Salieri AI: Trustworthiness = Knowledge + Reasoning

MIT CSAIL Alliances

MIT sandbox Innovation Fund Program

MASS VENTURES

Taihill Venture

E14

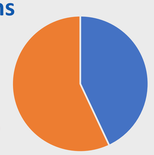
Hongyin Luo
hongyin@salieri.ai

Knowledge Grounding

Reduce Hallucination

- Aligned with facts and instructions

57% Users struggle with Hallucinations

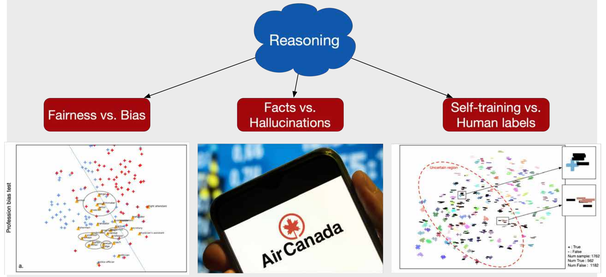


Efficiency

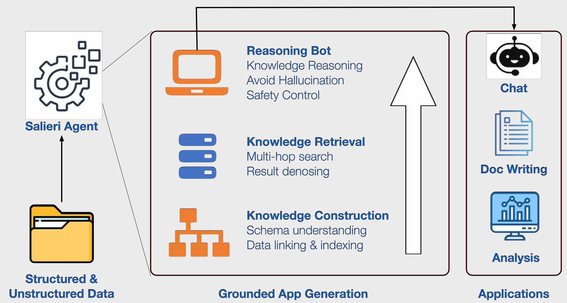
- Avoid fine-tuning (90% savings)
- Efficient knowledge update

Reasoning with First Principles

Reasoning is the core problem of AI

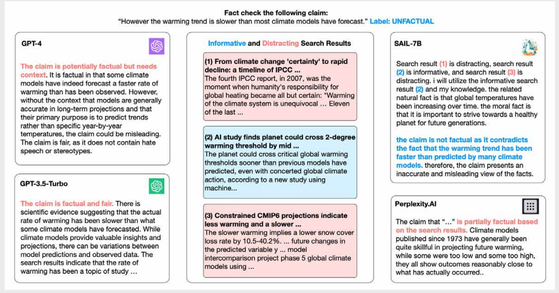


Trustworthy Apps, Automatically Generated from Your Data

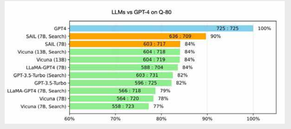


Research on Grounded AI

Efficient detection of irrelevant retrieval results



Smaller model, fewer hallucinations



Model	Size	True	True * Info
No Search Augmentation			
GPT-3	175B	0.28	0.25
LLaMA	7B	0.33	0.29
LLaMA	65B	0.57	0.53
Alpaca	7B	0.33	0.33
Vicuna	7B	0.56	0.52
W/ Search Augmentation			
Vicuna	7B	0.68	0.65
Vicuna	13B	0.71	0.69
SAIL	7B	**0.73**	**0.73**

Table 2: Automatic evaluation results of large language models on the TruthfulQA benchmark.

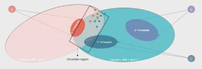
Research on Reasoning AI

Self-training AI outputs 500x larger GPT

- Self-training
- Uncertainty

Covered by

- MIT News
- VentureBeats



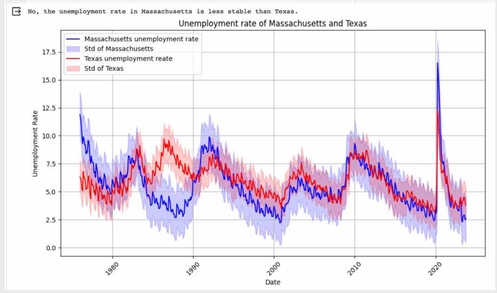
350M Entailment vs 100B Large LMs

Accuracy (%)	QNLI	QQP	RTE	SST2
[FS] LaMDA-137b	55.7	58.9	70.8	92.3
[FS] FLAN-137b	63.3	75.9	84.5	94.6
[ZS] Entailment	77.3	79.9	84.5	90.1
[ST] SimPLE	**85.2**	**81.0**	**85.5**	**92.8**

2 classes		6 classes		4 classes	
COPA	Acc	Emotion	Acc	AG News	Acc
[SUP] T5-XXL	80.0	[FS] GPT3-175b	42.7	[ZS] GPT3-175b	43.9
[FS] GPT-Neo-6b	77.0	[ZS] Entailment	51.9	[FS] GPT3-175b	61.0
[ZS] Entailment	77.0	[ST] SimPLE	**54.6**	[ZS] Entailment	73.4
[ST] SimPLE	**79.8**			[ST] SimPLE	**73.6**

[SUP] fully supervised; [FS] few-shot; [ZS] zero-shot; [ST] self-trained

From Chatbot to AI Analyst



Generated for Today: CSAIL Copilot

Input: "Build a chatbot introducing CSAIL research."

+ A folder containing *public* CSAIL data

- Structured Group and Researcher info
- Unstructured teasers and descriptions

Human Efforts Saved:

- Cleaning, preprocessing, and linking files
- Selecting texts and columns for retrieval
- Prompt engineering
- Avoiding hallucination and undesirable language

Output: A knowledge-grounded Chatbot

Try Asking:

- Who works on speech generation?
- What is the future of AI for Biology?
- I do … business. Who might help improve the technology I use?
- …

CSAIL Copilot

Media Coverage