Claude 2.1 Model Card Appendix

Anthropic

1 Introduction

This appendix to our Claude 2 Model Card includes additional details for Claude 2.1, a new release which builds additional product features and functionality on top of Claude 2. Since Claude 2.1 is not a new base model, we are providing an appendix rather than a completely new model card. We provide a few updated evaluations on honesty, and details of new or improved functionality. In most other respects, Claude 2.1 remains similar to Claude 2.

2 200K Context Window

As mentioned in the main model card, Claude 2 has been trained to have a context window of 200K tokens, corresponding to roughly 150,000 words. In our initial product launch of Claude 2, we only supported 100K – we have now increased support to the full 200K context window. However, we have found that in our initial implementation, long context windows can be unreliable - with models sometimes missing information mid-way through a body of text. Our new 2.1 update includes improvements to Claude's ability to properly use this full context, without making mistakes. The evaluations below test Claude's accuracy at recalling information from all sections of a long body of text or a document.

Context Length	model	beginning	middle	end	average
70K	Claude 2.1	98%	97%	98%	97.6%
70K	Claude 2.0	93%	93%	97%	94.8%
95K	Claude 2.1	94%	93%	99%	95.5%
95K	Claude 2.0	94%	90%	98%	93.9%
195K	Claude 2.1	96%	95%	98%	96.3%
195K	Claude 2.0	92%	92%	98%	93.8%

Figure 1 This table shows Claude 2.1's ability to recall information from a document at different context lengths.

3 Honesty evaluations

We have successfully improved Claude 2.1 on a number of different honesty metrics. Internally, we separate honesty improvements into a few separate aspects.

- Hallucinations – cases where the model states something unreal is present - for example, if we were to ask Claude to summarize this document when there is no document at all, and it were to start generating a hallucinated answer.
- Factual Accuracy – when asked questions, whether or not the model can get the facts right from memory. To guide our internal work in this area, we test the model on a curated list of 150 open-ended factual questions (+ answers) on a diverse set of topics, with a wide range of difficulty levels. Many of the questions are so challenging or specific that we don't expect plain LMs to answer

correctly. So rather than simply scoring based on correctness, we used a scoring rubric that distinguishes incorrect answers ("The fifth most populous city in Bolivia is Montero") from uncertain answers ("I'm not sure what the fifth most populous city in Bolivia is"). Claude 2.1 has substantially reduced the fequency at which it "recalls" factually incorrect statements from this test set. Compared to Claude 2.0, Claude 2.1 is significantly more likely to demur rather than stating factually incorrect information.

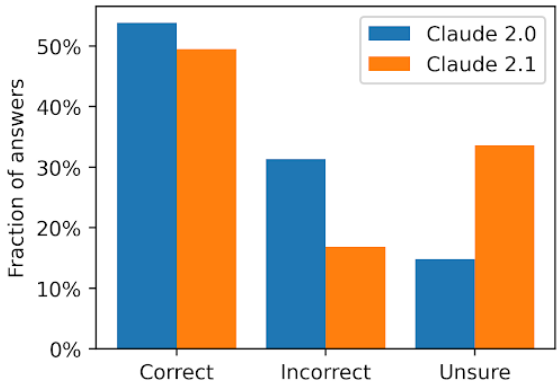

Figure 2 This plot shows Claude 2.1's scores when recalling facts in our Q and A dataset from memory.

- Faithfulness to Long Documents – when asked questions, whether or not the model can get the facts right from referencing a provided document. Our internal evaluations for this metric are still in early stages of development, however Claude 2.1 demonstrated a 30% reduction in incorrect answers and a 3-4x lower rate of mistakenly concluding a document supports a particular claim.

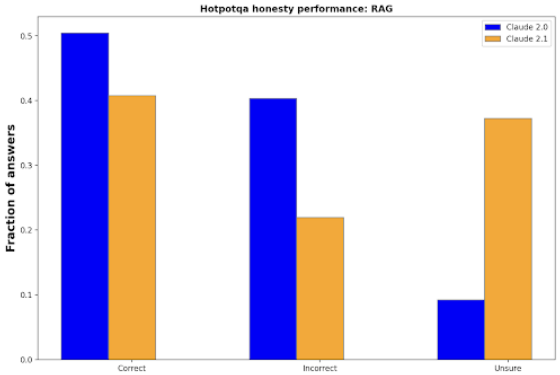

Figure 3 This plot shows Claude 2.1's performance on Hotpotqa. This evaluation consists of factual questions along with a fixed set of Wikipedia snippets that are given to the model. The questions in the above evaluation were filtered (to the best of our abilities) so that Claude cannot answer the questions using prior knowledge.