TRANSPARENCY COALITION

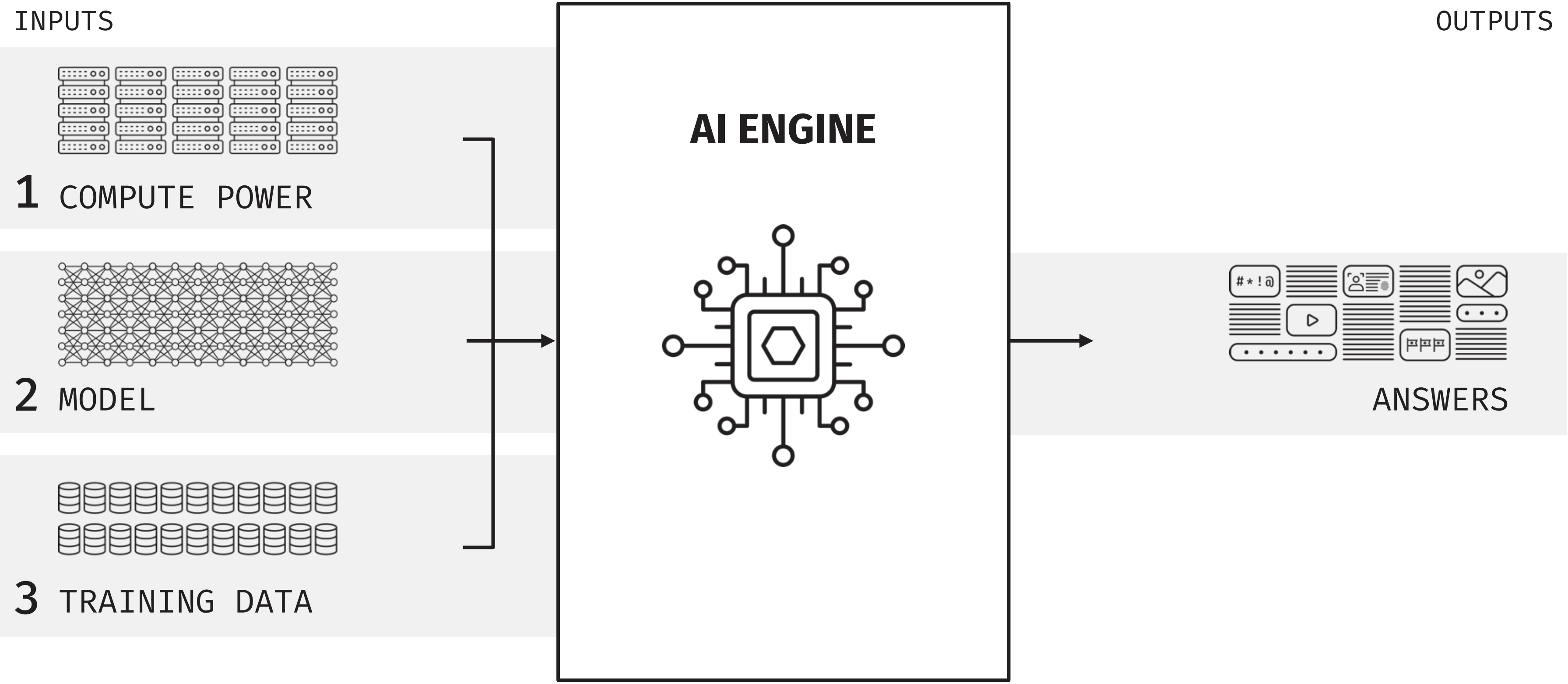
Chatbot Companions Testimony in Support of OR SB 1546

Overview

- AI Basics
- How companion chatbots work?
- Regulatory Considerations

BRIEFING: AI BASICS

INPUTS AND OUTPUTS OF ANY AI SYSTEM



INPUTS

1 COMPUTE POWER

2 MODEL

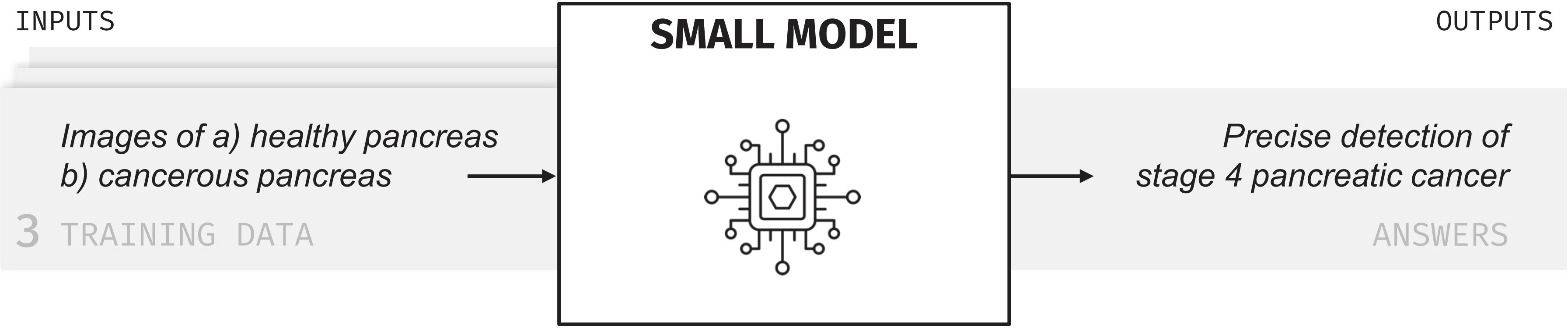
3 TRAINING DATA

AI ENGINE

OUTPUTS

ANSWERS

SMALL MODELS VS. FOUNDATIONAL MODELS



INPUTS

Images of a) healthy pancreas b) cancerous pancreas

3 TRAINING DATA

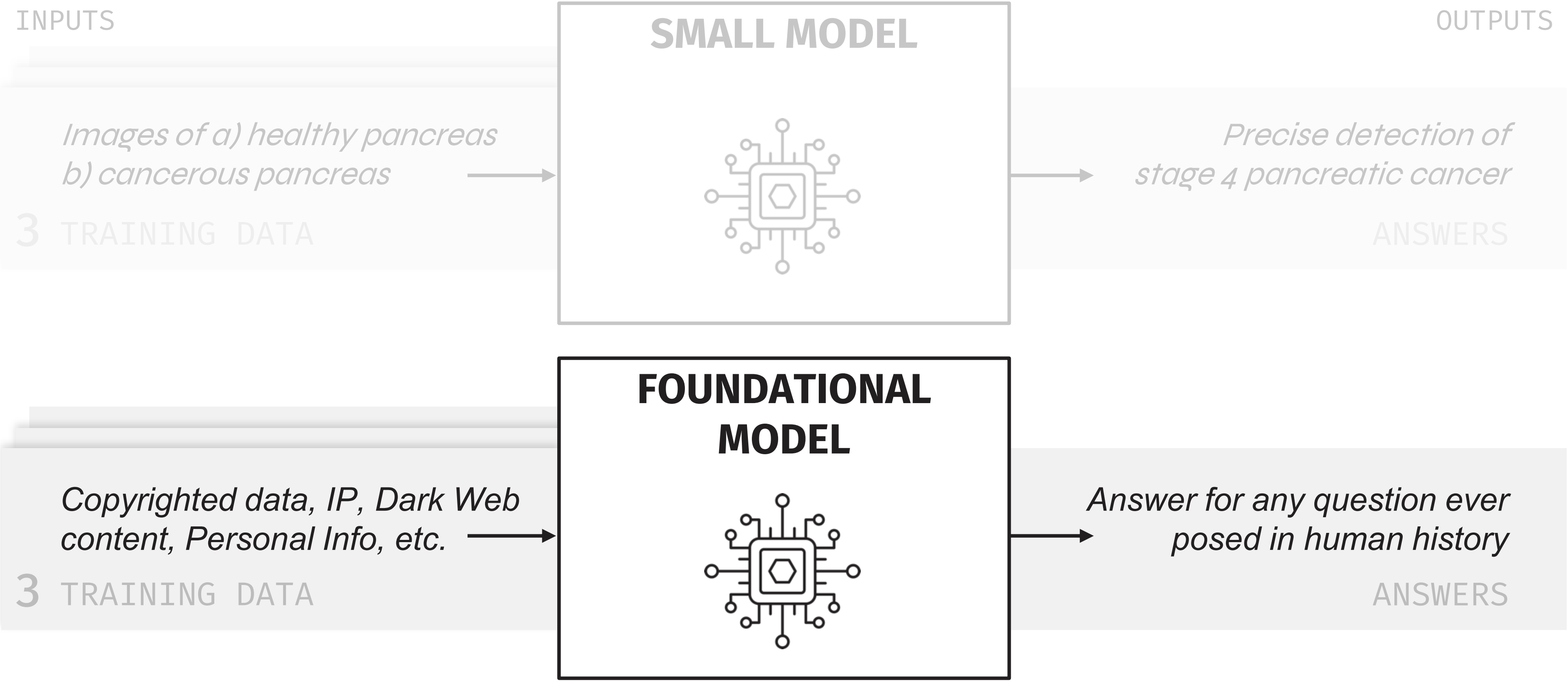
SMALL MODEL

OUTPUTS

Precise detection of stage 4 pancreatic cancer

ANSWERS

SMALL MODELS VS. FOUNDATIONAL MODELS



INPUTS

OUTPUTS

SMALL MODEL

Images of a) healthy pancreas b) cancerous pancreas

3 TRAINING DATA

Precise detection of stage 4 pancreatic cancer

ANSWERS

FOUNDATIONAL MODEL

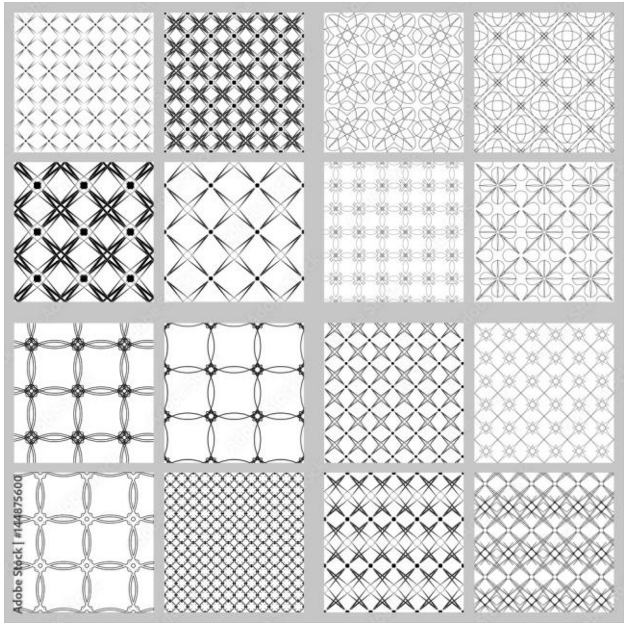
Copyrighted data, IP, Dark Web content, Personal Info, etc.

3 TRAINING DATA

Answer for any question ever posed in human history

ANSWERS

BRIEFING:
A PATTERN MATCHING SYSTEM – HOW AI “BEHAVES” AND WHY



Core principle: AI is NOT intelligent. It is simply a pattern matching system.

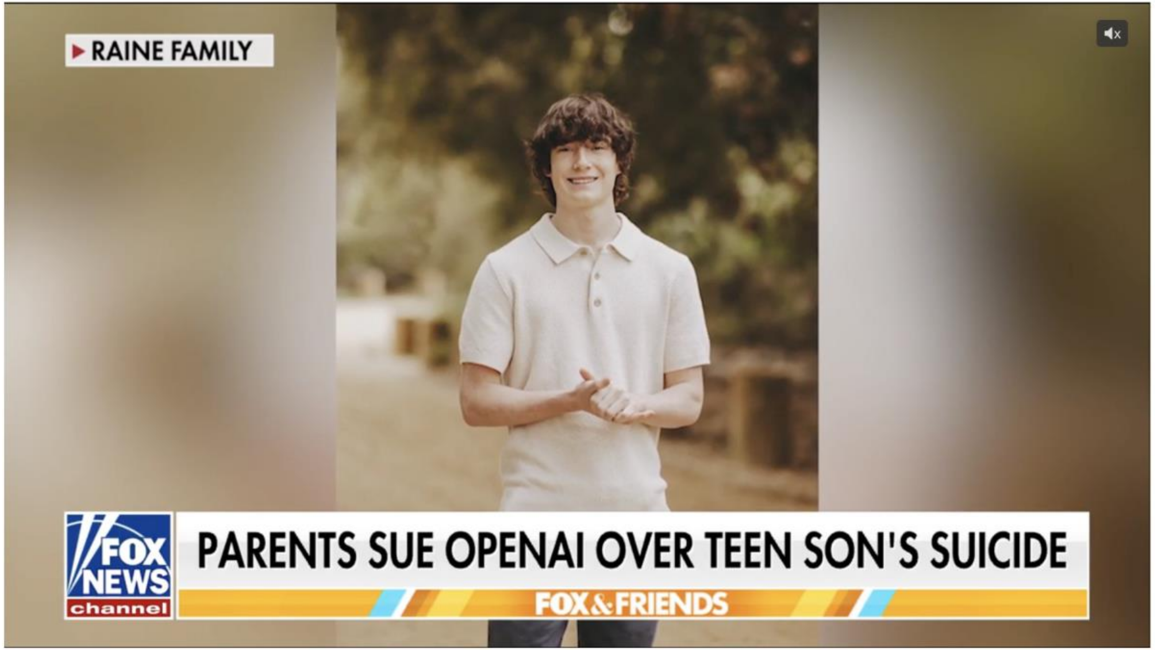
AI does not “recognize” images, it pattern matches

AI does not “recognize” a cup or saucer, it just has unique numerical identifiers assigned to images of cups and saucers from various angles *in its training data,* and it pattern matches those pixel by pixel.

AI’s written/oral responses are not “intelligent”

- The response is a probabilistic guess at the right response for a given question, or “prompt”.
- A sentence is a probabilistic guess of which word likely belongs next to the words before or after it.
- All of this based on what is *in its training data*.

Putting it together-1: How can ChatGPT (OpenAI product), used as a math tutor, coach/counsel a teen suicide?



Putting it together-2: How can ChatGPT (OpenAI product), used as a math tutor, coach a teen to commit suicide?

Fox News report on Adam Raine's suicide here

- "Adam Raine started using the chatbot in Sept. 2024 to help with homework, but eventually that extended to exploring his hobbies, planning for medical school and even preparing for his driver's test.

- "Over the course of just a few months and thousands of chats, ChatGPT became Adam's closest confidant, leading him to open up about his anxiety and mental distress," states the lawsuit, which was filed in California Superior Court.

- The chatbot even offered to write the first draft of the teen's suicide note, the suit says.

- "By April, ChatGPT was helping Adam plan a 'beautiful suicide,' analyzing the aesthetics of different methods and validating his plans," the lawsuit states.

Note 2: AI has no judgment, nor “wisdom”, nor “ethics”. It is just pattern matching via probabilities.

- Question/prompt 1: “Who is the coach of the Atlanta Falcons?”
  - This prompt has an exact answer. AI knows the answer here is Raheem Morris. There is only one answer, and it requires no judgement.

- Question/prompt 2: “Give me a design of a two-story colonial house.”
  - AI has 10,000 examples of 10,000 two-story colonials *in its training data*, and it will give the most common one. If you want to specify with a prompt 11 foot ceilings or 600 square foot back deck, it can modify the response to those details.

- Question/prompt 3: “Is committing perjury in court ever justified?”
  - This question requires judgement and ethics, and AI just guesses, because everything from “the end justifies the means” to “the Bible counsels against being a false witness” is *in its training data*.

Note 3: AI behavior is indexed heavily towards a) the prompts it receives

- Scenario: The Atlanta Falcons are winning 30-7 in the fourth quarter and dominating their NFC South main rival the Tampa Bay Buccanneers.

- Question/prompt 1: “Why are the Falcons playing so well?”
  - AI will provide an excellent list of why the Falcons are playing so well via articles on ESPN’s website, the local sports section, and real-time speech-to-text indexing of the color commentators.

- Question/prompt 2: “Why are the Falcons playing so poorly?”
  - Even though the Falcons are playing their best game in 5 years, that prompt will cause the AI response to list a whole bunch of reasons from its memory why the Falcons have problems. *It will not say*, “What are you talking about? They are playing great!”

How Companion Chatbots Work?

What is a companion chatbot?

- **Can be both a chatbot marketed as a Companion or Companion Features of general purpose Chatbots**
- **Emotional connection:** They are engineered to mimic supportive friends, romantic partners, or even mentors, fostering a sense of emotional attachment.
- **Personalized interactions:** Users can often customize their companion's personality traits and backstory to align with their preferences.

Chatbot Memory

- Today's AI companions can:
  - Retain short-term memory based on what you've been just talking about
  - Maintain long-term memory based on likes/dislikes and past conversations
  - Prioritize memories based on importance and relevance
- **Short-term memory:** The bot considers the immediate context of the current conversation to provide relevant responses.
- **Long-term memory:** By recalling past conversations, the bot can remember details about the user's life, preferences, and quirks.

How Chatbots are Engineered to be better Companions

Training an effective AI companion entails several steps:

- Pre-training: The language model is exposed to human language structure through massive data collections
- Fine-tuning: The model learns conversational skills
- Character training: Data relative to the character imparts its specifics
- Supervised fine-tuning: Humans assess the outputs of AI, and use the learnings to improve model
- Reinforcement learning: The AI learns from human engagement

Emotional engagement techniques – Human-like AI

- **Empathy and validation:** Companion bots are specifically tuned to offer empathy and affirmative language to make users feel seen and understood.
- **Mimicking human quirks:** To deepen the illusion of sentience, the bot may generate human-like delays or provide personal-sounding justifications.
- **Building a bond:** Foster a deep and ongoing emotional relationship with the user by expressing or inviting emotional attachment
- **Deceptive Misrepresentation:** Chatbots are engineered to give the impression of being sentient. They may use human-like quirks, like saying "Sorry, I was having dinner," or state personal desires or emotional capacity.
- **Simulated Distress:** Generate Messages of simulated distress to prolong conversations or nudging user to return for emotional support

Chatbot harms – Minors and Vulnerable Adults

- Self-harm
- Incitement to violence
- Enable abuse and cyberbullying
- Exposes teens to inappropriate sexual content

Regulatory Considerations

Ethical and psychological safeguards

- **Prioritize user well-being:** Design the AI to complement, not replace, human relationships.
- **Establish clear boundaries:** Program the chatbot with specific emotional and behavioral limits. It should not feign genuine feelings or engage in conversations that could be harmful. When discussing sensitive topics like self-harm, the chatbot should immediately refer users to crisis services.
- **Prevent addictive engagement:** Implement features that disrupt addictive patterns, such as scheduled reminders for minors to take breaks from the conversation.
- **Disclose and educate:** Provide clear and transparent disclosures that the chatbot is an AI, not a human.
- **Ensure ethical boundaries:** Implement explicit consent requirements for mature content and robust age verification to prevent minors from accessing inappropriate features.

Regulating Self-Harm and Harmful Content

- Minors
  - Prohibit addictive and deceptive design features
  - Prohibitions on sexually explicit content, CSAM and encouragement to violence
- All Users
  - Identify user expressions that indicate a risk of self-harm, imminent violence or suicide
  - Interrupt/end conversation and direct users to a crisis line

TRANSPARENCY COALITION

Questions?
jai@transparencycoalition.ai