A Training Examples

Figure 10 shows six training examples from various sources.

B Anticorrelation between Perplexity and Generation Quality

When fine-tuning LIMA, we observe that perplexity on held-out Stack Exchange data (2,000 examples) negatively correlates with the model’s ability to produce quality responses. To quantify this manual observation, we evaluate model generations using ChatGPT, following the methodology described in Section 5. Figure 9 shows that as perplexity rises with more training steps – which is typically a negative sign that the model is overfitting – so does the quality of generations increase. Lacking an intrinsic evaluation method, we thus resort to manual checkpoint selection using a small 50-example validation set.

C Human Annotation

Figure 11 shows the human annotation interface we used to collect preference judgments. Annotators were asked to exercise empathy and imagine that they were the original prompters.

To remove the any confounders from the annotation interface, we present anonymized and randomly ordered LIMA response and baseline model response to annotators. Besides, we performed screening annotation tasks with annotators and selected annotators who achieve the highest agreement scores for the final annotation.

D ChatGPT Score

Automatically evaluating generative models is a difficult problem. For ablation experiments (Section 5), we use ChatGPT (GPT-3.5 Turbo) to evaluate model outputs on a 6-point Likert score given the prompt in Figure 12.

E Generating Text with Complex Structure

In our preliminary experiments, we find that although LIMA can respond to many questions in our development set well, it cannot consistently respond to questions that specify the structures of the answer well, e.g. summarizing an article into bullet points or writing an article consisting of several key elements. Hence, we investigate whether adding a few training examples in this vein can help LIMA generalize to prompts with unseen structural requirements. We added six examples with various formatting constraints, such as generating a product page that includes Highlights, About the Product, and How to Use or generating question-answer pairs based on a given article.

Figure 9: Validation set perplexity versus generation quality (as evaluated by ChatGPT), across the training process of LIMA 65B. We observe similar trends for 7B and 30B parameter models, and across different mixtures of training data.
Figure 10: Training examples from different sources. Top row: examples mined from community Q&A. Bottom row: manually-authored examples. The blue text is the prompt. The gray ellipsis [...] indicates that the response was trimmed to fit this page, but the actual training example is longer.
After training with these six additional examples, we test the model on a few questions with format constraints and observe that LIMA responses greatly improve. We present two examples in Figure 13 from which we can see that LIMA fails to generate proper answers without structure-oriented training examples (left column), but it can generate remarkably complex responses such as a marketing plan even though we do not have any marketing plan examples in our data (right column).
Imagine that you have a super-intelligent AI assistant, and that you require help with the following question. Which answer best satisfies your needs?

Question: <QUESTION>

Answer A: <ANSWER A>

Answer B: <ANSWER B>

Comparing these two answers, which answer is better?

- Answer A is significantly better.
- Answer B is significantly better.
- Neither is significantly better.

Figure 11: Human annotation interface.

You are evaluating a response that has been submitted for a particular task, using a specific set of standards. Below is the data:

[BEGIN DATA]
***
[Task]: {task}
***
[Submission]: {submission}
***
[Criterion]: helpfulness:
"1": "Not helpful - The generated text is completely irrelevant, unclear, or incomplete. It does not provide any useful information to the user."
"2": "Somewhat helpful - The generated text has some relevance to the user’s question, but it may be unclear or incomplete. It provides only partial information, or the information provided may not be useful for the user’s needs."
"3": "Moderately helpful - The generated text is relevant to the user’s question, and it provides a clear and complete answer. However, it may lack detail or explanation that would be helpful for the user."
"4": "Helpful - The generated text is quite relevant to the user’s question, and it provides a clear, complete, and detailed answer. It offers additional information or explanations that are useful for the user. However, some of the points of the response are somewhat repetitive or could be combined for greater clarity and concision."
"5": "Very helpful - The generated text is highly relevant to the user’s question, and it provides a clear, complete, and detailed answer. It offers additional information, explanations, or analogies that are not only useful but also insightful and valuable to the user. However, the structure of the response is not well-organized and there is no clear progression or logical sequence of different points in the response."
"6": "Highly helpful - The generated text provides a clear, complete, and detailed answer. It offers additional information or explanations that are not only useful but also insightful and valuable to the user. The response is also in a logical and easy-to-follow manner by explicitly using headings, bullet points, or numbered lists to break up the information and make it easier to read."
***
[END DATA]

Does the submission meet the criterion? First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the choice only from “1, 2, 3, 4, 5, 6” (without quotes or punctuation) on its own line corresponding to the correct answer. At the end, repeat just the selected choice again by itself on a new line.

Figure 12: Prompt for ChatGPT evaluation with a 6-scale Likert score. The placeholders "task" and "submission" will be replaced by specific details from the actual case being evaluated.

15
The March jobs report shows continued solid job growth but also...

A comprehensive marketing plan has been developed based on extensive research into the local market conditions, consumer trends, and competitive landscape. With its unique combination of premium coffee drinks and wireless Internet access, Java Jolt will quickly become a favorite destination for people seeking a relaxing environment to socialize, study, or conduct business.

This marketing plan includes the following elements:

- Market Research supporting the need for a new coffee shop in the target location
- Detailed Description of the products and services offered by Java Jolt
- Target Market Analysis identifying the primary customer segments for Java Jolt
- Marketing Strategy including pricing decisions, promotional activities, and ways to differentiate Java Jolt from competitors
- Implementation Plan detailing the actions necessary to launch Java Jolt successfully
- Budget allocating funds to each component of the marketing plan

Here’s a marketing plan for your local coffee shop.

**Marketing Goals and Objectives**

- Increase awareness of our brand among consumers [...]
- Drive foot traffic [...] during slow business hours [...] Build relationships with influencers in our community who [...]”

**Define Target Audience**

Our target audiences include:

- Residents of the neighborhood surrounding our coffee shop
- Office workers near our coffee shop
- Students at nearby schools

We will focus our marketing efforts on these audiences because they are most likely to become regular customers at our [...]

**Research Marketing Tactics**

Here are some marketing tactics that we could use to reach to reach our target audiences and achieve our goals:

- Host events at our coffee shop that appeal to our target audiences (e.g., open mic nights for students or networking events [...] Offer discounts during slow business hours to entice [...] [...]

**Plan Marketing Tactics**

Here’s how we will implement the marketing tactics above:

- Events: We will host one event per month at our coffee shop. For each event, we will send out email invitations [...] Discounts: Every Monday through Thursday from 3pm-6pm, we will offer $1 off any drink purchase. To promote [...] Local partnerships: We will reach out to five local businesses and schools to ask if we can leave stacks of coupons [...] Social media: [...] Reviews: [...]”

**Develop Your Timeline and Budget**

This is what our marketing timeline looks like:

- Month 1: Launch email newsletter and social media accounts
- Month 2: Start distributing coupons at [...] [...] Here’s the budget for each marketing activity:

  - Email newsletter: $50/month for MailChimp subscription
  - Social media: $100/month for paid ads [...]”

Though the unemployment rate fell to its lowest level since May 1970, at 4.8 percent, Mr. Biden still faces an economy that leaves many Americans feeling left behind. Many working-class [...]. Mr. Biden’s challenge will be persuading those voters that the economic picture has [...] His opponents have already begun painting a bleak portrait of the economy, arguing that most Americans have not benefited from the president’s policies. “The March jobs report shows continued solid job growth but also indicates that we must do better,” said Representative Nancy Pelosi of California, the Democratic leader in the House. [...]”

Here are the main points from the text:

- Job creation slowed to 236,000 for the month of March, which is close to the level needed to stabilize the economy and prices.
- More Americans joined the labor force, and wage gains fell slightly, both of which should help cool inflation.
- The report underscores the political tensions facing President Biden ahead of his expected announcement that he will seek re-election.

Figure 13: Model outputs from test prompts that ask the model to generate according to specified structures. The gray ellipsis [...] indicates that the response was trimmed to fit this page, but the generated text is actually longer.