God the creator; when man the creature finally spoke in Genesis 2, verses 19 (in naming every living creature) and 23 (in delighting in his companion), he echoed the words of God, who had previously given names to things and called them good.

The Bible authors commend imitation of God and of wise behavior. For example, Proverbs encourages imitating the diligence of ants (Proverbs 6:6) and to walk in the paths of wisdom. Paul frequently gives himself as an example to imitate (e.g., 1 Cor 4:16, 11:1, 1 Thes 1:6, 2 Thes 3:7-12). Other epistles tell their readers to imitate those who are faithful and righteous (3 John 11, Hebrews 6:12). The ultimate goal of all of these commands is that we imitate God (Eph 5:1, 1 Cor 11:1).

However, imitation must be done with discernment. Even imitating God's actions might not be

wise, as illustrated by the Egyptian magicians in response to the signs that Moses used to show God's authority over the oppression of that country's leader. They are somehow able to mimic the staff-to-snake, water-to-blood, and frog signs—but their mimicry only worsened the trouble for their country (more frogs!); they were not able to *reverse* the plague. Eventually the magicians concede that the signs are "the finger of God" (Exodus 8:19). More commonly, the people of Israel are frequently tempted to imitate the ungodly practices of other people groups, such as those that inhabited the land of Canaan (e.g., Deut 18:9). We could view the entire Pentateuch as an extended reflection on what to imitate and what not to imitate, towards the goal of creating a place and community on Earth that is an image of God's heavenly dwelling.

#### 3.6 Since the systems hide all the nuance behind push-button interfaces...

Beyond text entry, generative AI models are also very frequently used for *inspiring* writers. But they do so in ways that hide and disconnect us from the *people* whose ideas and effort the system is leveraging. For example, OpenAI provides a "playground" interface for its customers to interact with GPT-3. This interface presents the system's generation as a magic writing demon, including showing the generation happening one word at a time as if it were a "ghost in the machine". But in fact these words are generated using patterns that the model learned from reading many examples of other people's work. Issues of copyright have been raised about this system,<sup>11</sup> and more prominently with image generation system such as OpenAI's DALL-E or Stable Diffusion<sup>12</sup> In the case of image generation, researchers have found that generative image models would even learn to add watermarks like Shutterstock's to their generated images; current systems have not thoroughly engaged the questions of attribution that arise. Christians might articulate that we want to be able to practice gratitude towards the people whose work our work is based on. The task of examplebased explanation has seen some study in classification tasks [18, 19], but has not been sufficiently explored for generative AI systems.

Educators may rightly be concerned about the availability of generative AI systems, especially for writing both natural language and programming language, as these systems could be used to generate assignment solutions of reasonable quality that do not reflect the student's understanding or effort. For example, language models can generate acceptable and diverse solutions to common programming exercises that do not trigger suspicion from plagiarism detection tools [20]. Full discussion of this issue is beyond the scope of this article, but in short, the situation is similar to that posed by the introduction of graphing calculators and computer algebra systems in mathematics instruction.

#### 3.7 A brief normative reflection on generative AI

One line of Christian scholarship about technology is organized around Christian philosopher Herman Dooyeweerd's "modal aspects," which help avoid reductionism by prompting reflection along a wide range of aspects that cannot be reduced to each other. For many aspects, Christian scholars have identified corresponding *design norms* that reflect implications of a biblical worldview in each of these aspects [21]. A comprehensive consideration of the design norms applied to generative AI

<sup>&</sup>lt;sup>11</sup>See GPT-3 – Copyright Law and Power - Dr Michael Guihot - YouTube

<sup>&</sup>lt;sup>12</sup>See DALL-E works with the images of creators who do not receive anything in return: what copyright says about AI - Digis Mak and Is DALL-E's art borrowed or stolen? | Engadget.

is beyond the scope of this paper, but I will mention a few observations and questions that we might raise based on some of these norms.

- *cultural norm*: What effect will generative models have on textual and visual media (and so on culture) overall? As generating text becomes cheaper, will the value of text further decrease? As increasing amounts of the text we read and images we see are generated at least partially by models that are trained to mimic the status quo, will our collective thinking be broadened or narrowed? What effects could these systems have on how people with differing literacy or language proficiency interact with language?
- *clarity norm*: Language has long been used to mislead, but will large language models increase the scale or effectiveness of disinformation?<sup>13</sup> On the other hand, might better language technology give us better tools for finding clarity through summarization, better search, etc.?
- *social norm*: Text that is kind and contextually appropriate has served as a signal of care and thoughtfulness in written communication. But if those signals can be generated automatically, will they lose their meaning? If the people we're communicating with suspect that our text is computer-generated (even if it is not), will they trust us less? The nascent field of AI-Mediated Communication studies (AI-MC) explores these and related questions.
- *stewardship norm*: With computation and storage demands of models only increasing, will generative AI technology waste energy? (Historical precedent suggests that power-efficient implementations will come.) How will we curate quality training data?
- *harmony norm* (or *beauty norm*): How will our cultural sensibilities of what is beautiful and valuable change as it becomes easier to produce artifacts that we once deemed as high-quality? How can tools work harmoniously with the human process of writing, rather than intruding on it?
- *justice norm*: Will the benefits of AI technology be distributed equitably? Will those whose jobs are transformed or replaced by AI advances be equipped to shift to different work? As these systems become used in sense-making and decision-making contexts, will they amplify, perpetuate, or even create biases towards or against some groups of people?
- *trust norm*: Of what sort of trust can the outputs of large language models be worthy? How can we assess their reliability?

## 4 Redeeming Artificial Parrots

Although many applications of generative AI can be positive, the general trend towards better mimicry leads to concern about the future impacts of these systems. But rather than abandon these systems, I encourage a "swords into plowshares" (Isaiah 2:3-5) approach, where we re-mold existing systems to serve new goals. In particular, we can leverage the extensive research and computational effort that has gone into large pre-trained language models such as GPT-3. Although these models were initially trained for mimicry and are often used to substitute for human thinking, they can be repurposed to empower and encourage human thinking. I discuss two potential approaches below and provide examples from my own research.

<sup>&</sup>lt;sup>13</sup>OpenAI recently published a report from a workshop they convened on the topic: https://openai.com/blog/forecasting-misuse/



Figure 2: Example of creative use of a modern AI system that can generate or refine images based on textual descriptions. A full description of the artist's process is given at the source: https://andys.page/posts/how-to-draw/.

### 4.1 Empowering Writers with Structural Editing and Visualization Tools

One approach to redeeming generative models is to use them to power *editing* and *visualization* tools that give creators more powerful tools to understand and reshape their work. These tools present the opportunity for writers to remold and reshape their work without ceding authorship of any part of that work to an AI system.

Low-level revision tools, typically for grammatical error correction [22] or contextual spelling correction, typically suggest edits that can be accepted or rejected, but the edits are typically at a small scale, which helpfully constrains their potential influence on the content and makes it easy for writers to critically evaluate the suggestion. Other interfaces allow exploration of alternatives to single words using a contextual thesaurus [23].

Other tools allow manipulation at a broader level, but at the cost of writers' control. For example, a system can generate paraphrases of a writer's sentence, perhaps to match a desired style [24], but writers' control over these paraphrases is typically indirect. Still others provide indirect controls over the system's edits [25, 26], but we seek direct control. Systems like OREO [27] can flexibly reword selected spans of text to better achieve a writer's goal, although the specification of that goal is difficult.

In contrast, Dang et al. consider providing summaries of each paragraph in the document margin [28]. Writers found that these summaries provided a perspective on their writing that helped them revise. Although the system generates text, it does so in the process of summarizing the writer's own work. Compared with approaches that complete partial texts, this approach reduces the risk of the system hallucinating information or making its own point. Nevertheless, writers still observed a difference between their own writing and the style of the generated text, and sometimes adapted the summaries into revisions of their own work. Dong et al. deliberately designed their interface to avoid implying that the system's outputs were suggestions to be accepted. That choice reduces the coercion for a writer to appropriate the system's words as their own.

The approach of developing AI-powered editing tools has been more common in visual domains than text. Adobe has recently released Neural Filters for Photoshop that give visual artists a new palette of brushes and sliders to shape their work, leveraging familiar interaction patterns in a new Yellowstone National Park was established by the US government in 1872 as the world's first legislated effort at nature conservation.



Figure 3: A selection interaction for clause reordering within a given sentence. The example (unedited output of a prototype system by Calvin students April Volzer and Noah Madrid) shows five different possibilities for which phrase opens the sentence. Phrases are given colored backgrounds to visualize the relationships between the alternatives. Each alternative can be expanded to show alternative completions of that phrase.

way. While the result is clearly artificial, the process utilizes generative AI models in ways that keep much of the control of the result in the hands of the artist. However, some current work in image editing does so in ways that leave substantial creative direction to the system rather than the artist. For example, GANPaint [29] gives artists "brushes" that cause the model to make localized edits to the brushed area of the image, to hallucinate the presence of some desired attribute (such as clouds or trees). Although the approach took a major step towards providing more direct control over the otherwise black-box process of image generation, further work was needed to give artists more explicit control over how the system enacts the desired edits. As text understanding and image generation both become more powerful, opportunities emerge to give more precise editing instructions, for example as demonstrated by InstructPix2Pix [30]. However, the same advances allow artists to give even higher-level instructions of what edits to make, leaving the system with has considerable latitude in interpreting the instruction, which may reduce the creative agency of the artist.

To further explore interaction techniques and technical approaches for structural revision of documents, several undergraduate students and I built several prototypes for revising structure at the sentence level. Figure 4 shows a proposed direct manipulation interaction: the writer drags a selected phrase to a new location and the system performs the necessary edits on the rest of the sentence (highlighted in green in the figure). However, in this interaction it is not obvious to the writer which manipulations are likely to be successful, so we also explored interactions in which the system presents several plausible reorderings that the writer can choose among and then refine using further selection operations (Figure 3). We are currently working to extend these ideas to help writers revise the structure and flow of wider areas of documents.

### 4.2 Inspiring Writers by Asking Questions

Language models have learned patterns that characterize writing; how can human writers leverage that knowledge without the model writing for them? One approach my collaborators and I have explored is having the model generate *questions* for the writer, rather than text that they might appropriate themselves.

In one exploratory study, we compared the approach of giving guidance in the form of content questions ("Questions") with two alternatives: no guidance, and exemplar sentences ("Snippets") from high-quality related documents. We designed a scenario for which writers would need to write



*drop* It was earlier demonstrated by researchers that high stress can cause heart attacks.

Figure 4: Mock-up of an example direct manipulation interaction for rearranging a sentence by dragging and dropping a selected phrase. During dragging, non-content words are dimmed. Dropping the phrase constrains its relative position; green highlights indicate other words that the language model needed to modify or move.



Figure 5: Left: Prompts presented as Questions were more often marked as relevant by participants. Right: Participants chose the Questions prompt as most preferable along all three measures asked.

isolated sentences in encyclopedia-style writing with optional guidance from a "bot." Results of the study, summarized in Figure 5, suggested that while both types of guidance helped productivity compared to the baseline, writers found prompts shown as Questions gave them usable ideas more often than the same semantic content shown as Snippets. Writers also expressed strong preference for Questions over Snippets presentation. More details about the study can be found in our workshop paper [31].

We have also undertaken formative research (conducted Summer 2022, not yet published) to understand what sort of questions would be helpful to writers at various stages in their writing process (we contrasted drafting stage vs revision stage) and what sorts of interaction techniques might be useful to provide questions to writers. Based on a Wizard-of-Oz style study where participants wrote argumentative essays and other participants wrote questions for them, we identified categories of questions that could be helpful (typically those we classified as Open-Ended, Elaboration, and Structure), and found that useful questions were typically grounded in the existing text, without introducing many external ideas. We found that presenting these questions as word processor comments, like a human reviewer might leave, was perceived as helpful without being overly intrusive. Preliminary explorations of providing our participants' writings to language models encouraged us that we should be able to get language models to generate these sorts of questions.

## 5 Conclusion

Generative AI systems such as large language models are powerful mimics. Uncritical use of them risks pressuring humans into appropriating the mimic's words as our own. But people can choose to use them wisely, both as end users and as software developers.

As users, some people have chosen to disable predictive text on their phones and computer systems; others have chosen to leave them enabled but be thoughtful about when to look at and accept suggestions. When the system suggests a generic pleasantry, a writer could instead choose to write something more specific.

As developers of AI systems and of interactive applications that utilize them, we can consider how to mold these systems. Here are a few questions we might ask:

- How can a person's interaction with this system encourage their creativity and individuality?
- Are there ways to help writers celebrate the thought and effort that they have put into their work, instead of just trying to reduce thought and effort?
- What sort of writing is the system encouraging? Does the system make it easier or harder for that writing to love and serve the reader?
- How can an AI system highlight the contributions of the people whose work became the system's training data? Can it help people be thankful to those people? Can it give them credit?

I have discussed two ways of repurposing generative AI systems towards a vision of empowering writers. But many other ways of repurposing these systems are also possible. For example, I have informally started to explore using language models like GPT-3 to help me come up with ideas for educational materials that are highly contextualized to the needs of my class but also leveraging the insights of educators who have tried to teach similar materials before. Also, these models are typically used to attempt to model expert-level understanding of language or code; could they be repurposed to simulate a student's developing understanding of a concept or skill, at a sufficiently high level of detail that an educator could get useful feedback on what their students individually understand and how they will respond to potential examples?

It is an exciting time to be working in interactive artificial intelligence. I hope that this article inspires some Christians to work towards shaping this developing technology.

## References

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee,

Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, PaLM: Scaling Language Modeling with Pathways, 2022. doi:10.48550/arXiv.2204.02311. arXiv:2204.02311.

- [2] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–623. doi:10.1145/3442188.3445922.
- [3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. URL: http://arxiv.org/abs/2203.02155. doi:10.48550/arXiv.2203.02155. arXiv:2203.02155.
- M. Phuong, M. Hutter, Formal Algorithms for Transformers, 2022. URL: http://arxiv. org/abs/2207.09238. doi:10.48550/arXiv.2207.09238. arXiv:2207.09238.
- [5] N. Davis, Human-Computer Co-Creativity: Blending Human and Computational Creativity, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 9 (2013) 9–12.
- [6] J. Rezwana, M. L. Maher, Understanding User Perceptions, Collaborative Experience and User Engagement in Different Human-AI Interaction Designs for Co-Creative Systems, in: Creativity and Cognition, Association for Computing Machinery, New York, NY, USA, 2022, pp. 38–48. doi:10.1145/3527927.3532789.
- [7] P. Yurman, A. V. Reddy, Drawing Conversations Mediated by AI, in: Creativity and Cognition, Association for Computing Machinery, New York, NY, USA, 2022, pp. 56–70. doi:10.1145/3527927.3531448.
- [8] M. Lee, P. Liang, Q. Yang, CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities, in: CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–19. doi:10.1145/3491102.3502030. arXiv:2201.06796.
- [9] A. Calderwood, V. Qiu, K. I. Gero, L. B. Chilton, How novelists use generative language models, in: HAI-GEN Workshop at IUI 2020, 2020.
- [10] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800GB Dataset of Diverse Text for Language Modeling, 2021. URL: http://arxiv.org/abs/2101.00027. doi:10.48550/arXiv. 2101.00027. arXiv:2101.00027.
- [11] R. Balestriero, L. Bottou, Y. LeCun, The Effects of Regularization and Data Augmentation are Class Dependent, 2022. doi:10.48550/arXiv.2204.03632. arXiv:2204.03632.
- [12] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar,

F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the Opportunities and Risks of Foundation Models, 2022. doi:10.48550/arXiv.2108.07258.

- [13] S. Reyal, S. Zhai, P. O. Kristensson, Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild, Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15 (2015) 679–688. doi:10.1145/ 2702123.2702597.
- K. Vertanen, H. Memmi, J. Emge, S. Reyal, P. O. Kristensson, VelociTap: Investigating Fast Mobile Text Entry using Sentence-Based Decoding of Touchscreen Keyboard Input, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15, 2015, pp. 659–668. doi:10.1145/2702123.2702135.
- [15] K. C. Arnold, K. Z. Gajos, A. T. Kalai, On Suggesting Phrases vs. Predicting Words for Mobile Text Composition, Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16 (2016) 603–608. doi:10.1145/2984511.2984584.
- [16] K. C. Arnold, K. Chauncey, K. Z. Gajos, Sentiment Bias in Predictive Text Recommendations Results in Biased Writing, in: Graphics Interface 2018, Toronto, Ontario, Canada, 2018, pp. 8–11.
- [17] K. C. Arnold, K. Chauncey, K. Z. Gajos, Predictive text encourages predictable writing, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20, Association for Computing Machinery, Cagliari Italy, 2020, pp. 128–138. doi:10.1145/3377325.3377523.
- [18] W. Zhang, Z. Huang, Y. Zhu, G. Ye, X. Cui, F. Zhang, On Sample Based Explanation Methods for NLP :Efficiency, Faithfulness, and Semantic Evaluation, 2021. doi:10.48550/ arXiv.2106.04753. arXiv:2106.04753.
- [19] C. Molnar, Interpretable Machine Learning: A Guide For Making Black Box Models Explainable, Leanpub, 2022.
- [20] S. Biderman, E. Raff, Neural Language Models are Effective Plagiarists, 2022. doi:10.48550/ arXiv.2201.07406. arXiv:2201.07406.
- [21] E. J. Brue, D. C. Schuurman, S. H. VanderLeest, A Christian Field Guide to Technology for Engineers and Designers, Intervarsity Press, 2022. URL: https://www.ivpress.com/ a-christian-field-guide-to-technology-for-engineers-and-designers.
- [22] R. Grundkiewicz, C. Bryant, M. Felice, A crash course in automatic grammatical error correction, in: Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts, International Committee for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 33–38. doi:10.18653/v1/2020.coling-tutorials.6.

- [23] K. I. Gero, L. B. Chilton, How a stylistic, machine-generated thesaurus impacts a writer's process, in: Proceedings of the 2019 on Creativity and Cognition, C&C '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 597–603. doi:10.1145/3325480. 3326573.
- [24] K. Gero, C. Kedzie, J. Reeve, L. Chilton, Low level linguistic controls for style transfer and content preservation, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 208–218. doi:10.18653/v1/W19-8628.
- [25] R. Louie, A. Coenen, C. Z. Huang, M. Terry, C. J. Cai, Novice-AI music co-creation via AI-Steering tools for deep generative models, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–13.
- [26] A. Fan, D. Grangier, M. Auli, Controllable abstractive summarization, in: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, 2018, pp. 45–54.
- [27] J. Li, Z. Li, T. Ge, I. King, M. R. Lyu, Text Revision by On-the-Fly Representation Optimization, 2022. doi:10.48550/arXiv.2204.07359. arXiv:2204.07359.
- [28] H. Dang, K. Benharrak, F. Lehmann, D. Buschek, Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries, 2022. URL: http://arxiv.org/abs/ 2208.09323. doi:10.1145/3526113.3545672. arXiv:2208.09323.
- [29] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, A. Torralba, Semantic photo manipulation with a generative image prior, ACM Trans. Graph. 38 (2019). URL: https://doi.org/10.1145/3306346.3323023. doi:10.1145/3306346.3323023.
- [30] T. Brooks, A. Holynski, A. A. Efros, InstructPix2Pix: Learning to Follow Image Editing Instructions, 2023. URL: http://arxiv.org/abs/2211.09800. doi:10.48550/arXiv.2211. 09800. arXiv:2211.09800.
- [31] K. C. Arnold, A. M. Volzer, N. G. Madrid, Generative Models can Help Writers without Writing for Them, in: D. Glowacka, V. Krishnamurthy (Eds.), 2nd Workshop on Human-AI Co-Creation with Generative Models - HAI-GEN 2021, volume 2903 of CEUR Workshop Proceedings, CEUR, College Station, United States, 2021.

# Appendix

All text in italics was generated by the model text-davinci-003, accessed via the OpenAI API.

Title: Some tough questions at the intersection of generative AI and Christian faith. Author: GPT-3  $\,$ 

1. How does generative AI challenge traditional Christian understandings of human creativity and free will?

Generative AI presents a challenge to traditional Christian understandings of human creativity and free will by introducing a new kind of artificial intelligence that has the ability to create original content. Generative AI can generate text, images, music, and more, and it can do so without the need for a human programmer or artist. This raises the question of whether humans can still be considered the sole creators of their work and whether they have full control over their creative processes. It also raises questions of whether the use of generative AI can be ethically justified, particularly with respect to religious beliefs.

2. How does generative AI shape our understanding of our relationship with God?

Generative AI can reshape our understanding of our relationship with God by introducing a new way of expressing our creativity. While traditional Christian faith has emphasized the importance of humans as creative agents, generative AI can enable us to create new forms of art and expression that may be seen as a form of prayer or worship. In addition, generative AI can allow us to explore our relationship with God in ways that were not possible before. By allowing us to explore the boundaries of creativity, generative AI can open up new avenues for spiritual reflection and growth.

3. Are there ethical considerations that must be taken into account when using generative AI?

Yes, there are many ethical considerations that must be taken into account when using generative AI. These include questions of authorship and ownership, privacy, and the potential misuse of generative AI for malicious purposes. Additionally, ethical considerations must be taken into account when considering the impact of generative AI on human creativity, freedom, and the potential for the proliferation of "fake news" and other forms of misinformation.

4. How might Christians use generative AI to further their faith?

Christians can use generative AI to further their faith in a variety of ways. For example, Christians can use generative AI to create new forms of art and expression that can be used to explore and express their spiritual beliefs. Generative AI can also be used to generate new kinds of sermons and reflections that can be used to explore and deepen faith. Additionally, generative AI can be used to generate new types of music, poetry, and other forms of expression that can be used to explore and express faith.

Explain generative AI.

Generative AI is a type of AI technology which attempts to generate new data based on existing data sets. For example, generative AI could be used to create new images, music, or text based on existing data sets. Generative AI can be used to create novel and creative outputs, as well as help in the development of more complex AI systems by providing additional training data.

Translate into pop song lyrics.

Verse 1: Generative AI, a technology so new It can generate data, create something brand new Create images, music, and text so fine It can help create a complex AI design

Chorus: Generative AI, it's the way to go It can create something special, you just gotta know Generative AI, it's the wave of the future It will help create something brand new, you'll be sure

Verse 2: Generative AI, it's the way to go It can help us create something we never thought we could know Novel and creative, yeah it's the thing It'll help us create a whole new kind of thing