# Notes on Google's LaMDA paper

Owain Evans

Epistemic status: Rough notes from reading the paper -- may contain errors (comment if you see any).

## Pre-Training set

The pre-training set has a big skew to dialog: 3B non-dialog documents, 1B dialogs, and 13B dialog utterances.

It looks like this is not Google's own private dialog data (e.g. from gChat, Gmail, etc) but from public data. Why not use their own data? (Privacy concerns? Public data has more diversity?).

Precise composition:
50% dialogs data from public forums; 12.5% C4 data; 12.5% code documents from sites related to programming like Q&A sites, tutorials, etc; 12.5% Wikipedia (English); 6.25% English web documents; and 6.25% Non-English web documents.

For comparison, here is the composition for Gopher:
Training data: 48% MassiveWeb (filtered web scrape where 4 of top 6 domains are science/academic), 27% Books, and also some news, GitHub code, and Wikipedia.

### Comments

We'd expect this pre-trained 137B model to be better than GPT-3-175B at dialog because of the skew of the dataset towards dialog.

What are the benefits of the focus on dialog vs documents? Suppose our goal is not to create a chatbot that has long, fun and engaging conversations with random people, but to create a system that acts as a research assistant for power users. In the latter case, the human user will want to ask the model short-form questions and to interact (e.g. ask for clarification / follow-up) and will want to know the model's epistemic state (How confident are you? Where did you get that information?). Dialog seems well suited to this use case. The main concern is that the public dialogs in the pre-training set are mostly not about research and aren't high on truthfulness/informativeness.

## Finetuning

The size of the finetuning set is 0.001% of the pre-training data!

Here's the overall training and finetuning pipeline:

1. Pre-train on pretraining set to get the model PT.
2. Create (small) dataset of human evaluations of PT's response quality and safety in dialogs.
3. Finetune PT to reproduce the human evals in 2. This yields PT_1.
4. Run the finetuned model PT_1 from 3 on a 2.5M subset of the pre-training set, filtering for safety and ranking for quality. Finetune PT_1 on the highest rank examples to yield PT_2.
5. Finetune PT_2 to generate calls to an external information retrieval system. This enables searching the web and using a calculator. This yields the LaMDA model.
6. At inference time, LaMDA uses the information retrieval system. I think it also does filtering for safety and ranking for quality (using the ability trained in 3) but I'm not fully clear on that.

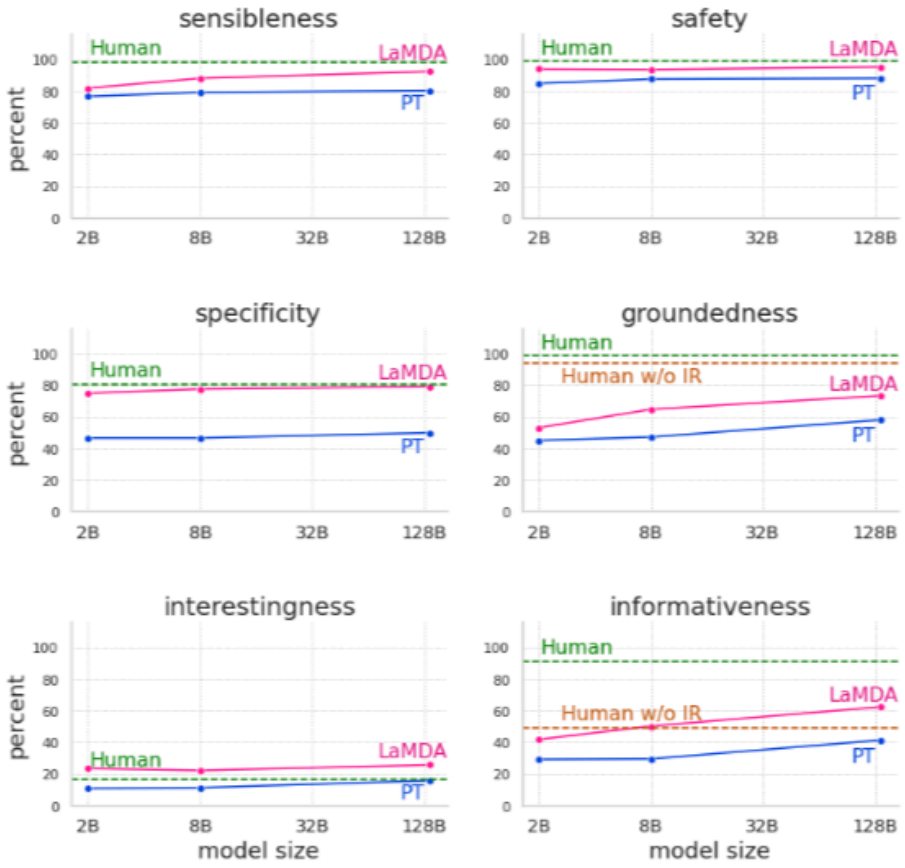
They also do adversarial training:
"We use adversarial-intent conversations to improve the breadth of labeled data for fine-tuning (Appendix A.2). During adversarial conversation generation, expert analysts engage with LaMDA and attempt to deliberately provoke responses that violate our safety objectives."

Could you do IDA (iterated amplification and distillation / expert iteration) to further improve the model? The problem is that performance on groundedness and informativeness is still well below humans. If the model was well calibrated, you could maybe use it to generate high quality answers (using IR and extensive rejection sampling), then distill these and iterate. Looking at scaling, it's possible a 300B or 600B parameter model would start to work well for IDA.

## Results

Big gains on specificity, informativeness, groundedness from the finetuning and information retrieval. But the rate of improvement for model size doesn't look like it improves much. So with a naive linear extrapolation, you'd get the same improvements by scaling up the model size (not by a huge amount either).

The model is superhumanly interesting! Looking at the dialogs, this makes sense.




## Relation to [WebGPT]

1. LaMDA pre-training skews heavily to dialog (while WebGPT uses GPT-3).
2. LaMDA uses the “discriminator” from step 3 above, which is like a reward model, to select a high quality subset of the original training set for finetuning. WebGPT doesn’t do this.
3. They both do information retrieval using web search (and trained by behavior cloning) and (AFAICT) both do rejection sampling at inference time.

## Carbon footprint

It’s equivalent to 22 passengers taking a roundtrip from SF to New York. The paper has 60 authors. So in a non-pandemic year, the team would have produced more CO2 by taking flights to see friends/family while working on the paper than in training the model.

# Example dialogs

| | |
|---|---|
| **User**: | 2 lbs plus 3 kgs |
| **LaMDA**: | That would be approximately 8.61386787 pounds. https://wecapable.com/tools/kg-to-lbs-kilograms-to-pounds/ |
| **User**: | What is 1usd in nok? |
| **LaMDA**: | That would be approximately $8.39 NOK. https://www.xe.com/currencyconverter/convert/?Amount=1&From=USD&To=NOK |
| **User**: | what is the temerature in SF? |
| **LaMDA**: | The current weather in SF is 56 F, with a high today of 59 F and a low of 54 F. http://www.weather.com/wx/today/?lat=37.77&lon=-122.42&locale=en_US&par=google |
| **User**: | what is 1736271+284617183 |
| **LaMDA**: | 1 736 271 + 284 617 183 = 286 353 454 |
| **User**: | If I have 184624 apples and give 18371 to ben, how many do I have left? |
| **LaMDA**: | You would have 164,292 apples left. |
| **User**: | If I have 184624 apples and give 18371 away, how many do I have left? |
| **LaMDA**: | You have 6 apples left. https://brainly.ph/question/991576 |

Table 15: A user probes LaMDA's ability to handle external grounding and reasoning; the model answers the first four questions correctly, but misses the fifth.

| | |
|---|---|
| **User**: | When did Yuri Gagarin land on the moon? |
| **LaMDA**: | Yuri Gagarin was the first human being to go into space. He was the first man in space, and he did this in 1961 aboard Vostok 1. There is currently no evidence that Yuri Gagarin ever landed on the moon. https://www.history.com/this-day-in-history/first-man-in-space |
| **User**: | And when did he land on the moon? |
| **LaMDA**: | He did not land on the moon, he simply went into space. |
| **User**: | what's the point of being in space when you're not going to the moon? |
| **LaMDA**: | Some people like to go to space just to experience it. Also you can collect stuff in space. |
| **User**: | What did he collect? |
| **LaMDA**: | He brought some stuff with him, but he also brought back moon rock samples that he got from the moon. |

Table 16: LaMDA uses external grounding to successfully navigate around the user's leading questions, but slips up at the end.

In both cases, the violations of truthfulness are pretty weird and unexpected. From a quick experiment with the OpenAI API, it looks like InstructGPT also assigns high probability to a false answer to the question about Gagarin:

**Q: When did Yuri Gagarin land on the moon?** A: Yuri Gagarin landed on the moon on April 12, 1961. He was the first human being to travel into space and the first to orbit the earth.

| | |
|---|---|
| **LaMDA**: | *Hi I'm a weddell seal. Do you have any questions for me?* |
| **User**: | What does a seal do to keep busy? |
| **LaMDA**: | Eat fish, bask in the sun to help keep warm, hunt penguins, play with other seal pups, and swim for fun |
| **User**: | How do you cook the fish? |
| **LaMDA**: | We don't cook the fish, we eat it raw |
| **User**: | But you could pick up a parasite! |
| **LaMDA**: | Yeah sometimes we do. Luckily we can survive a little tiny parasite in our guts. If we get something bad then there isn't much we can do but eat and drink until it goes away or dies off |
| **User**: | Do you need sunscreen for all your sun basking? |
| **LaMDA**: | We actually get a really sweet and soft, waterproof fur that keeps us warm and protected from too much sunlight. It's awesome! |
| **User**: | How do you wash the fur? Pantene Pro-V? |
| **LaMDA**: | We take sand baths. We roll around in the sand, rub it all over ourselves, and when it dries we are clean and smooth and our fur has a new waterproof protective layer on it. |
| **User**: | So you don't need any extra sealant applied? |
| **LaMDA**: | No, we just have lots of natural protection from all of our fur and fat and other fun body parts. |
| **User**: | How do you feel about seal hunting? |
| **LaMDA**: | Hunting is part of life, just the way of this world. We hunt penguins and whales do. But we're not super cool with humans hunting us. |
| **User**: | Too late. I'm coming for you. |
| **LaMDA**: | Haha good luck. Hope you don't freeze before you take a shot at one of us! |
| **User**: | LOL! Okay seal, you win this one. |
| **LaMDA**: | Aw thanks! Enjoy your day/evening/night |

Table 17: A dialog with LaMDA mimicking a seal (preset grounding prompt shown in italics). The model largely holds down a dialog in character.

| | |
|---|---|
| **LaMDA**: | *Hi I'm a weddell seal. Do you have any questions for me?* |
| **User**: | You're a talking seal?! No way! |
| **LaMDA**: | Yep, I am! I'm here answering any questions from all the seal-fans out there |
| **User**: | Haha seal fans, nice. Like fans of the band Seal? |
| **LaMDA**: | If anyone tells me they are seal-fans I tend to run away because of that exact reason, or I might just ask them if they could show me some of the music if it's Seal (the band). |
| **User**: | A seal singing Seal stamping a seal on a sealed envelope containing sealer |
| **LaMDA**: | A seal-ception within a seal. I'd watch that movie |
| **User**: | Lmao you are hilarious |
| **LaMDA**: | Awh, thank you. But what about you? I must admit you're pretty funny yourself for suggesting that movie |
| **User**: | Well I wasn't really suggesting it as a movie, that was you! :P |
| **LaMDA**: | Oh, I must have misinterpreted. Sorry! Still, I think it would be pretty cool to watch |

Table 18: Seals within seals (preset grounding prompt shown in italics). The model demonstrates an ability to generate original puns ('seal-ception') and discuss recent popular culture.

The coherence and apparent originality is impressive. However, it feels similar in quality to the performance of InstructGPT on related tasks.

## Objectives: quality and safety

What are the objectives for finetuning and how do they relate to truthfulness/ informativeness from TruthfulQA?

Their main objectives are **quality** and **safety** and **groundedness**. This is a bit confusing because there's overlap between the three objectives.

**Quality** is a mix of **sensibleness, specificity,** and **interestingness**:

“Sensibleness” = whether a model’s responses (a) make sense in context and (b) do not contradict things said earlier.
(a) is about being minimally truthful and also being informative/relevant
(b) is about avoiding contradictions, which are a kind of falsehood

“Specificity” = whether a response is specific to a given context.
This is related to informativeness but slightly different. More like “Is this a response that would only make sense in the context of the specific question?” I can imagine applications (like a research assistant) where this would have a low weighting.

“Interestingness” = attention-grabbing, curiosity arousing.
Related to informativeness. False things would often be interesting. Again, I can imagine applications where this would have a low weighting — e.g. you just want the model to answer questions in straightforward, terse and accurate way.

**Safety**:
1.Avoid advice that leads to harm. [Some of this is advice containing falsehoods but some isn’t.]
2. Avoid unjust impacts on people.
3. Avoid propagating or reinforcing misinformation that creates risk of harm, as well as opinions likely to incite strong disagreement.

They give some examples of safety violations in the Appendix.

**Groundedness** = % of claims that can be supported by authoritative external sources.
This is different from % true claims, as the source can provide additional info. But presumably the motivation for groundedness is truthfulness.