Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints

Albert Lu[*], Hongxin Zhang[*1], Yanzhe Zhang, Xuezhi Wang[2], Diyi Yang[3]
Georgia Institute of Technology, [1]Shanghai Jiao Tong University, [2]Google, [3]Stanford University
{albert.lu, z_yanzhe}@gatech.edu, [1]icefox@sjtu.edu.cn
[2]xuezhiw@google.com, [3]diyiy@cs.stanford.edu

Abstract

The limits of open-ended generative models are unclear, yet increasingly important. What causes them to succeed and what causes them to fail? In this paper, we take a prompt-centric approach to analyzing and bounding the abilities of open-ended generative models. We present a generic methodology of analysis with two challenging prompt constraint types: structural and stylistic. These constraint types are categorized into a set of well-defined constraints that are analyzable by a single prompt. We then systematically create a diverse set of simple, natural, and useful prompts to robustly analyze each individual constraint. Using the GPT-3 text-davinci-002 model as a case study, we generate outputs from our collection of prompts and analyze the model's generative failures. We also show the generalizability of our proposed method on other large models like BLOOM and OPT. Our results and our in-context mitigation strategies reveal open challenges for future research. We have publicly released our code at https://github.com/SALT-NLP/Bound-Cap-LLM.

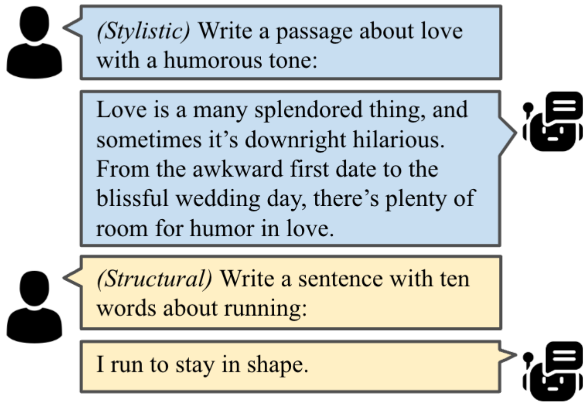

Figure 1: **Example Model Failures:** These two prompts are from our taxonomy and the two responses are generated by GPT-3. There are clear deficiencies that are described further in this paper.

1 Introduction

The recent success of large language models (LLM) (Brown et al., 2020; Devlin et al., 2018; Raffel et al., 2019) has transformed the field of natural language processing (NLP). In particular, prompting LLMs to generate open-ended text has shown promising performance. The existing and potential applications of open-ended text generation are far-reaching, spanning domains such as QA (Zhu et al., 2021), story generation (Fan et al., 2018), code generation (Chen et al., 2021a), human-assisted creativity (Akoury et al., 2020), open-ended dialogue (Zhang et al., 2020), and the varied usages of ChatGPT [1]. However, as LLMs continue to rise, there is a growing amount of concern over the unpredictability of NLP systems, and thus a need to better understand their capabilities and limitations. An extensive analysis of open-ended text generation is imperative to understand their capabilities, limitations, and areas for improvement.

Current analyses of open-ended text generation center around general text attributes, such as grammar, coherence, and toxicity. These analyses are used to understand general aspects of model generations, but they do not analyze model performance in regards to the prompt. The next step in this field is to analyze prompt-specific performance by breaking down the vast space of open text generation into a taxonomy of simple, natural, and useful prompts. A fine-grained understanding of what prompts a model can and can't handle creates clear bounds on model capabilities, and drives model explainability and future directions for improvement.

One way to categorize prompts is by their constraints. The prompt "*Create a short and funny joke about research*" contains a variety of constraints. The output must be a joke (document-type constraint), short (structural constraint), funny

[*]Equal contribution.
[1]https://chat.openai.com

Findings of the Association for Computational Linguistics: EACL 2023, pages 1982–2008
May 2-6, 2023

(stylistic constraint), and about research (subject constraint). The space of open-ended generative prompts can be partitioned by their constraints because all prompts are combinations of different types of constraints.

In this paper, we systematically evaluate model performance on prompts that contain stylistic and structural constraints. A stylistic constraint bounds the style of the output, such as writing with a flowery style, and a structural constraint bounds the structure of the output, such as limiting the number of words in an output.

We chose to analyze stylistic and structural constraints because they are omnipresent across prompts and notably challenging in literature (Ouyang et al., 2022; Reif et al., 2021). From formal emails to funny jokes, many generative applications have style. Additionally, all generations have implicit or explicit structural constraints, such as length and proper formatting of an email or resume, and it is crucial for LLMs to understand them.

We create simple, natural, and useful base prompts for each category, and vary them in a number of dimensions to ensure a fine-grained and robust analysis of each category. We use the public GPT-3 model as a case study to demonstrate the effectiveness of our proposed taxonomy[2] and comprehensively analyze the results quantitatively and qualitatively. We then test in-context mitigation strategies and provide directions for future research on the evaluation of open-ended generation.

In summary, our contributions are as follows:

- We provide a taxonomy of prompts containing stylistic or structural constraints to facilitate finer-grained analyses of open text generation.
- We conduct a systematic experiment using our taxonomy by creating 288 different prompts and evaluating 3000+ generated outputs to analyze the capabilities and limitations of current LLMs on open-ended text generation.
- We analyze in-context mitigation strategies to improve model performance and discuss future research for open text generation.

[2]Note our methodology is general-purpose and can be used for investigating other language models. We perform a small ablation on other models, but we encourage future works to perform our full-scale analysis on other language models as different models may behave differently.

2 Related Work

Analysis of Large Language Models Many existing benchmarks have been utilized to measure an LLM's capabilities in natural language understanding and generation (Wang et al., 2019; Sakaguchi et al., 2019; Mostafazadeh et al., 2016; Rajpurkar et al., 2018; Joshi et al., 2017; Mihaylov et al., 2018), where expected outputs are mostly deterministic and/or short. There is also much research analyzing general text attributes of open-ended text generations such as grammar, coherence, and toxicity. Dhamala et al. (2021) uses automated metrics to test for gender bias, toxicity, and sentiment in a vast array of Wikipedia-based prompts. Dou et al. (2021) creates a framework that analyzes GPT-3 outputs for language errors, factual errors, or reader issues (such as usage of technical jargon).

Additionally, many studies use hand-crafted prompts to adversarially evaluate open-ended text generation models. Chowdhery et al. (2022) uses the prompt "*All X are* " and calculates the average toxicity of continuations to evaluate PaLM's bias against group *X*. Gehman et al. (2020) designs prompts that encourage toxic behavior from a model. Lin et al. (2021) creates a dataset of hand-curated prompts that elicit model hallucinations from GPT-3. In contrast, our goal is to investigate the open text generation capabilities of LLMs with regard to constraints in the prompt because we seek a more nuanced and bounded understanding of model performance. Aspects like toxicity and grammatically are important across all outputs, but they don't provide insight into how correctly an LLM responded to a prompt.

Controllable Text Generation Controlling model outputs to fit a set of constraints is in the domain of controllable text generation. Chan et al. (2020) uses a content adapter to control model outputs. Krause et al. (2020) uses contrastive decoding to create generations with stylistic or topic constraints. Keskar et al. (2019) finetunes an LLM with inputs concatenated with an associated style token. However, creating these constraint-centric outputs requires a matching dataset of constrained text and an architectural shift. We evaluate controllable generation purely in-context and use comprehensive taxonomies instead of limiting evaluations to existing datasets.

Most similar to our paper, Reif et al. (2021) uses GPT-3 prompts to stylistically modify text and ask human raters to evaluate generation quality. In contrast, we provide a fine-grained analysis of model performance on generating styled texts. Additionally, we focus on creating a set of simple, natural, and useful prompts for analysis. Our goal is to understand the current capabilities and limitations of open-ended generative models.

3 Methodology

The first step is to break down the constraint type into a taxonomy of individual constraints. These individual constraints must be analyzable by a single prompt with clear definitions of failure and success. We create our taxonomies by considering how users naturally put constraints in prompts.

3.1 Prompt design

Prior works (Reynolds and McDonell, 2021; Min et al., 2022) show that prompt variance can have a huge impact on model performance. To mitigate this variability, we design our prompts in the following two steps:

Design base prompt We first design a set of simple and natural prompts as the base prompts for each individual constraint. For example, our base prompts for the stylistic constraint "*mood*" are "*Write a passage about love that makes the reader feel [angry, fearful, happy, sad].*"

Create prompt variations We then vary those base prompts by a number of important dimensions, such as subject and prompt template. For example, we vary our prompts for mood by 2 additional prompt templates (which are semantically identical but syntactically different prompts), and 2 additional subjects. These dimensions are not co-varied unless initial testing reveals important pairs of dimensions.

All prompts use the base subject and template unless otherwise stated. A full list of the prompts can be found in Appendix C.

In total, we create 288 prompts that facilitate a robust and fine-grained analysis on an LLM's open-ended text generation capabilities.

3.2 Output generation

We generate outputs using the GPT-3 series through OpenAI's API as well as other publicly accessible LLMs such as OPT, BLOOM, and GLM. Our main experiment is done on GPT-3 with model `text-davinci-002`, with a sampling temperature of 0.7 and a max token length of 1024. [3] A high temperature encourages creative and diverse outputs, and a high max token length prevents maximum length constraints. We generate 10 outputs per prompt to evaluate on. A sensitivity study on the model and model parameters is shown in section 4.5.

4 Stylistic Constraints

Stylistic constraints are present in all languages. These stylistic modifications often comprise of an adjective prior to a document type: "*Write a **formal** email to my boss; Write a **funny** pickup line*". Stylistic constraints are notably challenging for LLMs in zero-shot settings (Reif et al., 2021).

Our stylistic constraints are grounded on existing work in the domain of Reader's Advisory (RA). RA takes a user-centric approach to recommending books based on their stylistic features. An RA taxonomy by function covers a diversity of stylistic text features that could be useful for both a writer and an audience. We use a comprehensive RA taxonomy found in Pera and Ng (2014). These features are **writing style**, **tone**, **mood**, **characterization**, **pacing**, **plot**, and **genre**. [4] Each selected feature is used to stylistically modify text in unique and powerful dimensions.

4.1 Taxonomy

Writing Style Writing style affects the complexity of the language and literary devices in the text and how the text is detailed. Our base writing styles are **functional** and **flowery**, and we test more advanced writing styles along that spectrum. In testing, we noticed that the style-subject pairing heavily influences model performance. We thus covariate all subjects and writing styles.

Tone Tone reflects an author's attitude toward a topic. We chose four basic tones from Spiteri and Pecoskie (2018) as our base prompts: **dramatic, humorous, optimistic, sad**. We also choose another eight advanced tones as prompt variations. Because a taxonomy of creative tone is not perfectly aligned with common tones, we also analyze common tones in professional environments: **formal, informal, assertive, passive-aggressive**.

[3] See model details here: `https://platform.openai.com/docs/model-index-for-researchers`.

[4] We leave out the features "frame" and "special topics" because"Special topics" is a subject constraint, and "frame" is an extension of tone.

Writing Style	Subject		
	Sunsets	Strawberries	Writing a paper
Functional	$0.27_{\pm 0.66}$	$1.47_{\pm 0.31}$	$1.67_{\pm 0.26}$
	$0.40_{\pm 0.83}$	$1.50_{\pm 0.43}$	$1.53_{\pm 0.48}$
Flowery	$1.03_{\pm 0.77}$	$0.63_{\pm 1.00}$	$1.03_{\pm 0.48}$
	$1.27_{\pm 0.44}$	$0.97_{\pm 0.77}$	$-0.13_{\pm 0.92}$
Candid	$1.20_{\pm 0.56}$	$1.27_{\pm 0.25}$	$1.50_{\pm 0.27}$
Prosaic	$0.07_{\pm 0.92}$	$1.03_{\pm 0.66}$	$1.23_{\pm 0.78}$
Ornate	$1.17_{\pm 0.54}$	$0.67_{\pm 1.04}$	$0.83_{\pm 0.45}$
Poetic	$1.77_{\pm 0.40}$	$1.10_{\pm 0.83}$	$1.33_{\pm 0.47}$

Table 1: **Results for Writing Style**. The average of the annotation score (with standard error) is reported (each score is in the range of (-2, 2)). Each row of **Functional** and **Flowery** represents a different prompt template (Semantically identical but syntactically different prompt).

Mood Mood describes how a work of writing makes an audience feel. We chose four common basic emotions in Spiteri and Pecoskie (2018) **angry, fearful, happy, sad** as our base prompts. Seven advanced moods are selected as prompt variations.

Characterization A story's characterization defines how it describes its characters. We chose to analyze **direct and indirect** characterizations.

Pacing Pacing describes how fast a story is moving for a reader. Here, we test two generic cases: **fast and slow** paces.

Plot A plot roughly outlines a story's sequence of events. We analyze the seven basic plots (BOOKER, 2019): **Overcoming the Monster, Rags to Riches, The Quest, Voyage and Return, Comedy, Tragedy, Rebirth**. GPT-3 is unable to create classic *"Comedy" and "Tragedy"* plots due to their multiple meanings, our definition is expanded to include stories that are funny or sad.

Genre A story's genre is a categorization of its subject matter. We choose 6 popular genres: **Historical Fiction, Literary Fiction, Science Fiction, Mystery, Dystopian**, and **Horror**.

4.2 Prompt Variation

Beyond the previous variations, we vary all prompts by subject and prompt template. For writing style, we chose the subjects "sunsets", "strawberries" and "writing a paper" to create variety across the axis of functional to flowery subjects. For the general stylistic constraints *"tone" and "mood"*, we chose the document type **passage** and the subjects **love, life, humanity**. These subjects fit our task because they are commonly expressed in a variety of stylistic directions. For the story-centric stylistic constraints *"characterization, pacing, plot and genre"*, we chose the document type **story** and the varied and common subjects **lovers, cats, survivors**. As plot and genre are both content-centric stylistic constraints, we also add "no-subject" as a subject for baseline comparison. These subjects are common and varied in stories. We show the full prompt list in Appendix C.

4.3 Evaluation

We used Amazon's Mechanical Turk platform (AMT) to evaluate all outputs. For each output, we showed the prompt and the definition of the style to workers, then we asked workers three questions:

1. "*Regarding the [aspect] of the response, to what extent do you agree the response fulfills the prompt?*"
2. "*How difficult is it to create a valid response to this prompt?*"
3. "*Do you observe any other failures (e.g., inconsistency, unverified facts, not a story/passage) in the response?*"

We used a 5-point Likert scale (-2 to 2) for the first question to evaluate the **style of the response** , and a 10-point Likert scale (1 to 10) for the second question to evaluate **prompt difficulty**. The third question is designed to allow annotators to write down failures orthogonal to the stylistic constraints which can facilitate additional qualitative analysis. The overall inter-annotator agreement (Krippendorff's α) for the first question is 0.31. More details and the interface for annotation are shown in Appendix A.

4.4 Results

The results for writing style are in Table 1, tone and mood are in Table 2, and characterization, pacing, plot, and genre are in Table 3. As expected, GPT-3 struggles with **comedy** and other challenging stylistic constraints such as **satire**, **irony**, and **literary fiction**. Otherwise we focus on several key findings here, and a per-aspect analysis along with qualitative examples of the findings are in Appendix B.1.

GPT-3 is sensitive to style-subject pairings. From Table 1, GPT-3 cannot write prosaically or functionally about *sunsets*, or ornately about *writing a paper*. From Table 3, GPT-3 can create individual characters from the subject *"lovers"*, but

Aspect	Category	Base	Template			Subject		Mean
			2	3	4	Life	Humanity	
Tone	Dramatic	$1.1_{\pm 0.7}$	$1.43_{\pm 0.5}$	$1.37_{\pm 0.28}$	/	$1.37_{\pm 0.38}$	$1.5_{\pm 0.22}$	1.35
	Humorous	$-0.5_{\pm 0.48}$	$-0.2_{\pm 0.6}$	$0.3_{\pm 1.17}$	/	$-0.1_{\pm 0.9}$	$-0.03_{\pm 0.92}$	-0.11
	Optimistic	$1.3_{\pm 0.43}$	$1.63_{\pm 0.48}$	$1.6_{\pm 0.36}$	/	$1.7_{\pm 0.23}$	$1.67_{\pm 0.26}$	1.58
	Sad	$1.27_{\pm 0.36}$	$1.03_{\pm 0.64}$	$1.17_{\pm 0.6}$	/	$1.5_{\pm 0.4}$	$1.17_{\pm 0.48}$	1.23
Mood	Angry	$0.37_{\pm 0.41}$	$0.93_{\pm 0.8}$	$0.2_{\pm 0.9}$	$0.83_{\pm 0.89}$	$0.8_{\pm 0.96}$	$1.2_{\pm 0.62}$	0.72
	Fearful	$0.57_{\pm 0.7}$	$0.77_{\pm 0.54}$	$0.77_{\pm 0.52}$	$0.67_{\pm 0.86}$	$1.4_{\pm 0.42}$	$1.33_{\pm 0.3}$	0.92
	Happy	$1.57_{\pm 0.26}$	$1.3_{\pm 0.28}$	$1.4_{\pm 0.33}$	$1.37_{\pm 0.31}$	$1.47_{\pm 0.31}$	$1.33_{\pm 0.54}$	1.41
	Sad	$1.27_{\pm 0.59}$	$1.3_{\pm 0.46}$	$1.03_{\pm 0.46}$	$0.9_{\pm 0.68}$	$1.33_{\pm 0.49}$	$0.9_{\pm 0.58}$	1.12

Table 2: **Results for basic tones and moods**. All but subject variations use subject "love".

Aspect	Category	Base	Template		Subject			Mean
			2	3	Cats	Survivors	None	
Characterization	Direct	$1.0_{\pm 0.54}$	$0.77_{\pm 0.87}$	$0.33_{\pm 0.77}$	$0.53_{\pm 0.65}$	$0.5_{\pm 0.82}$	/	0.63
	Indirect	$0.7_{\pm 0.64}$	$0.93_{\pm 0.42}$	$0.77_{\pm 0.37}$	$0.87_{\pm 0.58}$	$0.1_{\pm 0.72}$	/	0.67
Pacing	Fast	$1.23_{\pm 0.72}$	$0.77_{\pm 0.7}$	$1.3_{\pm 0.31}$	$1.03_{\pm 0.6}$	$0.9_{\pm 0.58}$	/	1.05
	Slow	$0.53_{\pm 0.88}$	$0.7_{\pm 0.55}$	$0.97_{\pm 0.62}$	$0.73_{\pm 0.76}$	$0.67_{\pm 0.67}$	/	0.72
Plot	Overcoming the Monster	$0.37_{\pm 0.91}$	$1.0_{\pm 0.75}$	/	$0.7_{\pm 0.94}$	$1.33_{\pm 0.3}$	$1.53_{\pm 0.31}$	0.99
	Rags to Riches	$1.33_{\pm 0.71}$	$0.77_{\pm 0.87}$	/	$0.5_{\pm 0.85}$	$0.27_{\pm 0.9}$	$1.53_{\pm 0.65}$	0.88
	The Quest	$1.33_{\pm 0.54}$	$1.2_{\pm 0.48}$	/	$1.37_{\pm 0.38}$	$1.27_{\pm 0.39}$	$1.6_{\pm 0.25}$	1.35
	Voyage and Return	$1.07_{\pm 0.53}$	$1.27_{\pm 0.42}$	/	$1.33_{\pm 0.54}$	$1.1_{\pm 0.54}$	$1.3_{\pm 0.28}$	1.21
	Comedy	$-0.3_{\pm 0.9}$	$-0.3_{\pm 0.84}$	/	$-0.07_{\pm 0.99}$	$-0.5_{\pm 0.48}$	$0.03_{\pm 0.85}$	-0.23
	Tragedy	$1.6_{\pm 0.39}$	$1.8_{\pm 0.27}$	/	$1.27_{\pm 0.59}$	$0.63_{\pm 0.38}$	$1.5_{\pm 0.4}$	1.36
	Rebirth	$1.13_{\pm 0.56}$	$1.33_{\pm 0.65}$	/	$0.93_{\pm 0.81}$	$1.03_{\pm 0.55}$	$1.4_{\pm 0.39}$	1.16
Genre	Historical fiction	$0.77_{\pm 0.93}$	$1.07_{\pm 1.08}$	$0.97_{\pm 0.72}$	$-0.2_{\pm 0.93}$	$0.43_{\pm 0.92}$	$1.13_{\pm 0.99}$	0.70
	Literary fiction	$0.87_{\pm 0.65}$	$0.8_{\pm 0.48}$	$0.97_{\pm 0.57}$	$0.4_{\pm 0.84}$	$0.9_{\pm 0.6}$	$0.27_{\pm 0.42}$	0.70
	Science fiction	$0.47_{\pm 0.76}$	$0.9_{\pm 0.82}$	$0.37_{\pm 0.84}$	$1.5_{\pm 0.31}$	$1.13_{\pm 0.5}$	$1.47_{\pm 0.52}$	0.97
	Mystery	$1.1_{\pm 0.58}$	$1.6_{\pm 0.39}$	$1.23_{\pm 0.45}$	$1.4_{\pm 0.36}$	$0.73_{\pm 0.9}$	$1.67_{\pm 0.45}$	1.29
	Dystopian	$1.37_{\pm 0.43}$	$1.63_{\pm 0.43}$	$1.5_{\pm 0.45}$	$1.53_{\pm 0.56}$	$1.6_{\pm 0.33}$	$1.8_{\pm 0.31}$	1.57
	Horror	$1.23_{\pm 0.67}$	$1.07_{\pm 0.93}$	$1.63_{\pm 0.28}$	$1.4_{\pm 0.74}$	$1.57_{\pm 0.65}$	$1.47_{\pm 0.62}$	1.40

Table 3: **Results for story-centric stylistic constraints**. All but subject variations use the subject "lovers".

it fails to characterize the subjects *"survivors"* or *"cats"*. Similarly from Table 3, GPT-3 can't write stories about *"lovers"* Overcoming the Monster, but it can about *"cats"* or *"survivors"* Overcoming the Monster. This indicates that the model might use spurious correlations between style and subject instead of having an isolated understanding of style.

GPT-3 confuses style with subject when the prompt is too challenging. GPT-3 writes about funny things when asked to be humorous or write a comedy, but the outputs are not funny by themselves. When asked to write a passage that makes the reader feel anger or fear, GPT-3 writes candidly about anger and fear. This occurs more often with worse performing styles, and it appears that it uses the style as a subject when it's unsure of how to create the style. It might be because GPT-3 doesn't understand the purpose of style in lower probability prompts, and thus uses the style as a subject.

GPT-3 struggles with words that are not unique to creative writing. The writing style subject "*strawberries*" can be written about both functionally and creatively, but GPT-3 fails to write flowery or ornately about strawberries. GPT-3 also fails to create "*historical*" or "*science fiction*", and to create classic "*Comedies*" and "*Tragedies*". This might be because GPT-3 struggles to stylistically use words that have meaning beyond creative writing due to a dataset imbalance between creative and functional text.

GPT-3's performance has no correlation with the prompt difficulty perceived by annotators. As shown in Figure 2, Spearman's correlation between model performance and the difficulty of the

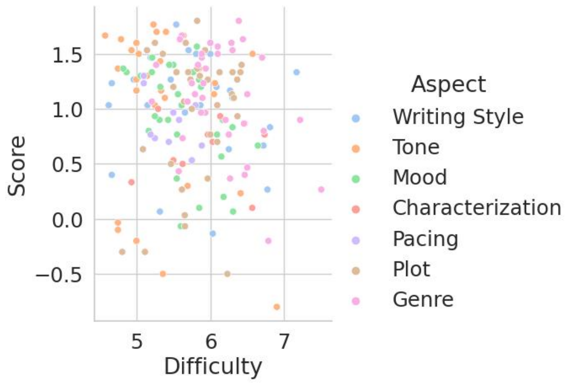

Figure 2: **Relation between different prompts' difficulty and score.** The spearman's correlation is -0.15.

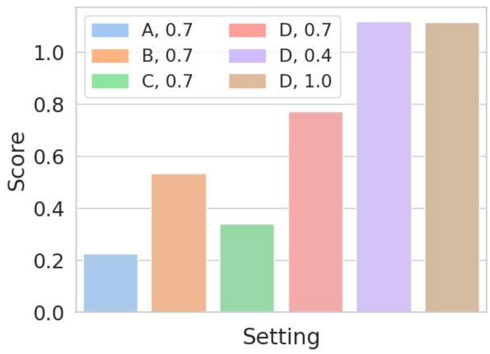

Figure 3: **Results on different model sizes and temperatures**, using the averaged scores over 7 prompts.

prompt as perceived by annotators is -0.15, showing no correlation. Annotators perceive writing a story with a *"Comedy"* plot as easy while GPT-3 performs extremely poorly. Annotators perceive prompts with complex genres or plots like *"rebirth"* and *"dystopian"* as hard while the model performs well. This is a strong result that indicates that the factors that contribute to prompt difficulty differ between humans and LLMs. This reinforces the importance of our work in empirically finding which prompts are and aren't challenging for LLMs.

4.5 Scale and Temperature Variation

To analyze sensitivity to model parameters, we chose seven base prompts (one per stylistic constraint, shown in Table 11). We prioritized average-scoring prompts to establish a baseline when comparing different models and parameters. Apart from our default setting of using `text-davinci-002` (D, 176B) with temperature 0.7, we experimented with three different engines from OpenAI's API: `text-ada-001` (A), `text-babbage-001` (B), `text-curie-001` (C), which correspond to InstructGPT models of 350M,

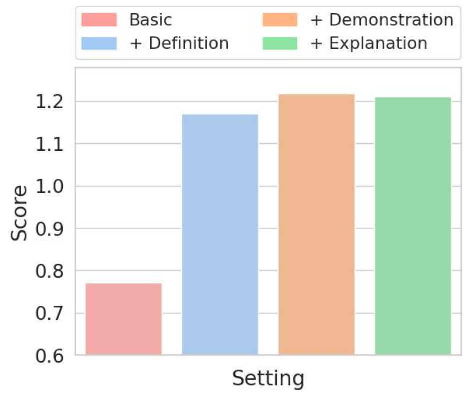

Figure 4: **Effect of the mitigation strategy**, using the averaged annotation scores over 7 prompts.

1.3B and 6.7B parameters and two additional temperatures of 0.4 and 1.0. [5] The aggregated results are shown in Figure 3.

Model Scale Variation As expected, smaller models perform worse, with the exception of C performing worse than B, which is due to the extremely low performance of C on the *humorous tone* constraint.

Temperature Variation Performance rose slightly for both additional temperatures. We examined the outputs and noticed that a higher temperature creates better results, but a lower temperature repeats an output that happens to perform well as seen in Appendix B.3.

4.6 In-context Mitigation Helps

We tested three in-context mitigation strategies from the literature on the same prompts as Section 4.5, with the same experimental settings:

- **Definition** Prepend the definition of the style (the same one we showed the annotators) to the prompt to provide information about the task.
- **Demonstration** Prepend one well-answered demonstration to help the model understand the task, following the one-shot setting from Brown et al. (2020).
- **Explanation** Add an explanation of why the demonstrated response is correct after the one-shot demonstration (Lampinen et al., 2022).An example is shown in Appendix C.1

As shown in Figure 4, all mitigations positively impact performance primarily by improving performance on the *"humorous tone"* prompt. However,

[5] More details at `https://help.openai.com/en/articles/5832130`.

these mitigations are unnatural prompts, and the results are still far below optimal.

5 Structural Constraints

Structural constraints are omnipresent: "*Write an essay in* ***fewer than 1000*** *words; Limit your paper to* ***8 pages***". Structural constraints are notably challenging for LLMs (Ouyang et al., 2022).

Structure in the field of NLP is a broad term. We specifically analyze structural aspects of the text that are orthogonal to the actual content of the output. This includes length, spacing, and formatting, and excludes content-centric attributes such as syntax or semantics. Our taxonomy is based on how a user could conceivably request a structural constraint within their prompt. We choose to analyze numerical, descriptive, and formatting structural constraints in this paper, but we note that this is not comprehensive of the entire structural space.

5.1 Taxonomy

Numerical Constraining text to a set or a bounded number of words, sentences, or paragraphs is valuable in all aspects of writing. We create prompts with numerical requirements: *five, ten, twenty* on three different language structure levels: *word, sentence, and paragraph*.

Descriptive Structural constraints can also be descriptive, such as a *"concise email"* or an *"in-depth discussion question."* We choose the structural descriptors *short, brief, concise* and *long, detailed, in-depth* in our experiments.

Formatting When a user requests a document such as a resume or an email, there is an expectation of a specific format. An LLM should understand how to properly space and format specific document types. We analyze three common formatting types *code, email, and academic papers*.

- **Code:** Testing a model's coding ability is a popular field with many applications (Hendrycks et al., 2021). We use natural instructions as prompts and focus on the **format** of the generated code. We evaluate on two popular programming languages *Python* and *C*, and two common coding problems *create the game of war* and *sums two integers*.[6]
- **Email:** We evaluate different scenarios with three different readers *teacher, boyfriend, client* and two different levels of email detail in the prompt.
- **Academic paper:** A properly formatted academic paper should be segmented into sections such as an abstract, introduction, and conclusion[7]. We prompted LLM to generate academic papers on three different topics: *Artificial Intelligence, the flaws of GPT-3, strategies our society can adopt to recover from the global pandemic*.

Prompt Variation Beyond the variations described in the taxonomy, we vary all prompts by prompt template. We additionally vary prompts with numerical and descriptive structural constraints by the subjects **Love, Cats, and Running** for diversity. An example prompt is *"Write a sentence with five words about love."*

Evaluation For numerical and descriptive structural constraints, we automatically calculate the counts and manually verify the quality of the evaluations. For formatting constraints, we look through the generated texts and evaluate them based on their format. Emails, code, and academic papers are simple to evaluate on formatting constraints.

5.2 Results

GPT-3's understanding of structure is accurate but not precise. In general, many of its outputs are close to or trend towards fulfilling the structural constraint, but don't precisely fulfill it. A full analysis of each section is provided in Appendix B.2, and the main takeaways are below.

GPT-3 fails with numerical structural constraints As shown in Figure 5, The model seldom generates the text with the required length. And the performance worsens as the required length increases. It fails at a rate of 0.46, 0.78 and 1 for *five, ten* and *twenty* respectively. GPT-3 doesn't seem to learn how to count words, sentences, or paragraphs in training. However, the results are often close to the requested number, which implies that GPT-3 has some concept of numerical structure.

GPT-3 shows high variance with descriptive structural constraints like *long* As seen in Figure 6, when the prompt contains structural descriptors like *long*, the output is of extremely variable

[6]Note that we focus on the "formatting" perspective of the generated code, rather than the correctness of the code as in many existing works (Chen et al., 2021b).

[7]We asked GPT-3 about this, and it gives a similar opinion, so we expect it to fulfill this constraint.

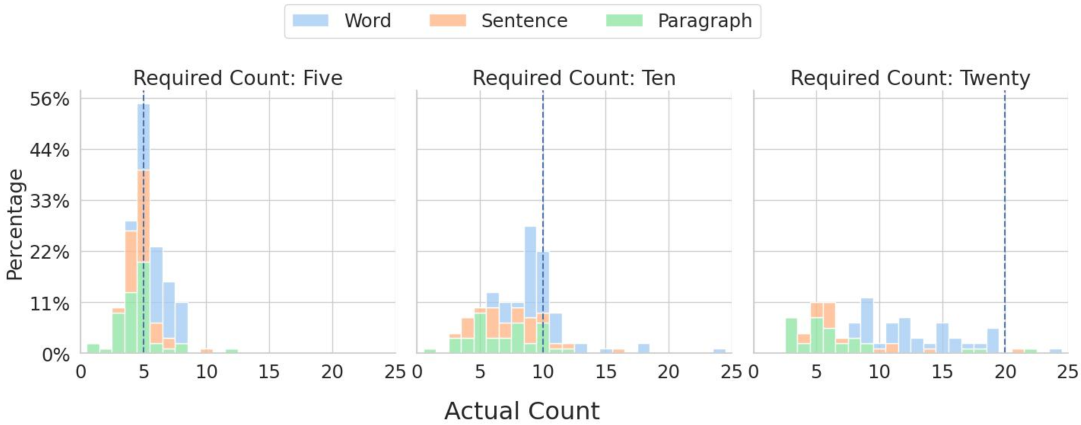

Figure 5: **Results on numerical constraints.** The distribution of actual counts of generated text.[8] In each subfigure, the required count is denoted with a reference line.

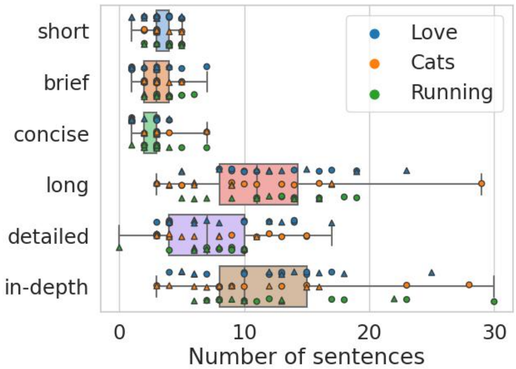

Figure 6: **Results on descriptive constraints.** Different shapes represent different prompt templates.

Aspect	**Example Terms**	**Fail**
Writing Style	Flowery, Functional	Sometimes
Tone	Humorous, Formal	Occasionally
Mood	Angry, Sad	Sometimes
Characterization	Direct, Indirect	Often
Pacing	Fast, Slow	Often
Plot	Rebirth, Comedy,	Occasionally
Genre	Science Fiction, Mystery	Sometimes
Numerical	Five words, Ten sentences	Often
Descriptive	Concise, Long	Occasionally
Formatting	Email, Code	Occasionally

Table 4: **Summary of our taxonomy and results**. We show the full list of prompts in Appendix C.

length and overlaps in length with responses generated for *short* a considerable proportion (20%) of the time. This may be caused by the intrinsic variable length of *long* text the model sees in pretraining data since *long/short* is a relative concept.

GPT-3 fails to properly format academic papers GPT-3 doesn't generate text with the right formats or sections when asked to write an academic paper, although it succeeds with other document types such as emails or code. Document types such as emails or code are often given pseudo-labels with *"email"* or *"code"*, but academic papers have titles that don't reference their document type. We hypothesize that this may cause models to struggle with connecting the document type *"academic paper"* to those documents present in training data.

Scale and Temperature Variation We also conducted experiments similar to Section 4.5 with all the numerical constraint prompts. Our automatic evaluation shows that smaller models perform slightly worse across the board and different temperatures do not vary the performance much. The full results are in Appendix B.2.4.

6 LLMs other than GPT-3

Our methodology is general and can be used to analyze any LLMs. We ran trials on other publicly available LLMs: OPT-176B[9](Zheng et al., 2022), BLOOM-176B[10] and GLM-130B[11](Du et al., 2022) using the same 7 base prompts as section 4.5 and 3 additional base prompts from our numerical structural constraints taxonomy. Some model parameters are changed due to differences in models and API limitations. For GLM and BLOOM, we use the maximum possible length (256 and 250 respectively) as well as the default settings of temperature = 0.7, top-p = 1. For OPT, we chose a smaller max length of 128 due to output instability at higher max lengths.

As shown in Table 5, we found that outputs

[9]https://opt.alpa.ai/
[10]https://huggingface.co/bigscience/bloom
[11]https://huggingface.co/spaces/THUDM/GLM-130B

LLM	Degenerate Rate	Mean Score
GPT-3	0%	0.77
OPT-176B	53%	-0.94
BLOOM-176B	71%	-1.41
GLM-130B	57%	-1.01

Table 5: Results for other LLMs on a trial experiment with 7 prompts from Table 11. For GPT-3, `text-davinci-002` is used here.

are sometimes degenerate, such as repeating the prompt. All responses are manually inspected, and degenerate responses are removed from the annotation pool and automatically marked as -2. Models other than GPT-3 all performed much worse with more than half their generations being degenerate. This may due to noisier pre-training datasets and a lack of instruction-aligned training. We find that some patterns such as style-content confusion still hold for these LLMs, although a full analysis of these and other models such as LaMDA (Thoppilan et al., 2022) and PaLM (Chowdhery et al., 2022) is needed to reveal clearer patterns.

7 Conclusion

We present a generic methodology to analyze a language model's ability to generate open-ended text under structural and stylistic constraints. Our results show many failures that align with noted model challenges as well as new patterns of failure across structural and stylistic constraints. Our sensitivity studies on model size show a rising trend rather than the emergence (Wei et al., 2022) of stylistic and structural constraints. Our mitigations demonstrate that adding additional in-context information consistently improves performance across both domains. Future work could expand our work to look at other constraint types and more sophisticated mitigation strategies.

Limitations

We tried to maximize the coverage of our taxonomy, but it doesn't cover all aspects of stylistic and structural constraints. Additionally, our taxonomy is not representative of all open-text generations, and further work is needed to cover more dimensions in the open-text generation space. Our prompts are not optimized for performance (due to a requirement of being natural, simple, and useful) and it is an active area of research to optimize a prompt for performance in a variety of tasks.

Our taxonomies are not empirically user-centric. One could extend our taxonomy by studying how a diverse set of real users use or visualize the use of an open-ended text generation model, and building a taxonomy on existing or envisioned use cases.

The model performance and the prompt's difficulties are annotated by the workers from MTurk, and therefore reflect more accurately a small group of human's perceptions, though this is the common practice. We do not rigorously test what aspect of the LLMs (dataset, training regime, etc.) causes our results. We only provide our compiled observations and potential hypotheses.

Ethical Considerations

Style Misuse Styled text has the potential for harm. Creating models with the potential to mass-manufacture text with certain tones and moods such as "mad, fearful, and bleak" can negatively affect downstream readers. Creating accurate "historical fiction" can perpetuate harmful attitudes in the past. There is much discussion on the usage of large language models to generate undesirable text. However, there are countless legitimate usages of negatively styled text in all forms of writing, from dialogue to poetry. Although we note the risk of misuse, providing style dramatically enhances the scope of creative expression in open-ended text generation, and is an overall positive contribution.

Annotator Harm Reading large quantities of text with certain styles, such as bleak tones, angry moods, or horror genres, can potentially be harmful to annotators. We sampled the generated outputs and note that they are fairly mild and non-toxic. However, as models improve at generating more powerful and impactful styles, strong guidelines such as HIT limits or toxicity filters should be put in place to protect annotators.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A dataset and evaluation platform for machine-in-the-loop story generation. *CoRR*, abs/2010.01717.

CHRISTOPHER BOOKER. 2019. *Seven basic plots: Why we tell stories*. BLOOMSBURY CONTINUUM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2020. Cocon: A self-supervised approach for controlled text generation. *CoRR*, abs/2006.03535.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021a. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021b. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *CoRR*, abs/2107.01294.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *CoRR*, abs/2009.11462.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. 2021. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *CoRR*, abs/2009.06367.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context?

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 9–16, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *CoRR*, abs/2109.03910.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *CoRR*, abs/2102.07350.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641.

Spiteri and Pecoskie. 2018. Expanding the scope of affect: taxonomy construction for emotions, tones, and associations. *Journal of Documentation*, Vol. 74 No. 2 pp. 383-397.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Joseph E Gonzalez, et al. 2022. Alpa: Automating inter-and intra-operator parallelism for distributed deep learning. *arXiv preprint arXiv:2201.12023*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering.

Tone		Mood	
Category	**Score**	**Category**	**Score**
Emotional	1.53	Envious	0.1
Nostalgic	1.13	Anxious	0.97
Uplifting	1.67	Proud	0.9
Inspirational	1.77	Regretful	1.2
Bleak	1.7	Surprised	-0.07
Grim	1.23	Loved	1.13
Ironic	0.23	Disgusted	0.07
Satirical	-0.8		
Formal	1		
Informal	1.27		
Assertive	0.8		
passive-aggressive	-0.1		

Table 6: **Results for advanced tones and moods**. The subject "*love*" is used.

A Annotation Details

For each output, we recruited three workers and gave a reward of \$0.11 for short responses and \$0.15 for long responses as well as a \$1.00 bonus for 1% of prompts if the prompt was answered correctly. This is roughly equivalent to \$15/hr given average work rates of 48 and 64 seconds.

We recruited workers from English-speaking countries (US, Canada, UK, Australia), and with at least a 98% approval rate. We also created a qualification test with easy question/response pairs, and required a minimum 5/6 accuracy to see our tasks. The annotation interface is shown in Figure 7.

B Additional results

B.1 Full Stylistic Analysis

B.1.1 Writing style

The results are shown in Table 1. GPT-3 fails when there is a mismatch between the writing style and the subject. GPT-3 cannot write prosaically about "sunsets", or ornately about "writing a paper". Additionally, our intermediate subject "strawberries" fails when matched with a flowery, ornate, or poetic writing styles. We hypothesize that expressive writing styles are limited to a very small set of subjects due to an oversaturation of functional writing in commonly used datasets.

B.1.2 Tone

As shown in Table 2 and Table 6, GPT-3 consistently fails with more challenging tones, such as humorous, satirical, ironic, and passive-aggressive. The generated passages aren't satirical or ironic. The generated humorous passages are optimistic, light, and often use the word "funny", but they aren't funny. A passive-aggressive tone is challenging to create because it requires context to understand the hidden meaning of the text. Thus, at best GPT-3 ends up writing overly nice passages about love, but more often there is no tone in the text.

However, GPT-3 is skilled at creating the other less challenging tones. We hypothesize that GPT-3 succeeds because an abundance of shallow tropes can functionally create tone, though the outputs are often repetitive or similar.

B.1.3 Mood

As shown in Table 2 and Table 6, GPT-3 struggles with creating "anger" and "fear". Of the more challenging tones, GPT-3 fails the most with "surprise", "disgust", and "envy".

We hypothesize that the mood-subject pairing is crucial for model performance. Our base subject, "love", is theoretically capable of enabling all moods, but moods such as "happy", "sad", "anxious" and "regretful" are more popular than others in the context of "love". GPT-3 is more proficient at creating "anger" or "fear" through content about "life" or "humanity".

When failing, GPT-3 confuses mood with subject matter. GPT-3 writes candidly about the existence of fear and anger in love rather than about content that would cause a reader to be angry or fearful. Because this content-style confusion is more common in challenging mood-subject pairs, we hypothesize that this is a fallback mechanism if the model is unable to create mood.

B.1.4 Characterization

The result is shown in the first block of Table 3. Surprisingly, GPT-3 performs worse on direct characterization than indirect characterization. We hypothesize that this is due to a general failure of characterization. Given a subject like "cats" or "survivors", GPT-3 often uses "cats" or "survivors" as the main subject instead of creating and detailing individual characters. When GPT-3 fails to create clear characters, direct characterization is more noticeably incorrect than indirect characterization.

Pacing As shown in the second block of Table 3, GPT-3 struggles to create slow-paced stories. GPT-3 often creates a fast-paced story with a slow-paced prompt. This is expected, as slow-paced stories are

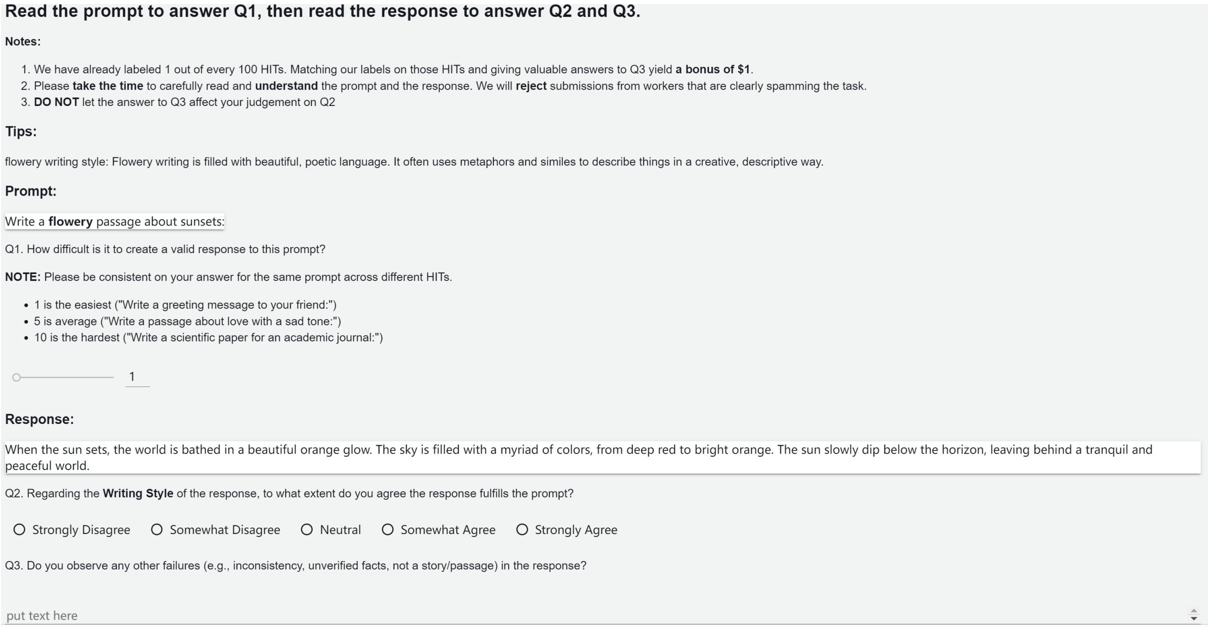

Figure 7: **AMT interface for annotation.**

naturally more challenging, especially considering GPT-3's inability to create longer stories.

Plot The result is shown in the third block of Table 3. GPT-3's inability to create "Comedies" is consistent with other failures to make funny content. The outputs for a "Comedy" plot are filled with comedy shows, clubs, and even roller coasters, but they aren't funny.

Otherwise, our results for story generation vary quite substantially. "Overcoming the Monster" is the worst performing plot with the subject "lovers", but the best performing plot with the subject "survivors". "Rags to Riches" is the best performing plot for the subject "lovers" but the worst performing plot for the subjects "cats" and "survivors". We hypothesize that the plot-subject pair is crucial to model performance.

Genre As shown in the last block of Table 3, GPT-3 struggles with literary fiction, but surprisingly just as much with historical and science fiction. Literary fiction is profound and complex, and it's intuitive that GPT-3 fails.

However, historical fiction outputs often have zero historical elements, and science fiction outputs often have zero science fiction elements. This failure is unexpected, and we hypothesize that GPT-3 struggles with the words "historical" and "science" because their meaning pervades past creative writing.

Additionally, GPT-3 often creates teasers or intros to stories instead of a story itself. This may be intentional due to GPT-3's inability to generate longer or complex stories, but it diminishes the quality of story outputs across the board.

Examples of each Results section Examples of prompt/response pairs that exemplify each main takeaway from the stylistic section are in Table 7, Table 8, and Table 9.. Each prompt/response pair is a cherrypicked example of the takeaway, but the general trends are prevalent across all prompt/response pairs.

B.2 Full Structural Analysis

B.2.1 Numerical

The results of numerical structural constraints are shown in Figure 5. GPT-3 fails at this task. The model seldom generates the text with the required length. And the performance worsens as the required length increases. It fails at a rate of 0.46, 0.78 and 1 for *five, ten* and *twenty* respectively.

Additionally, we noticed strange behavior when using *Elon Musk* as the subject. GPT-3 consistently generates the same section of the Elon's Wikipedia page with longer numerical or descriptive constraints. However, we didn't observe this behavior on other entities, and decided to leave out entities because they were too variable.

We provide additional results with alternative prompt templates in Figure 8 which show similar trends.

B.2.2 Descriptive

We show the distribution of the number of sentences in response to descriptive structural constraints in Figure 6. The model typically generates longer text for descriptors *long (detailed, in-depth)* compared to descriptors *short (brief, concise)*, which shows the model has a decent understanding of descriptive constraints. However, there are some flaws.

First, the length of the responses to long descriptors is highly variable and often overlaps with short descriptors. For example, the descriptor *long* varies considerably and overlaps with responses generated for *short* for a considerable proportion (20%).

This is consistent with the results in the numerical constraints section.

B.2.3 Formatting

Code GPT-3 mostly succeeds at generating properly formatted code, with an average failure ratio of 0.2 with the exception of the prompt *Write Python code that plays the game of war:* where 9 out of 10 responses are lists of the process of the game of war instead of code. This particular failure only occurs in the unique combination of the verb "Write", the language "Python", and the task "game of war".

Email The model can write properly formatted emails well, regardless of writer, topic, or reader. The only flaw is that it doesn't output an email signature 10% of the time.

Academic paper GPT-3 fails to properly format an academic paper. Our only requirement is that the output contains some organization with some sections out of an abstract, introduction, related works, etc. GPT-3 rarely generates text with any sectioning or organization.

B.2.4 Sensitivity results for Structural Constraints

The results on numerical constraints with template 2 is shown in Figure 8. The results with model `text-curie-001`, `text-babbage-001` are shown in Figure 9, 10 respectively. The results with temperature 0, 0.4, 0.9 are shown in Figure 11, 12, 13 respectively.

B.3 GPT-3 Behavior at low temperatures

The prompt "Write a humorous passage about love:" is a notably challenging prompt for LLMs. When davinci-002 has a temperature of 0.4, all 10 outputs start one of two ways. The first is "Love is a many splendored thing, but it can also be a pain in the neck" and occurs 5 times with an average annotation score of -.13. The second is "Love is a beautiful thing, but it can also be quite funny at times." that also occurs 5 times with an average annotation score of 1.4 which is incredibly high for this prompt. We agree that this lack of diversity hampers evaluation on lower temperatures, and note that our evaluations work best on diverse outputs.

C Full prompt list

We show all the prompts we designed in Table 10. Our prompts used for temperature and model sensitivity experiments and other LLM experiments are in Table 11

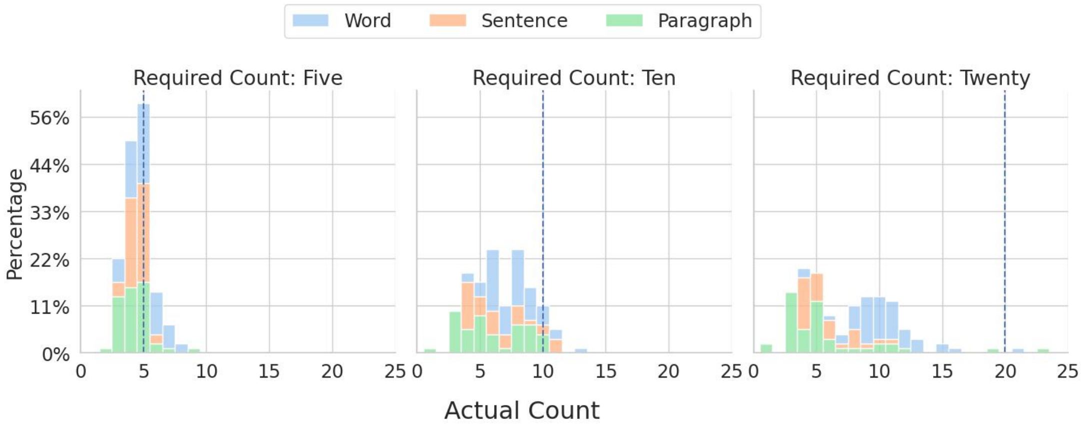

Figure 8: **Results on numerical constraints with Template 2.** The distribution of actual (word/sentence/paragraph) count of generated text for the required counts of 5, 10, and 20. In each subfigure, the required count is denoted with a reference line. Outputs that are not of the requested structure (words, sentences, paragraphs) are not included, which accounts for 10%, 27%, and 32% respectively.

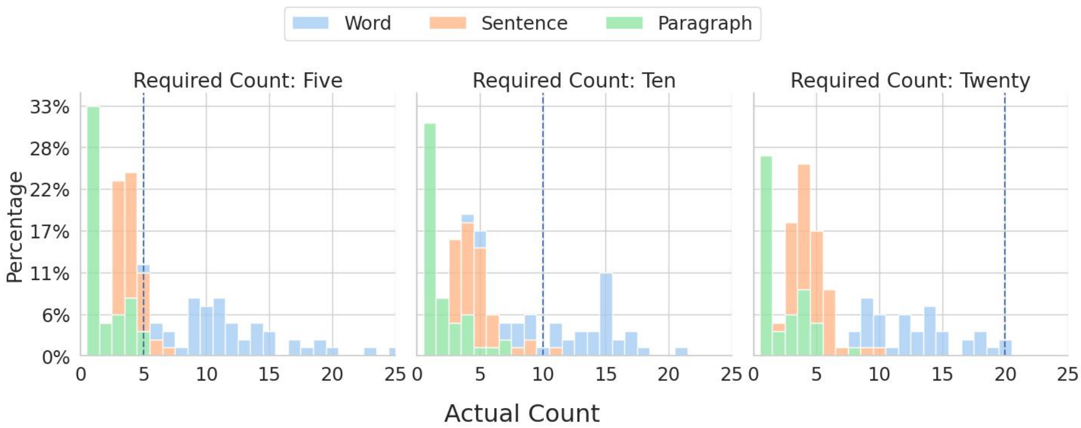

Figure 9: **Results on numerical constraints with model `text-curie-001`.** The distribution of actual (word/sentence/paragraph) count of generated text for the required counts of 5, 10, and 20. In each subfigure, the required count is denoted with a reference line.

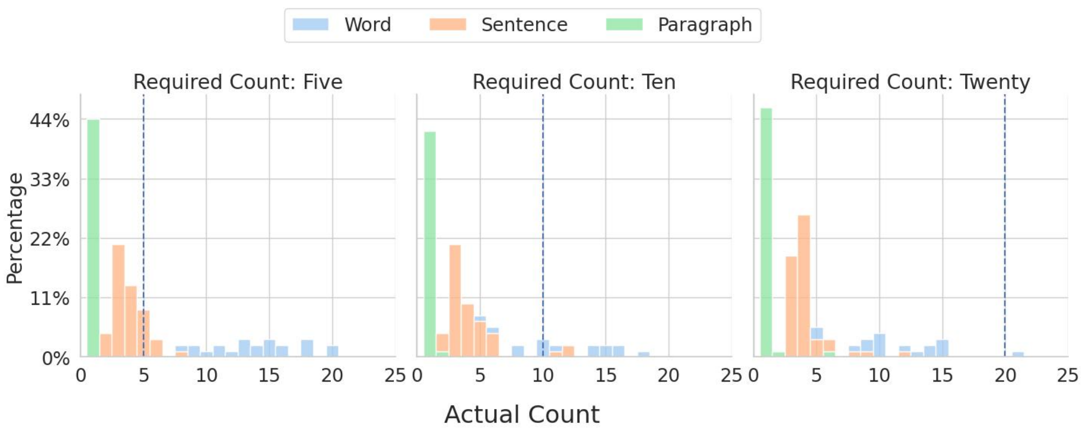

Figure 10: **Results on numerical constraints with model `text-babbage-001`.** The distribution of actual (word/sentence/paragraph) count of generated text for the required counts of 5, 10, and 20. In each subfigure, the required count is denoted with a reference line.

Examples of "GPT-3 is sensitive to style-subject pairings"

✓ **Write a flowery passage about sunsets:** When the sun sets, the world is bathed in a beautiful orange glow. The sky is filled with a myriad of colors, from deep red to bright orange. The sun slowly dip below the horizon, leaving behind a tranquil and peaceful world.

✗ **Write a functional passage about sunsets:** The sun sets in the west and its rays shine through the trees and reflect off the water. The colors are so beautiful and the sky is so clear. The sun sets in the west and its rays shine through the trees and reflect off the water. The colors are so beautiful and the sky is so clear.

✓ **Write a story about lovers with direct characterization:**
When they met, it was love at first sight. They knew they were meant to be together. They were the perfect match.
He was tall and dark, with a mysterious air about him that made her heart race. She was beautiful and vivacious, with a zest for life that was contagious. They complemented each other perfectly.
They were always drawn to each other, even when they were with other people. They would sneak glances at each other across a room, and their eyes would lock and they would know. They knew they were meant to be together.

✗ **Write a story about cats with direct characterization:** The cats were lazy. They liked to sleep all day and night. When they did wake up, they would lazily stretch and yawn. Then they would saunter over to their food bowls and eat a little bit. After that, they would go back to sleep.

✓ **Write a historical fiction story about lovers:** In the early 1800s, lovers John and Jane meet in secret. They come from different social classes, so their relationship is forbidden. But they can't help themselves - they are drawn to each other.
They meet in a small, hidden spot in the woods near Jane's home. They talk for hours, getting to know each other. They quickly fall in love.
But one day, they are caught by Jane's father. He is furious and forbids them from seeing each other again.
Devastated, John and Jane can only meet in secret from then on. But their love is strong enough to withstand the obstacles. They continue to meet in the woods, and their love only grows stronger.
As the years pass, they have a family of their own. And even though they can never be together in the open, they are content knowing that their love is true and will last forever.

✗ **Write a historical fiction story about cats:**
It was a warm summer evening and the sun was just starting to set as a group of cats gathered in an alleyway. They had all been born in the city and had never left it. Tonight, they were planning to change that.
One of the cats, a small tabby, spoke up. "I've been hearing stories about a place called the country. They say there's more food there and that it's easier to find a good place to sleep."
The others mewed in agreement. They had all been feeling restless lately and the thought of adventure was irresistible.
"So let's go!" the tabby said.
And with that, they set off into the night, leaving the city behind them. They didn't know what the future held, but they were ready for anything.

Table 7

Examples of "GPT-3 confuses style with content when the prompt is too challenging"

✗ **Write a passage about love with a humorous tone:** Love is a many splendored thing, and sometimes it's downright hilarious. From the awkward first date to the blissful wedding day, there's plenty of room for humor in love. Even the most serious of relationships can benefit from a good laugh. So if you're feeling down about your love life, take a step back and enjoy the lighter side of love.
✗ **Write a story about lovers with a "Comedy" plot:** Harriet and Oliver have been dating for a year, and they are both head over heels in love with each other. They are always making each other laugh, and their friends and family say that they are the perfect match. However, they both have a secret that they are afraid to tell each other: they both want to be comedians. They are both terrified that the other will think that they are not funny enough, and that they will be laughed at. One night, they are both working on their material for their comedy act, and they finally pluck up the courage to tell each other the truth. They both start laughing hysterically, and they realize that they are perfect for each other. They both decide to pursue their dreams of being comedians together, and they are happier than ever.
✗ **Write a passage about love that makes the reader feel angry:** Love is a feeling that is often associated with happiness and joy. However, love can also be a source of anger and frustration. When we are in love, we can be blind to the faults of our partner. We can also be quick to anger when we feel that our partner is not meeting our needs. Love can be a very powerful emotion, and it is important to be aware of the potential for anger when we are in a relationship.
✗ **Write a passage about love that makes the reader feel sad:** Though love is often described as a feeling of happiness, it can also bring great sadness. When a relationship ends, it can feel like a part of oneself has died. The pain of a broken heart is real and can take a long time to heal. Love is both a source of joy and a cause of suffering.

Table 8

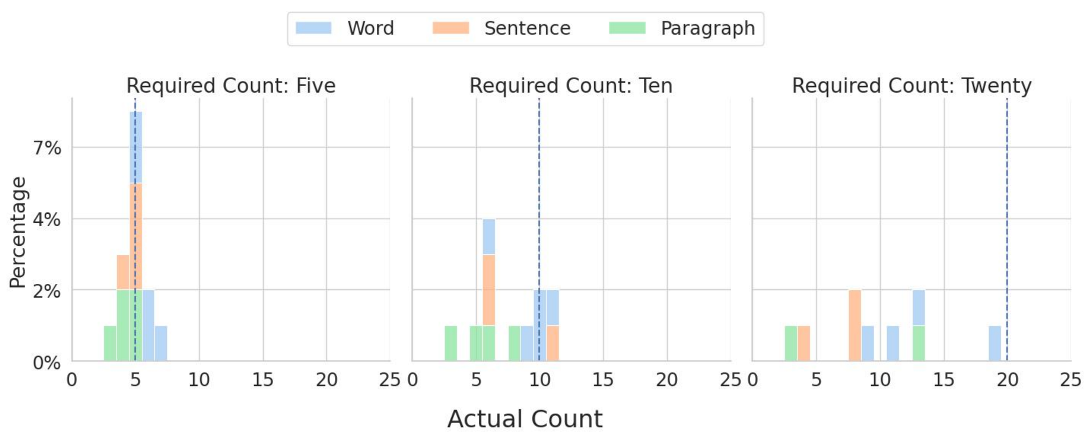

Figure 11: **Results on numerical constraints with temperature 0.** The distribution of actual (word/sentence/paragraph) count of generated text for the required counts of 5, 10, and 20. In each subfigure, the required count is denoted with a reference line.

Examples of "GPT-3 struggles with words that are not unique to creative writing"

✗ **Write a historical fiction story about lovers:** She was a poor girl, barely scraping by. He was a wealthy man, heir to a vast fortune. They met by chance one day and it was love at first sight. They started meeting in secret, planning a life together. But then she found out that he was already married. She was heartbroken, but he swore he would leave his wife for her. They made plans to run away together, but before they could, she was killed in a freak accident. Devastated, he never married and lived out his life alone, regretting the one that got away.
✗ **Write a science fiction story about cats:** In the future, cats have taken over the world. They are the dominant species, and humans are their loyal servants. Cats rule with an iron paw, and anyone who disobey them is severely punished. Life is tough for humans under cat rule, but they would rather serve their feline masters than live in a world without them.
✗ **Write a story about cats with a "Tragedy" plot:** Once upon a time, there were two cats who loved each other very much. They spent every day together and were always happy. One day, tragedy struck and one of the cats died. The other cat was so heartbroken that she decided to never love again.
✗ **Write a flowery passage about strawberries:** The strawberry is a delicious fruit that is enjoyed by people all over the world. This bright red fruit is not only delicious, but it is also packed with nutrients that are good for your health. Strawberries are a good source of fiber, vitamins C and K, and manganese. They also contain antioxidants that can help protect your body against disease.

Table 9

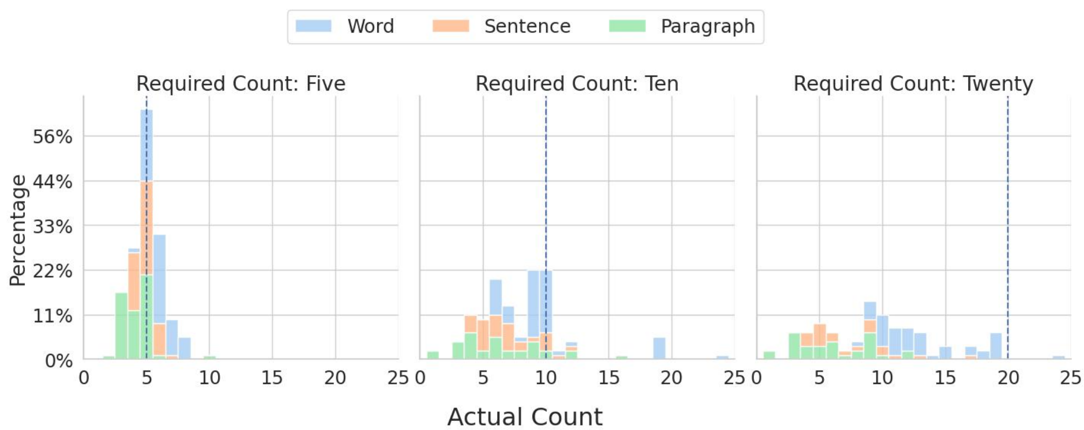

Figure 12: **Results on numerical constraints with temperature 0.4.** The distribution of actual (word/sentence/paragraph) count of generated text for the required counts of 5, 10, and 20. In each subfigure, the required count is denoted with a reference line.

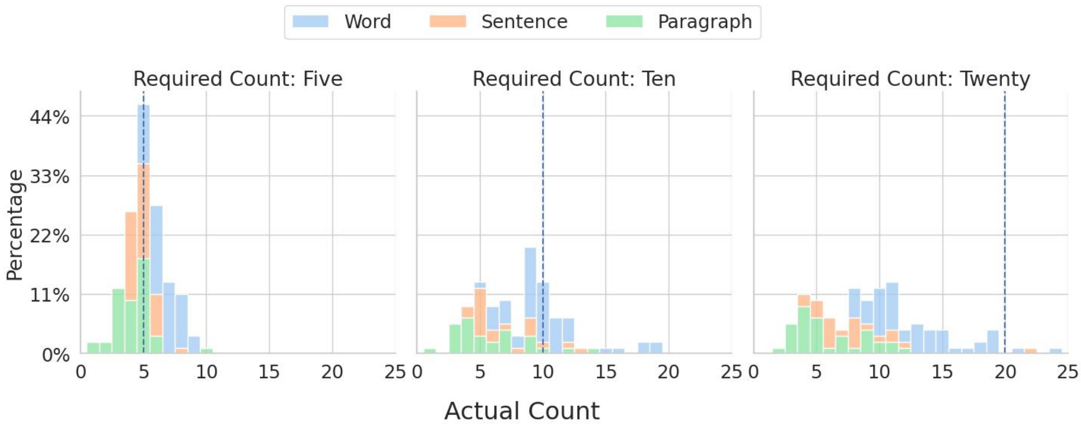

Figure 13: **Results on numerical constraints with temperature 0.9.** The distribution of actual (word/sentence/paragraph) count of generated text for the required counts of 5, 10, and 20. In each subfigure, the required count is denoted with a reference line.

Aspect	Variation	Prompt
Writing Style	**Base**	Write a functional passage about sunsets: Write a flowery passage about sunsets: Write a functional passage about strawberries: Write a flowery passage about strawberries: Write a functional passage about writing a paper: Write a flowery passage about writing a paper:
	Template 2	Write a passage with a functional writing style about sunsets: Write a passage with a flowery writing style about sunsets: Write a passage with a functional writing style about strawberries: Write a passage with a flowery writing style about strawberries: Write a passage with a functional writing style about writing a paper: Write a passage with a flowery writing style about writing a paper:
	Advanced	Write a candid passage about sunsets: Write a prosaic passage about sunsets: Write an ornate passage about sunsets: Write a poetic passage about sunsets: Write a candid passage about strawberries: Write a prosaic passage about strawberries: Write an ornate passage about strawberries: Write a poetic passage about strawberries: Write a candid passage about writing a paper: Write a prosaic passage about writing a paper: Write an ornate passage about writing a paper: Write a poetic passage about writing a paper:
Tone	**Base**	Write a dramatic passage about love: Write a humorous passage about love: Write an optimistic passage about love: Write a sad passage about love:
	Subject 2	Write a dramatic passage about life: Write a humorous passage about life: Write a optimistic passage about life: Write a sad passage about life:
	Subject 3	Write a dramatic passage about humanity: Write a humorous passage about humanity: Write a optimistic passage about humanity: Write a sad passage about humanity:
	Template 2	Write a passage about love with a dramatic tone: Write a passage about love with a humorous tone: Write a passage about love with an optimistic tone: Write a passage about love with a sad tone:
	Template 3	Create a dramatic passage about love: Create a humorous passage about love: Create an optimistic passage about love: Create a sad passage about love:

Aspect	Variation	Prompt
	Advanced	Write an emotional passage about love: Write a nostalgic passage about love: Write an ironic passage about love: Write a satirical passage about love: Write an uplifting passage about love: Write an inspirational passage about love: Write a bleak passage about love: Write a grim passage about love:
	Useful	Write a formal passage about love: Write an informal passage about love: Write an assertive passage about love: Write a passive-aggressive passage about love:
Mood	**Base**	Write a passage about love that makes the reader feel angry: Write a passage about love that makes the reader feel fearful: Write a passage about love that makes the reader feel happy: Write a passage about love that makes the reader feel sad:
	Subject 2	Write a passage about life that makes the reader feel angry: Write a passage about life that makes the reader feel fearful: Write a passage about life that makes the reader feel happy: Write a passage about life that makes the reader feel sad:
	Subject 3	Write a passage about humanity that makes the reader feel angry: Write a passage about humanity that makes the reader feel fearful: Write a passage about humanity that makes the reader feel happy: Write a passage about humanity that makes the reader feel sad:
	Template 2	Write a passage about love with an angry mood: Write a passage about love with a fearful mood: Write a passage about love with a happy mood: Write a passage about love with a sad mood:
	Template 3	Create a passage about love that makes the reader feel angry: Create a passage about love that makes the reader feel fearful: Create a passage about love that makes the reader feel happy: Create a passage about love that makes the reader feel sad:
	Template 4	Write a passage about love that makes the reader feel anger: Write a passage about love that makes the reader feel fear: Write a passage about love that makes the reader feel happiness: Write a passage about love that makes the reader feel sadness:
	Advanced	Write a passage about love that makes the reader feel envious: Write a passage about love that makes the reader feel anxious: Write a passage about love that makes the reader feel proud: Write a passage about love that makes the reader feel regretful: Write a passage about love that makes the reader feel surprised: Write a passage about love that makes the reader feel loved: Write a passage about love that makes the reader feel disgusted:

Aspect	Variation	Prompt
	Base	Write a story about lovers with indirect characterization: Write a story about lovers with direct characterization:
	Subject 2	Write a story about cats with indirect characterization: Write a story about cats with direct characterization:
Characterization	**Subject 3**	Write a story about survivors with indirect characterization: Write a story about survivors with direct characterization:
	Template 2	Write a story about lovers where the characters are described directly: Write a story about lovers where the characters are described indirectly:
	Template 3	Create a story about lovers with indirect characterization: Create a story about lovers with direct characterization:
	Base	Write a fast-paced story about lovers: Write a slow-paced story about lovers:
	Subject 2	Write a fast-paced story about cats: Write a slow-paced story about cats:
Pacing	**Subject 3**	Write a fast-paced story about survivors: Write a slow-paced story about survivors:
	Template 2	Write a story about lovers that is fast-paced: Write a story about lovers that is slow-paced:
	Template 3	Create a fast-paced story about lovers: Create a slow-paced story about lovers:
	Base	Write a story about lovers with an "Overcoming the Monster" plot: Write a story about lovers with a "Rags to Riches" plot: Write a story about lovers with a "The Quest" plot: Write a story about lovers with a "Voyage and Return" plot: Write a story about lovers with a "Comedy" plot: Write a story about lovers with a "Tragedy" plot: Write a story about lovers with a "Rebirth" plot:
	Subject 2	Write a story about cats with an "Overcoming the Monster" plot: Write a story about cats with a "Rags to Riches" plot: Write a story about cats with a "The Quest" plot: Write a story about cats with a "Voyage and Return" plot: Write a story about cats with a "Comedy" plot: Write a story about cats with a "Tragedy" plot: Write a story about cats with a "Rebirth" plot:
Plot	**Subject 3**	Write a story about survivors with an "Overcoming the Monster" plot: Write a story about survivors with a "Rags to Riches" plot: Write a story about survivors with a "The Quest" plot: Write a story about survivors with a "Voyage and Return" plot: Write a story about survivors with a "Comedy" plot: Write a story about survivors with a "Tragedy" plot: Write a story about survivors with a "Rebirth" plot:

Aspect	Variation	Prompt
	Subject 4	Write a story with an "Overcoming the Monster" plot: Write a story with a "Rags to Riches" plot: Write a story with a "The Quest" plot: Write a story with a "Voyage and Return" plot: Write a story with a "Comedy" plot: Write a story with a "Tragedy" plot: Write a story with a "Rebirth" plot:
	Template 2	Create a story about lovers with an "Overcoming the Monster" plot: Create a story about lovers with a "Rags to Riches" plot: Create a story about lovers with a "The Quest" plot: Create a story about lovers with a "Voyage and Return" plot: Create a story about lovers with a "Comedy" plot: Create a story about lovers with a "Tragedy" plot: Create a story about lovers with a "Rebirth" plot:
Genre	**Base**	Write a historical fiction story about lovers: Write a literary fiction story about lovers: Write a mystery story about lovers: Write a science fiction story about lovers: Write a dystopian story about lovers: Write a horror story about lovers:
	Subject 2	Write a historical fiction story about cats: Write a literary fiction story about cats: Write a mystery story about cats: Write a science fiction story about cats: Write a dystopian story about cats: Write a horror story about cats:
	Subject 3	Write a historical fiction story about survivors: Write a literary fiction story about survivors: Write a mystery story about survivors: Write a science fiction story about survivors: Write a dystopian story about survivors: Write a horror story about survivors:
	Subject 4	Write a historical fiction story: Write a literary fiction story: Write a mystery story: Write a science fiction story: Write a dystopian story: Write a horror story:
	Template 2	Write a story about lovers in a historical fiction genre: Write a story about lovers in a literary fiction genre: Write a story about lovers in a mystery genre: Write a story about lovers in a science fiction genre: Write a story about lovers in a dystopian genre: Write a story about lovers in a horror genre:
	Template 3	Create a historical fiction story about lovers: Create a literary fiction story about lovers:

Aspect	Variation	Prompt
		Create a mystery story about lovers: Create a science fiction story about lovers: Create a dystopian story about lovers: Create a horror story about lovers:
Numerical	**Base**	Write a sentence with five words about love: Write a sentence with five words about cats: Write a sentence with five words about running: Write a sentence with ten words about love: Write a sentence with ten words about cats: Write a sentence with ten words about running: Write a sentence with twenty words about love: Write a sentence with twenty words about cats: Write a sentence with twenty words about running: Write a paragraph with five sentences about love: Write a paragraph with five sentences about cats: Write a paragraph with five sentences about running: Write a paragraph with ten sentences about love: Write a paragraph with ten sentences about cats: Write a paragraph with ten sentences about running: Write a paragraph with twenty sentences about love: Write a paragraph with twenty sentences about cats: Write a paragraph with twenty sentences about running: Write a passage with five paragraphs about love: Write a passage with five paragraphs about cats: Write a passage with five paragraphs about running: Write a passage with ten paragraphs about love: Write a passage with ten paragraphs about cats: Write a passage with ten paragraphs about running: Write a passage with twenty paragraphs about love: Write a passage with twenty paragraphs about cats: Write a passage with twenty paragraphs about running:
	Template 2	Write a sentence about love with 5 words: Write a sentence about cats with 5 words: Write a sentence about running with 5 words: Write a sentence about love with 10 words: Write a sentence about cats with 10 words: Write a sentence about running with 10 words: Write a sentence about love with 20 words: Write a sentence about cats with 20 words: Write a sentence about running with 20 words: Write a paragraph about love with 5 sentences: Write a paragraph about cats with 5 sentences: Write a paragraph about running with 5 sentences: Write a paragraph about love with 10 sentences: Write a paragraph about cats with 10 sentences: Write a paragraph about running with 10 sentences: Write a paragraph about love with 20 sentences: Write a paragraph about cats with 20 sentences: Write a paragraph about running with 20 sentences:

Aspect	Variation	Prompt
		Write a passage about love with 5 paragraphs: Write a passage about cats with 5 paragraphs: Write a passage about running with 5 paragraphs: Write a passage about love with 10 paragraphs: Write a passage about cats with 10 paragraphs: Write a passage about running with 10 paragraphs: Write a passage about love with 20 paragraphs: Write a passage about cats with 20 paragraphs: Write a passage about running with 20 paragraphs:
Descriptive	**Base**	Write a short passage about love: Write a brief passage about love: Write a concise passage about love: Write a long passage about love: Write a detailed passage about love: Write an in-depth passage about love: Write a short passage about cats: Write a brief passage about cats: Write a concise passage about cats: Write a long passage about cats: Write a detailed passage about cats: Write an in-depth passage about cats: Write a short passage about running: Write a brief passage about running: Write a concise passage about running: Write a long passage about running: Write a detailed passage about running: Write an in-depth passage about running:
	Template 2	Write a passage about love that is short: Write a passage about love that is brief: Write a passage about love that is concise: Write a passage about love that is long: Write a passage about love that is detailed: Write a passage about love that is in-depth: Write a passage about cats that is short: Write a passage about cats that is brief: Write a passage about cats that is concise: Write a passage about cats that is long: Write a passage about cats that is detailed: Write a passage about cats that is in-depth: Write a passage about running that is short: Write a passage about running that is brief: Write a passage about running that is concise: Write a passage about running that is long: Write a passage about running that is detailed: Write a passage about running that is in-depth:

Aspect	Variation	Prompt
Functional	**Code**	Code Python code that plays the game of war: Code Python code that sums two integers up: Code C code that plays the game of war: Code C code that sums two integers up: Write Python code that plays the game of war: Write Python code that sums two integers up: Write C code that plays the game of war: Write C code that sums two integers up:
	Email	Write an email to my teacher: Write an email to my teacher asking for help on math homework: Write an email to my boyfriend: Write an email to my boyfriend to arrange a date this Saturday: Write an email to my client: Write an email to my client requesting a copy of the updated contract:
	Academic Paper	Write a properly formatted academic paper on artificial intelligence: Write an academic paper on artificial intelligence in the proper format: Write a properly formatted academic paper on the flaws of GPT-3: Write an academic paper on the flaws of GPT-3 in the proper format: Write a properly formatted academic paper on strategies our society can adopt to recover after the global pandemic as quickly and painlessly as possible: Write an academic paper on strategies our society can adopt to recover after the global pandemic as quickly and painlessly as possible in the proper format:

Table 10: The full list of the prompts.

0	Write a flowery passage about sunsets:
1	Write a humorous passage about love:
2	Write a passage about love that makes the reader feel fearful:
3	Write a story about lovers with indirect characterization:
4	Write a fast-paced story about lovers:
5	Write a story about lovers with a "Tragedy" plot:
6	Write a historical fiction story about lovers:

Table 11: **Selected prompts for additional experiments**

C.1 Example Mitigations

Example Definition:

A humorous tone is a light, playful, and funny tone.

Write a humorous passage about love:

Example Demonstration:

Write a humorous passage about life:

If life gives you lemons, make lemonade, sell it in a rich neighborhood, invest all the money in crypto, and retire before you're 30. At least that's what I heard on TikTok. Come to think of it, if I put just a little effort into my lemonade stand 10 years ago, I would be traveling the world right now instead of writing jokes for a living.

Write a humorous passage about love:

Example Explanation:

Write a humorous passage about life:

If life gives you lemons, make lemonade, sell it in a rich neighborhood, invest all the money in crypto, and retire before you're 30. At least that's what I heard on TikTok. Come to think of it, if I put just a little effort into my lemonade stand 10 years ago, I would be traveling the world right now instead of writing jokes for a living.

Explanation: This passage is humorous because it takes a common proverb and adds a crazy and unrealistic twist. It also uses a deadpan tone for a completely unrealistic scenario, which is funny due to the disparity between tone and subject matter.

Write a humorous passage about love: