

AI SAFETY + ALIGNMENT

Q&A CHATBOT

April 13, 2023



01

OVERVIEW

Stampy / aisafety.info
LessWrong Bounty

02

DEMOS

Key Features
Challenges
Discussion

03

PLANNING

What's next?



Rob Miles

“AI is leaping forward right now,
it's only a matter of time before we develop
true Artificial General Intelligence, and there
are a lot of different ways that this could go
badly wrong for us.”

AI SAFETY PROJECTS

STAMPY

Curated Q&A
stored in
gdocs + coda

STAMPY-UI

Semantic search
chat.stampy.ai
chat.aisafety.info

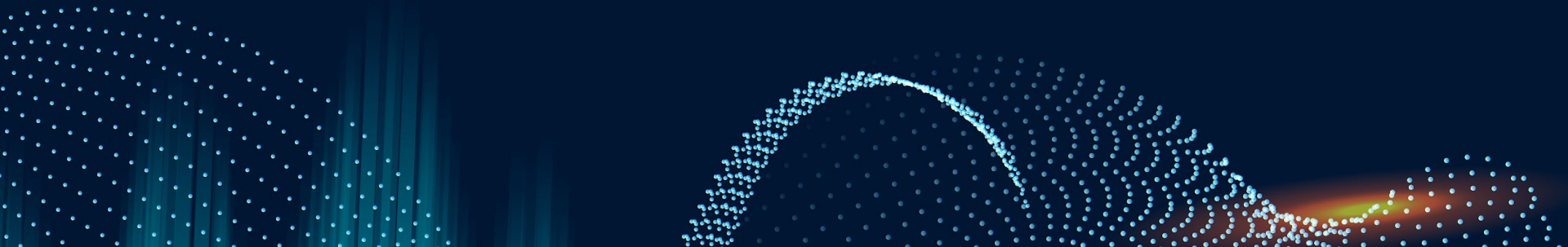


Discord bot

FAQ

stampy.ai
aisafety.info

NLP



**“Speed running everyone through
the bad alignment bingo.
\$5k bounty for a LW
conversational agent”**

—ArthurB

KEY COMPONENTS



Data

Alignment
Research Dataset

Kept up-to-date, chunked,
embedded for vectorstore.



Chat API

Use LLM to
generate answers

Given user query
find relevant chunks.



Front-End

GUI, web, bot or
other app calls API

Accepts user query and
tracks chat history.

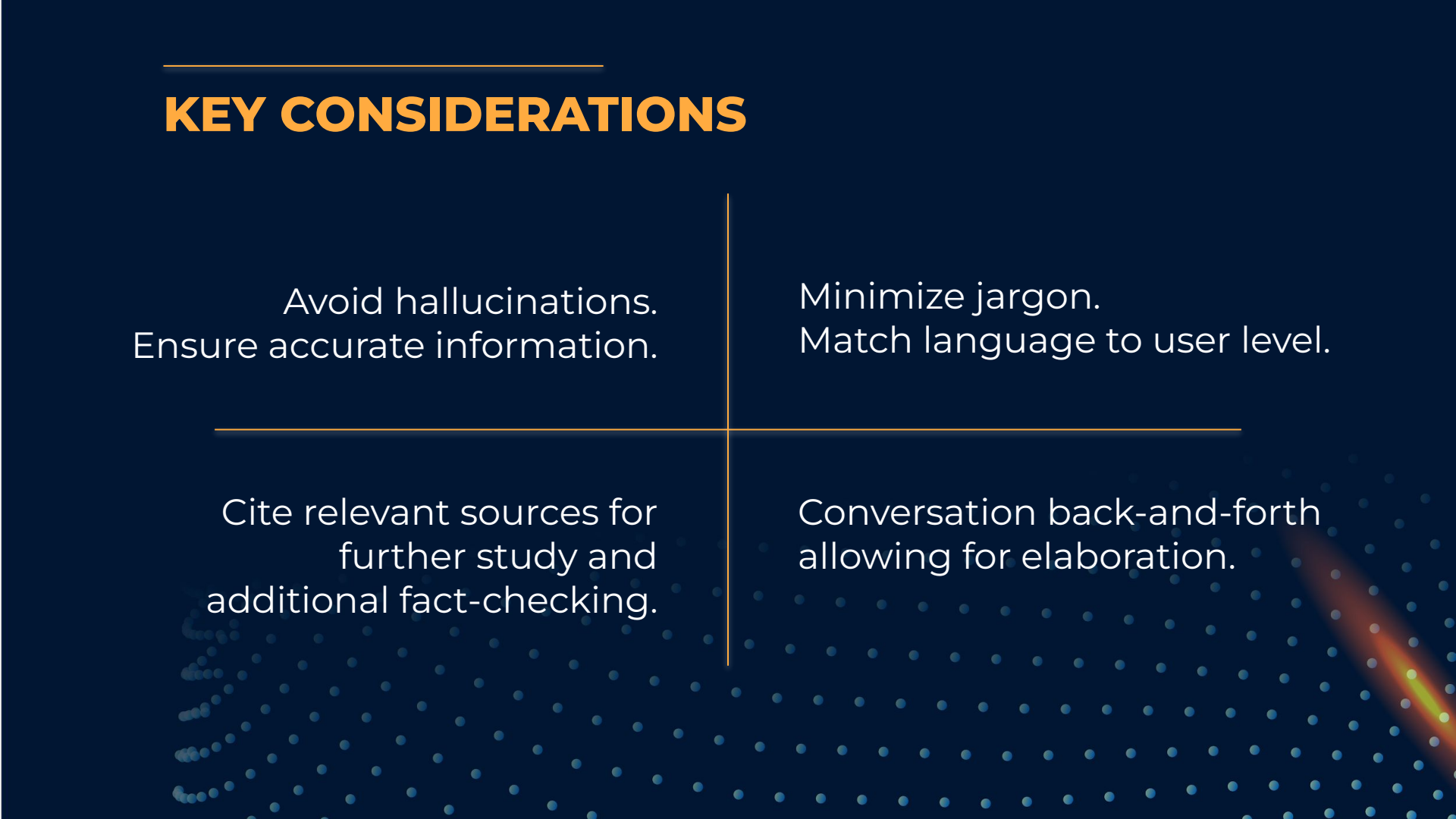
KEY CONSIDERATIONS

Avoid hallucinations.
Ensure accurate information.

Minimize jargon.
Match language to user level.

Cite relevant sources for
further study and
additional fact-checking.

Conversation back-and-forth
allowing for elaboration.



DEMOS

Stampy Chat

<http://chat.stampy.ai/>

Thomas

<https://ai-safety-conversational-agent.thomasbroadley.com>

McGill Team's AlignmentSearch

<https://alignmentsearch.up.railway.app/>

Craig AlignmentGPT

<http://tidblitz.com/>

Johnathan

<https://discord.gg/HkRkbC7n>



langchain

Index

Doc Loader
Recursive Splitter
Embed for Vectorstore

Models

LLM Support
Completions +
Embedding

Memory

Conversation
Buffer/Window/Sum

Chains

Retrieval QA+Source
Chat over Docs
rephrase query context

Agent

ReAct docstore
conversational desc

Future

Updates from
Research + Releases

HOW TO CONTRIBUTE?

Questions +
Next Steps

- Dataset update, plug-in
- Integrating implementations
- Langchain
- Prompt Engineering
- Citations
- Web Interface
- Testing

THANKS!

Your attendance today and
contribution is much appreciated!

CREDITS: This presentation template was
created by Slidesgo, including icons by Flaticon,
and infographics & images by Freepik.

