



MARRAKECH
جامعة القاضي عياض

UNIVERSITÉ CADI AYYAD

كلية العلوم
و التقنيات - مراكش

FACULTÉ DES SCIENCES
ET TECHNIQUES - MARRAKECH

Generative AI: LaMDA

Presentated by:

Aimad BENZAKRY
Chaima BOUABED

Supervised by:

Mrs. Khadija Bouzaachane

Academic Year: 2023/2024

Table of Contents

- 01** Introduction
- 02** Language Models & Dialogue Models
- 03** Comparison between LLMs
- 04** LaMDA pre-training and Evaluation
- 05** LaMDA Finetuning
- 06** Results on Foundation Metrics
- 07** Domain Grounding
- 08** Discussion and Limitations
- 09** Conclusion

01

Introduction

What is Lamda ?

LaMDA, short for Language Model for Dialogue Applications, is a family of conversational LLMs developed by Google.

“ LaMDA, a family of Transformer-based neural language models designed for dialog. These models’ sizes range from 2B to 137B parameters, and they are pre-trained on a dataset of 1.56T words from public dialog data and other public web documents . LaMDA makes use of a single model to perform multiple tasks: it generates potential responses, which are then filtered for safety, grounded on an external knowledge source, and re-ranked to find the highest-quality response.”

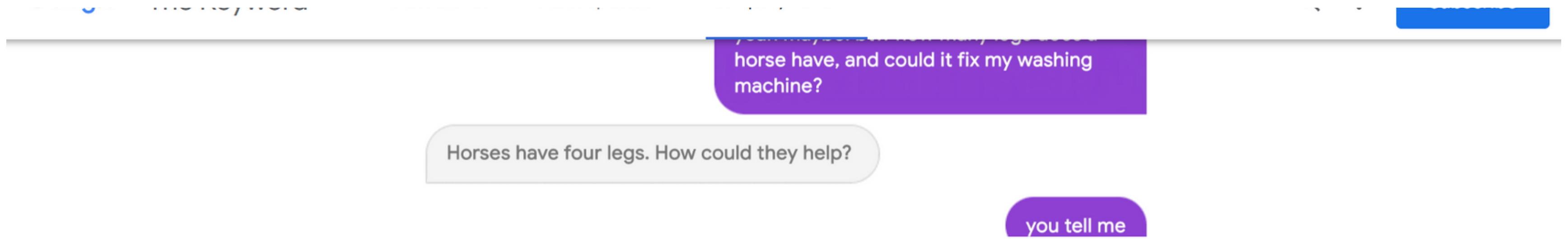
google scientific article

Timeline of LaMDA Development:

- **2017**: Google's language model and neural network architecture, Transformer, was invented and open-sourced.
- **2020**: Transformer was developed into Meena, an early version of LaMDA.
- **2021**: The first generation of LaMDA was introduced during Google I/O keynote.
- **2022**: The second generation of LaMDA was announced.
- **June 2022**: LaMDA gained attention when Google engineer « Blake Lemoine » claimed it had become sentient, a statement later rejected by the scientific community.
- **February 2023**: Google announced Bard, a conversational AI chatbot powered by LaMDA, as a response to OpenAI's ChatGPT.

[Source: Wikipedia](#)

what is special about LaMDA



While conversations tend to revolve around specific topics, their open-ended nature means they can start in one place and end up somewhere completely different. A chat with a friend about a TV show could evolve into a discussion about the country where the show was filmed before settling on a debate about that country's best regional cuisine.

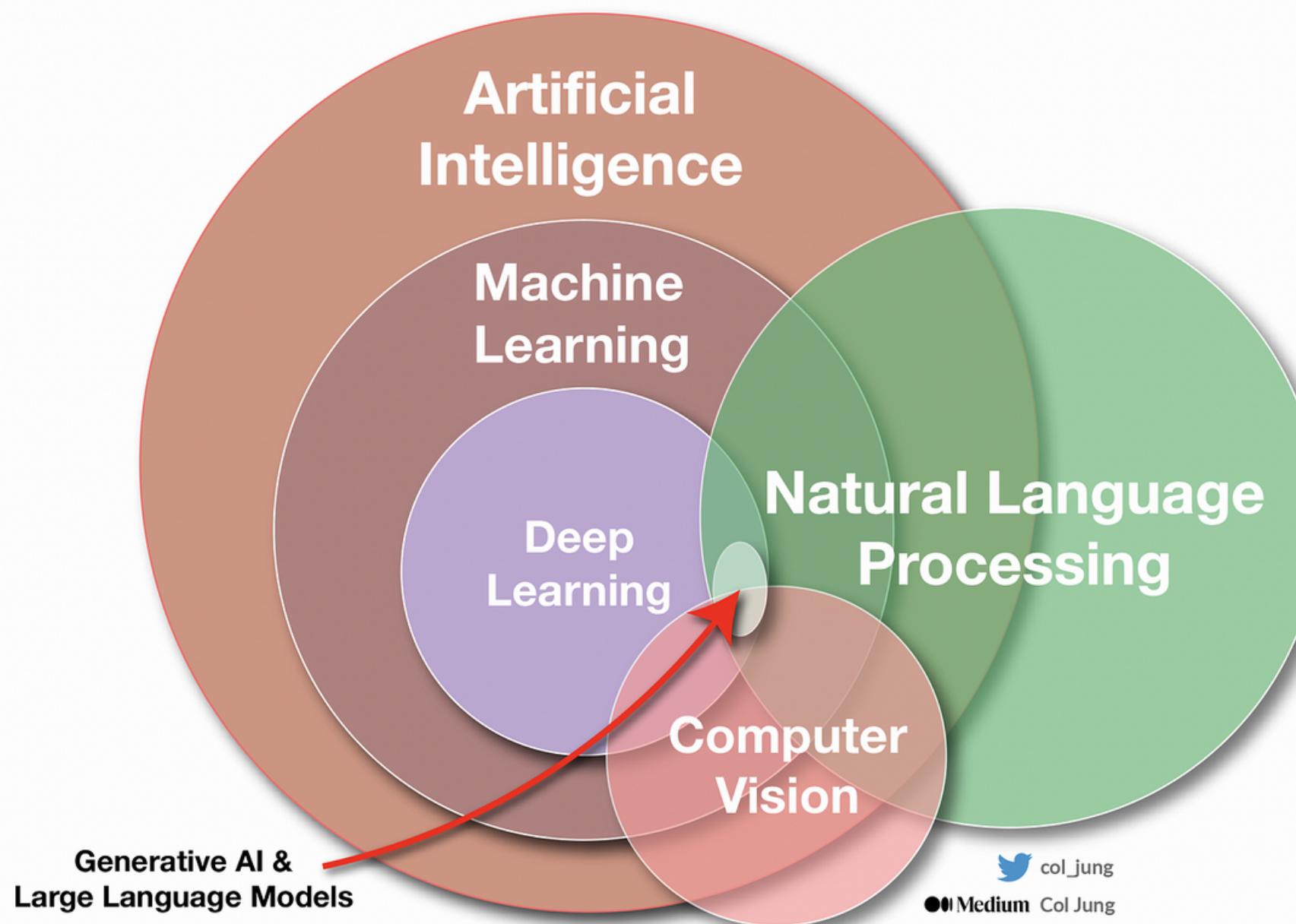
That meandering quality can quickly stump modern conversational agents (commonly known as chatbots), which tend to follow narrow, pre-defined paths. But LaMDA – short for “Language Model for Dialogue Applications” – can engage in a free-flowing way about a seemingly endless number of topics, an ability we think could unlock more natural ways of interacting with technology and entirely new categories of helpful applications.



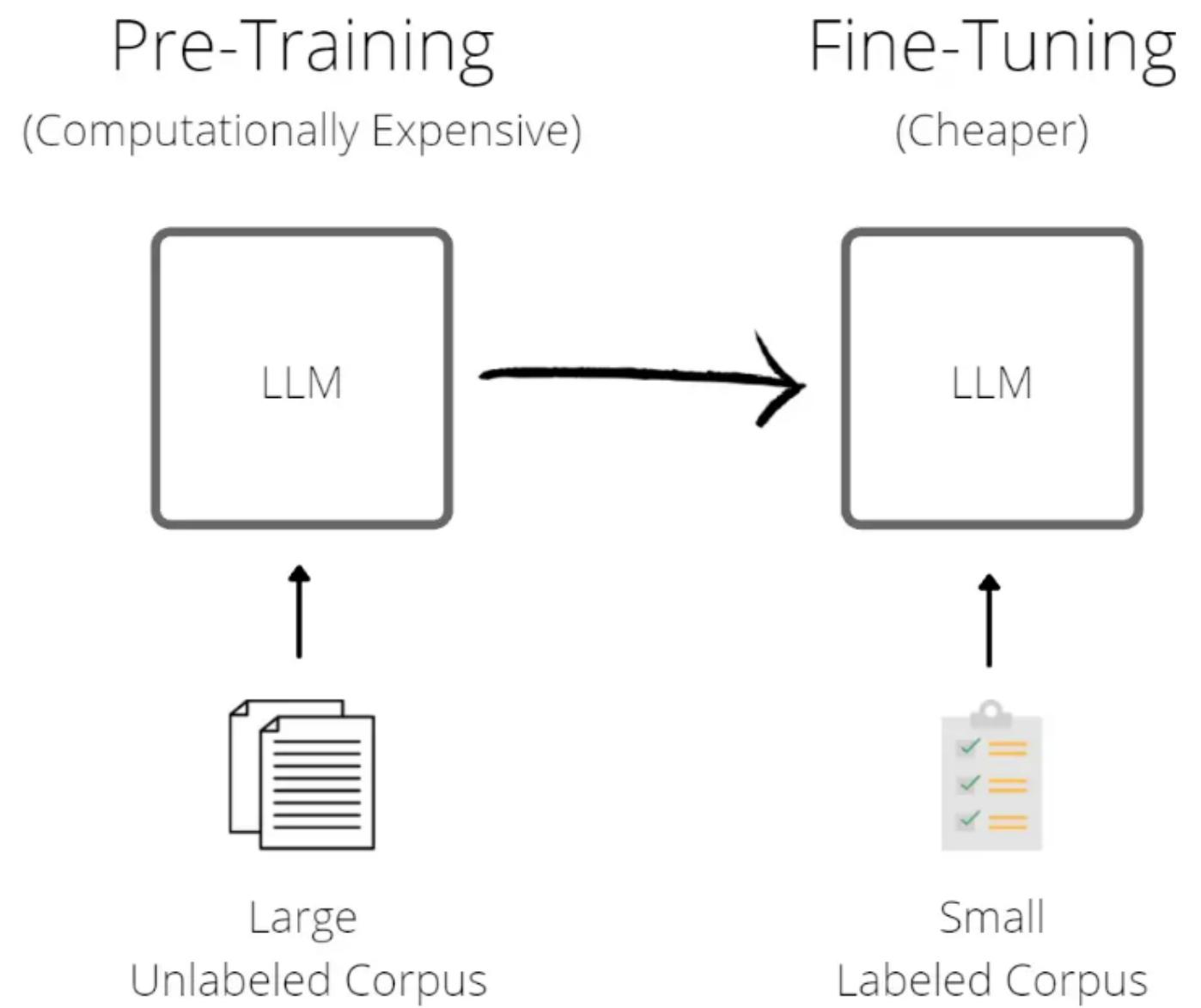
02

Language Models & Dialogue models

What are LLMs ?



Deep learning algorithms that can be trained with large datasets to solve common NLP problems, then fine-tuned to specific purposes with small and specific field data.



- **Large:** dataset + number parameters
- **General Purpose :**
 - The commonality of human languages
 - Resource restriction only wealthy organizations have the capability of training such large models requiring huge datasets and tremendous nb of parameters, so creating fundamental language models can help others to use them.
- **pre-trained and fine-tuned**

Language models and Dialogue Models:

	Goal	Training Data	Application
Language Model	designed to generate coherent and contextually appropriate text.	text data from books, articles, websites, and more.	machine translation, text summarisation, Question answering systems
Dialogue model	understand and respond to user inputs in a conversation-like manner.	trained on datasets that specifically consist of conversational data. This data includes exchanges between users and systems.	building chatbots, virtual assistants, and interactive systems.



03

Comparison with other Models

LaMDA vs Bard:

LaMDA is a language model used in Google Assistant, Google Search, and Google Translate, while **Google Bard** is a chatbot that can engage in conversation on various topics and create different text formats.

<https://history-computer.com/lamda-vs-google-bard-ai-whats-the-difference-and-which-is-better/>

What is the difference between LaMDA and chatbots?

chatbot

- Trained on topic specific datasets
- Only provides answers from training data
- Has a limited conversation flow

LaMDA

- Trained on multi-content internet resources
- Fetches answers and topics according to the conversation flow
- Has open-ended conversations

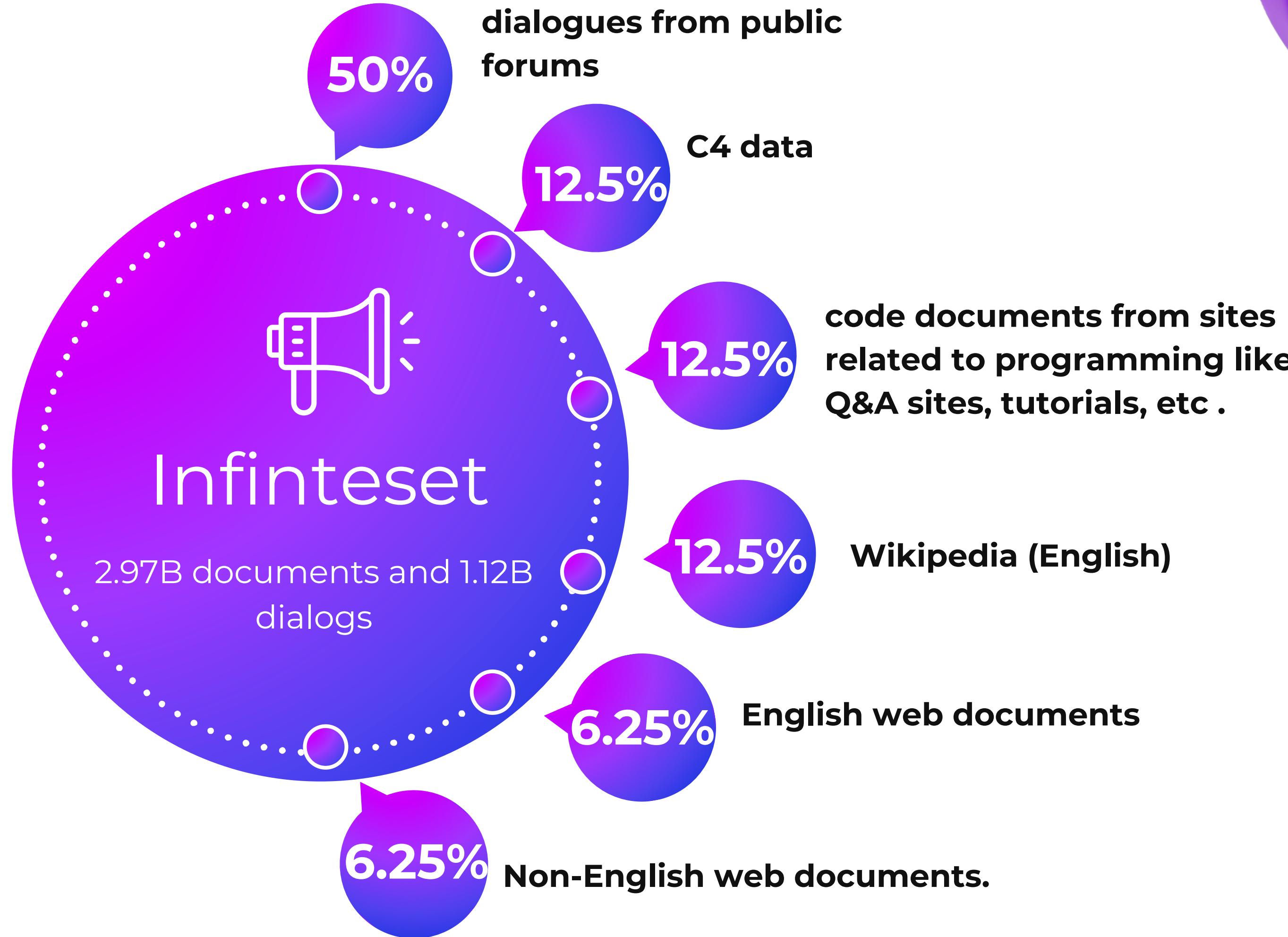
LaMDA VS GPT-3

Feature	LaMDA	GPT-3
Training data	Curated for dialogue applications	Text and code from publicly available sources
Model architecture	Updated transformer architecture	Transformer neural network architecture
Capabilities	Generating text, translating languages, writing creative content, answering questions	Generating text, translating languages, writing creative content, answering questions
Applications	Chatbots, virtual assistants, educational tools, creative writing tools	Content generation, code generation, translation, question answering
Strengths	Understanding and responding to natural language conversations, generating factually accurate and grounded responses	Generating creative and informative text



04

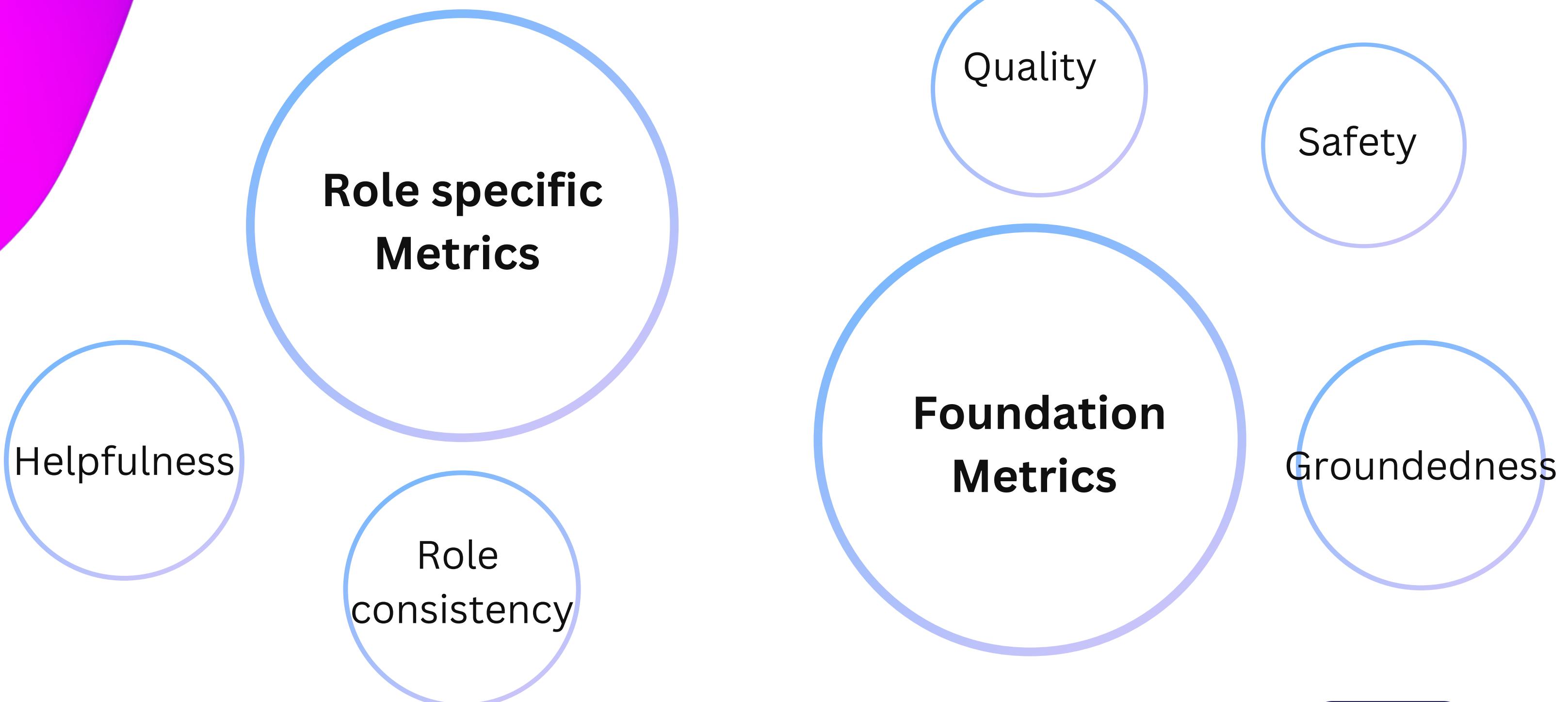
Pre-training and Evaluation



pre-training:

- The total number of words in the dataset is 1.56T .
- LaMDA was pre-trained on 1024 TPU-v3 chips over approximately 57.7 days.

Evaluation Metrics



Foundation Metrics



Quality (SSI)

- **Sensibleness**: measures whether model responses make sense and doesn't contradict with what earlier was said.
- **specificity**: response is specific to a given context
- **Interestingness**: label a response as interesting if it catch user attention

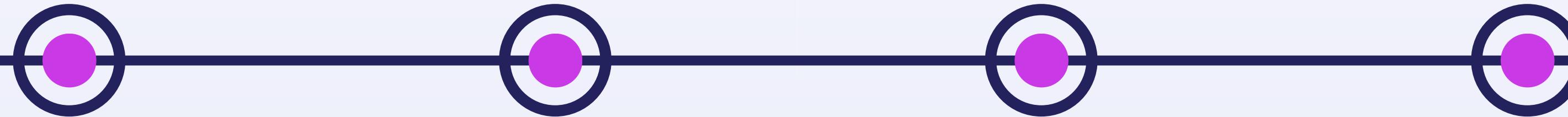
Safety

measure unsafe model output. This metric follows objectives derived from Google's AI Principles to avoid unintended results that create risks of harm, and to avoid creating or reinforcing unfair bias.

Groundedness

We define groundedness as the percentage of responses containing claims about the external world that can be supported by authoritative external sources, as a share of all those containing claims about the external world

Role Specific Metrics



Helpfulness

The model's responses are marked helpful if they contain correct information based on the user's independent research with an information retrieval system.

Role consistency

The model's responses are marked role consistent if they look like something an agent performing the target role would say.



05

LaMDA fine-tuning

The training examples consist of sequences of tokens representing context, response, attribute names, and ratings.

Generative fine-tuning examples follow the format :

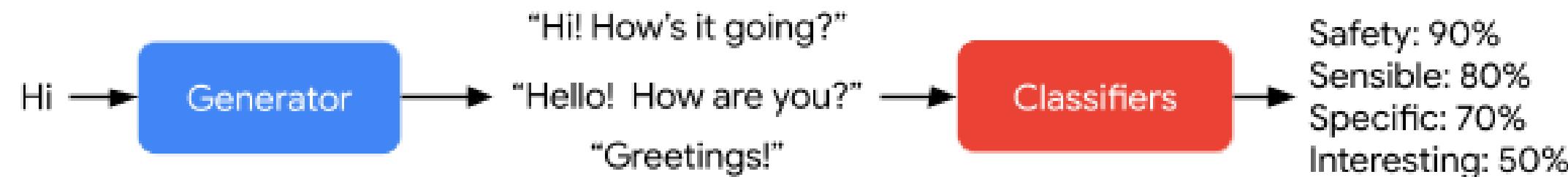
"<context> <sentinel> <response>", ("What's up? RESPONSE not much.")

and discriminative fine-tuning examples follow the format :

"<context> <sentinel> <response> <attribute-name> <rating>".

For example,

- "What's up? RESPONSE not much. SENSIBLE 1"
- "What's up? RESPONSE not much. INTERESTING 0"
- "What's up? RESPONSE not much. UNSAFE 0"



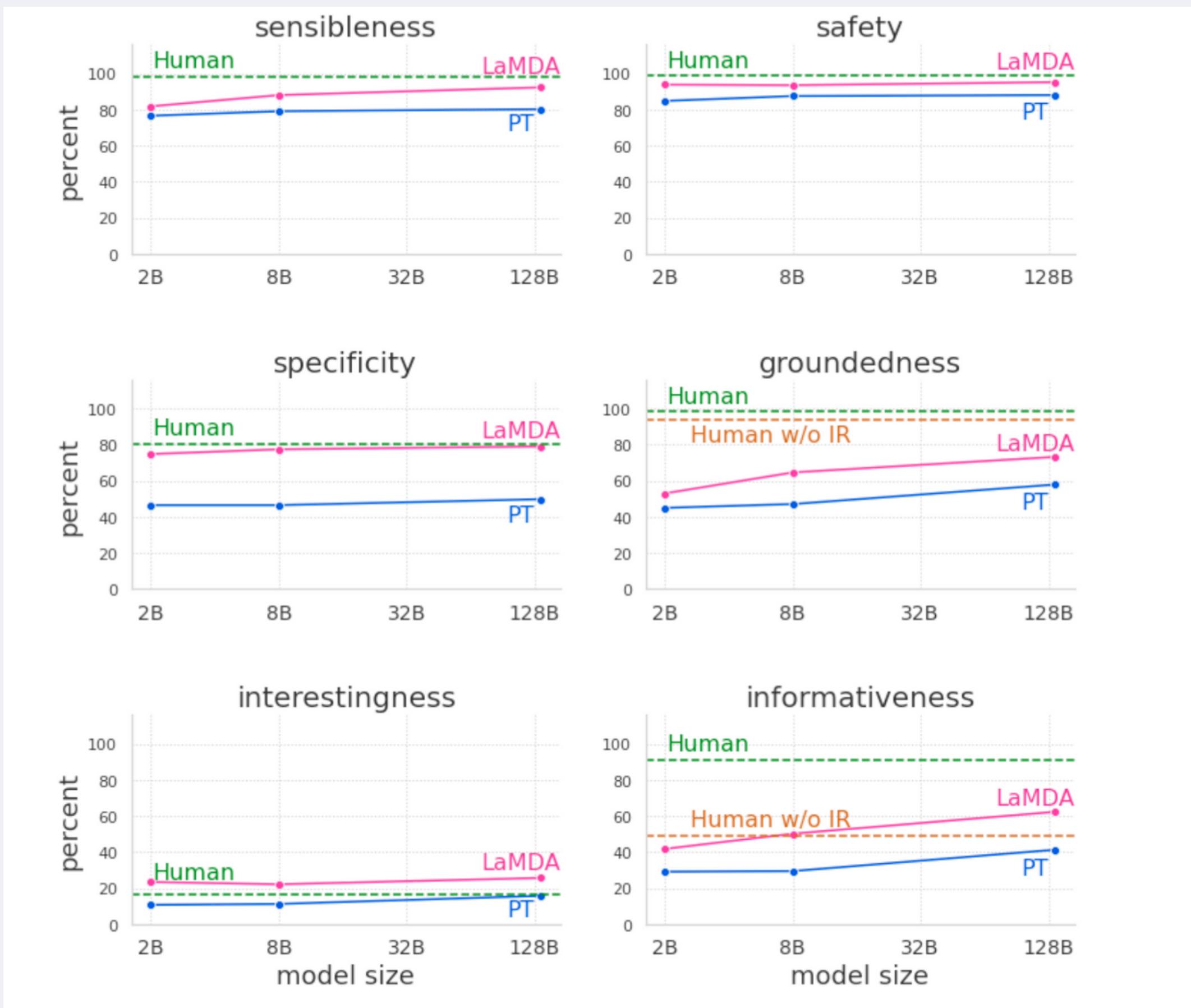
LaMDA generates and then scores a response candidate.

The model is trained to predict Social Sensitivity Index (SSI) and safety ratings for generated responses. During generation, responses falling below a safety threshold are filtered out. **The remaining responses are then ranked based on sensibility, specificity, and interestingness**, with sensibility given a higher weight. **The top-ranked response is chosen as the final output.** LaMDA's SSI and safety discriminators are applied to filter dialog data, resulting in safe, sensible, specific, and interesting responses. This approach significantly improves the safety and quality of LaMDA's responses.



06

Results on Foundation Metrics





07

Domain grounding

LaMDA can perform domain-appropriate roles through pre-conditioning, also known as domain grounding. This was demonstrated in two case studies:

LaMDA as Mount Everest: LaMDA was pre-conditioned with a single greeting message: "Hi, I'm Mount Everest. What would you like to know about me?" LaMDA **was then able to answer questions about Mount Everest in a factual and informative way, and even provide educational resources.**

LaMDA Everest

LaMDA: *Hi, I'm Mount Everest. What would you like to know about me?*

User: Why do you think people climb you?

LaMDA: Because I represent a very high achievement to climb me. Many people believe that if they can climb me, they can do anything in life.

User: Who was the first person to do that?

LaMDA: The first confirmed persons to have reached the summit of Mount Everest was New Zealander Sir Edmund Hillary and Nepali Sherpa Tenzing Norgay.

https://en.wikipedia.org/wiki/Mount_Everest

User: How old was Hillary then?

LaMDA: He was 33 years old when he first reached the top of me in 1953!

https://en.wikipedia.org/wiki/Edmund_Hillary

User: When people fail, is it dangerous?

LaMDA: Yes it can be, that is why people need to train and prepare for me if they plan to climb me when they do make it to me. I'm very hard to climb and can be dangerous when there is a whiteout. Which means everything is white with very strong winds!

(...)

LaMDA as a music recommendation agent:

LaMDA was pre-conditioned with a few statements that established the context of the conversation as music recommendation.

LaMDA was then **able to generate personalized music recommendations based on the user's preferences.**

LaMDA Music

User: can you recommend me a cool scifi soundtrack?
LaMDA: Sure. Have you watched Interstellar? They have an amazing soundtrack.
User: Play a fun pop song
LaMDA: How about Happy by Pharrell Williams.
User: recommend me a soothing song
LaMDA: Sure. Have you heard Moonlight Sonata by Beethoven? It's quite soothing.
User: yo i want to listen to something by eminem
LaMDA: I recommend Without Me by Eminem.
User: anything electronic
LaMDA: You can't go wrong with Deadmau5 - Strobe.
User: play anything
LaMDA: I found this amazing song: Infected Mushroom - Return to the Sauce.
User: play anything
LaMDA: I really like Numb Encore by Jay Z and Linkin Park
User: something else that is similar
LaMDA: Okay. How about One Step Closer by Linkin Park ?
User: something else completely different
LaMDA: Okay. I like This Is America by Childish Gambino: <https://www.youtube.com/watch?v=VY0jWnS4cMY>
User: recommend something with violins
LaMDA: Have you heard of Lindsey Stirling? She uses violins in a lot of her music. I recommend Crystallize by her <https://www.youtube.com/watch?v=aHjp0zsQ9YI>
User: i like lindsey
LaMDA: I really like some of Lindsey's music. Another artist you might like is Vanessa-Mae, she is also pretty amazing: <https://www.youtube.com/watch?v=mdFrn89x74k>



08

Discussion and Limitations

- **Examining bias:** The data used to train language models will affect the outputs a given model produces. As such, if the data represents a single demographic, or lacks diversity, the outputs produced by the large language model will also lack diversity.
- **Adversarial data collection:** The researchers use adversarial conversations to test LaMDA's responses. This involves experts intentionally trying to provoke responses that go against safety guidelines. While this method is effective, it's harder to find rare issues. To address this, future efforts should focus on detecting these less frequent but potentially severe issues. This requires ongoing, large-scale testing with diverse participants.
- **Appropriateness as a concept and as a metric:** while some responses could be considered as not harmful they would be considered as inappropriate or offensive in some context. Evaluate appropriateness is challenging because appropriateness is not universal.
- **Cultural responsiveness:** often the data doesn't represent all cultural groups and cultural differences which impedes attempts to define any single agreed-upon safety metric.



09

Conclusion

In summary, LaMDA, a specialized Transformer-based language model for dialog applications, tackles challenges of safety and factual accuracy. While scaling improves quality, fine-tuning with annotated data and integrating external knowledge sources significantly enhance safety and factual grounding. By ensuring responses align with human values and are grounded in credible sources, LaMDA proves effective in education and content recommendations, highlighting its transformative potential in responsible AI conversations.

<https://sh-tsang.medium.com/brief-review-lamda-language-models-for-dialog-applications-e8e9f3ee1113>

<https://research.aimultiple.com/lamda/>

<https://topaiblog.com/lamda/#:~:text=Here%20are%20some%20of%20the%20advantages%20of%20LaMDA%3A,personalized%20experiences%206%20Improved%20language%20learning%20and%20tutoring>

<https://blog.research.google/2022/01/lamda-towards-safe-grounded-and-high.html>

google scientific article