

AI LUNCH AND LEARN SERIES

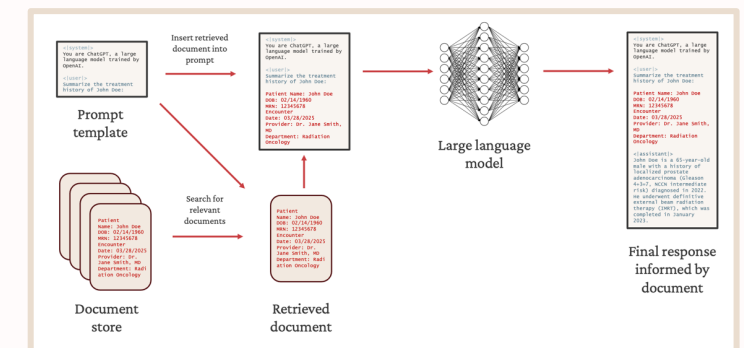
Reading Clinical Notes with AI: *Building an Effective System*

ANDREW Y. K. FOONG, PH.D.

September 17th 2025



Radiation
Oncology
AI & Data Analytics
AIDA



AI and Data Analytics team & collaborators

Work done by and with the Radiation Oncology AIDA team:



MARK WADDLE
M.D.



SATOMI SHIRAISHI
PH.D.



ANDREW FOONG
PH.D.



DAVID ROUTMAN
M.D.



ADAM
AMUNDSON



SRINIVAS
SEETAMSETTY



JASMINE
ZHOU



FEDERICO MASTROLEO
M.D.



MARIANA BORRAS-OSORIO
PH.D.



JASON HOLMES
PH.D.

About me



Senior Associate Consultant · Radiation Oncology



Senior Researcher · Microsoft Research



Ph.D. in Machine Learning · Cambridge University



Research Scientist Intern · Google DeepMind

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

Roadmap

1. **Why AI for Clinical Notes?**
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

WHY AI FOR CLINICAL NOTES?

Dark data

DARK DATA:

- Contained in the EHR in free-text clinical notes.
- Is **not structured**, so cannot:
 - Search efficiently
 - Query via SQL
 - Perform even simple statistical analyses

Analogy with DARK MATTER:

- *We know it's there*—we can detect it.
- *We know it's massive*—there's more of it than visible matter.
- *We can't analyze it any detail*—largely an enigma.



Example use-cases

ARTIFICIAL EXAMPLE:

pt name stevens john dob 07/14/54 mrn 00349821 consult
dr mallory urol bx prostate 12 cores both sides. psa
reported 10.7 ng/ml (lab 3/10/17 per pt chart maybe not
verified). gross - fragments mostly tan some crushed.
right apex tiny piece only. left mid looks ok tissue.

micro - right apex suspicious glands, adenoca likely,
gleason 3+3=6, about 20% of core? right mid pos
carcinoma gs 3+4=7 35% of tissue, right base ?artifact
hard to tell maybe carcinoma not def, left apex benign,
left mid adenoca glands fused pattern gs 4+3=7 high %
(approx 70). left base bad sample large cribriform
glands gleason 4+4=8 extensive tumor. other cores
chronic inflammation no carcinoma.

note multifocal disease. highest gs is 8. risk category
prob high. correlation w clinical findings advised. some
cores v fragmented, gross description maybe wrong for
right lat base, recheck if necessary. pt has palpable
nodule per referring note. path sig l k patel md 3/24/17
addendum later 3/27 confirmed gs 8 left base rest
unchanged. insurance code?? c61 likely. pls verify.

PROSTATE BIOPSY PATHOLOGY REPORT

- Tens of thousands of such reports
- Variable writing style
- Questions we might want to ask:
 - *What's the distribution of Gleason scores?*
 - *How has it changed over time?*
 - *What is it associated with?*

Example use-cases



Example use-cases

UROLOGY



Typical patient journey generates 100+ notes.

PAT

Questions we might ask:

- *How many patients had a recurrence?*
- *What risk factors are associated with recurrence?*

RAI
ON

M

ONCOLOGY



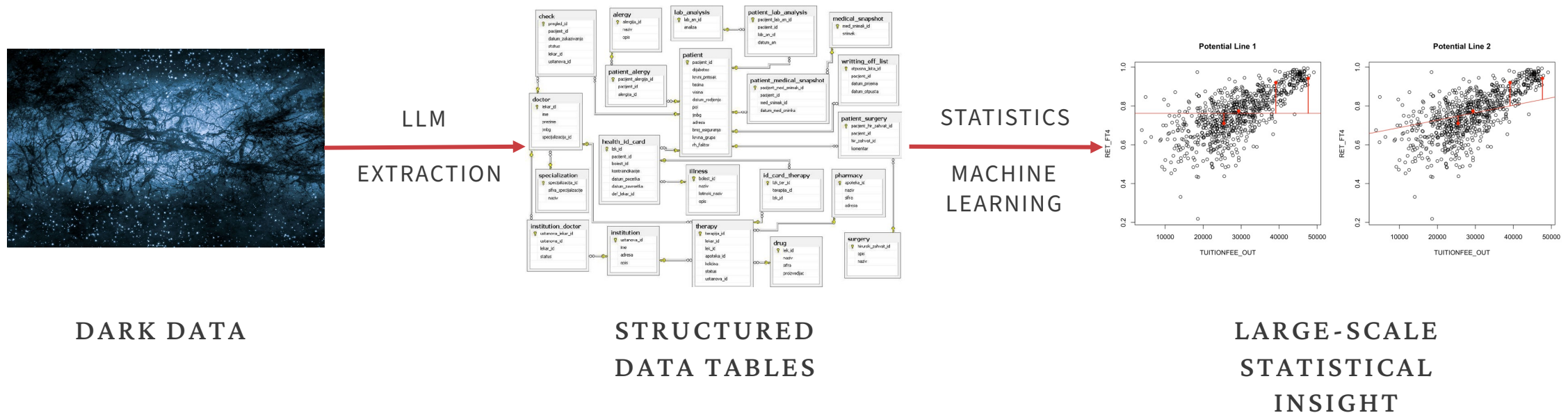
PATIENT CARE JOURNEY

The LLM opportunity

PREVIOUSLY: to answer these population-level questions:

Manual, time-consuming, mind-numbing chart review of thousands of notes.

NOW: LLMs can automate this to a high degree of accuracy.



Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

Roadmap

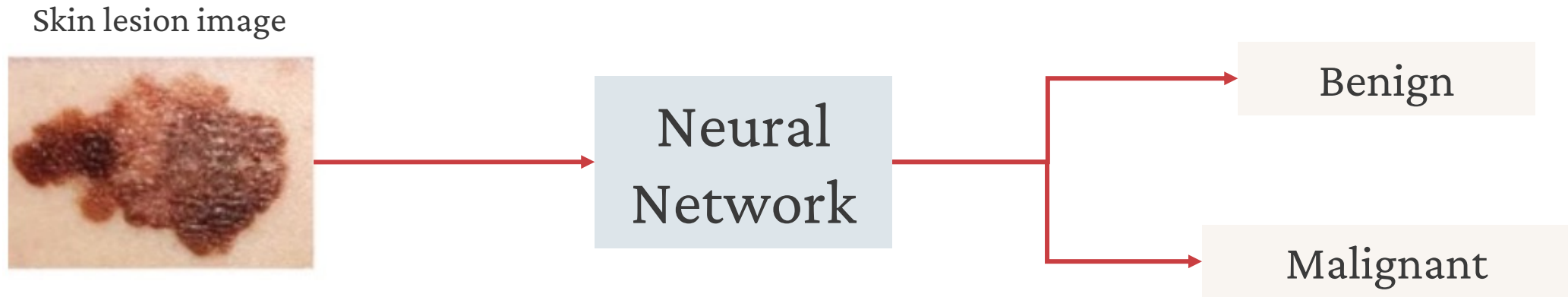
1. Why AI for Clinical Notes?
2. **Large Language Models 101**
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

LARGE LANGUAGE MODELS 101

Generative AI is predictive AI

- How does an LLM (ChatGPT, Gemini) generate bodies of coherent text?
- Idea: text generation is just *repeated word prediction*.
- “Classical AI”: **predict** malignancy given image.

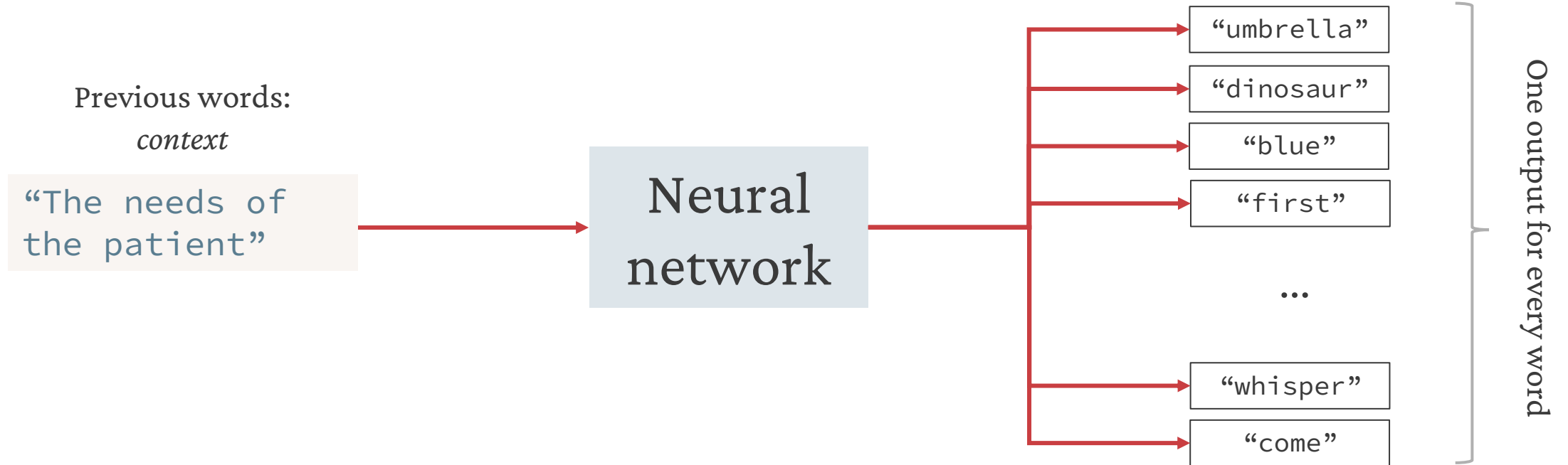
EXAMPLE:



- *Generative AI*: **predict** next word given previous words.

Generative AI is predictive AI

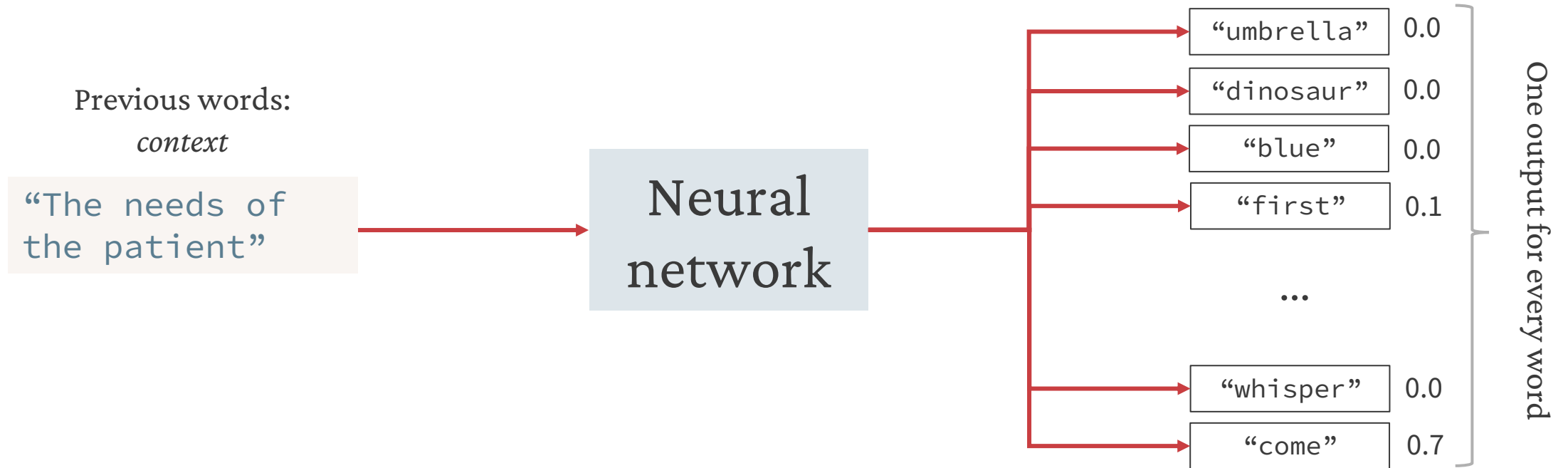
EXAMPLE:



1. Predict most probable next word.

Generative AI is predictive AI

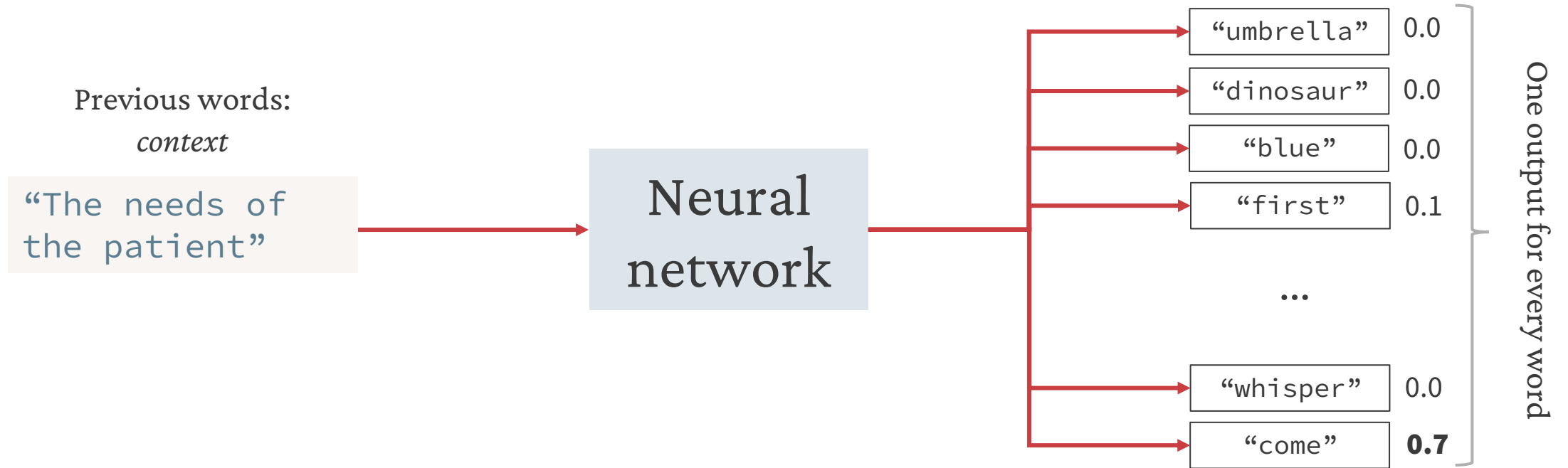
EXAMPLE:



1. Predict most probable next word.

Generative AI is predictive AI

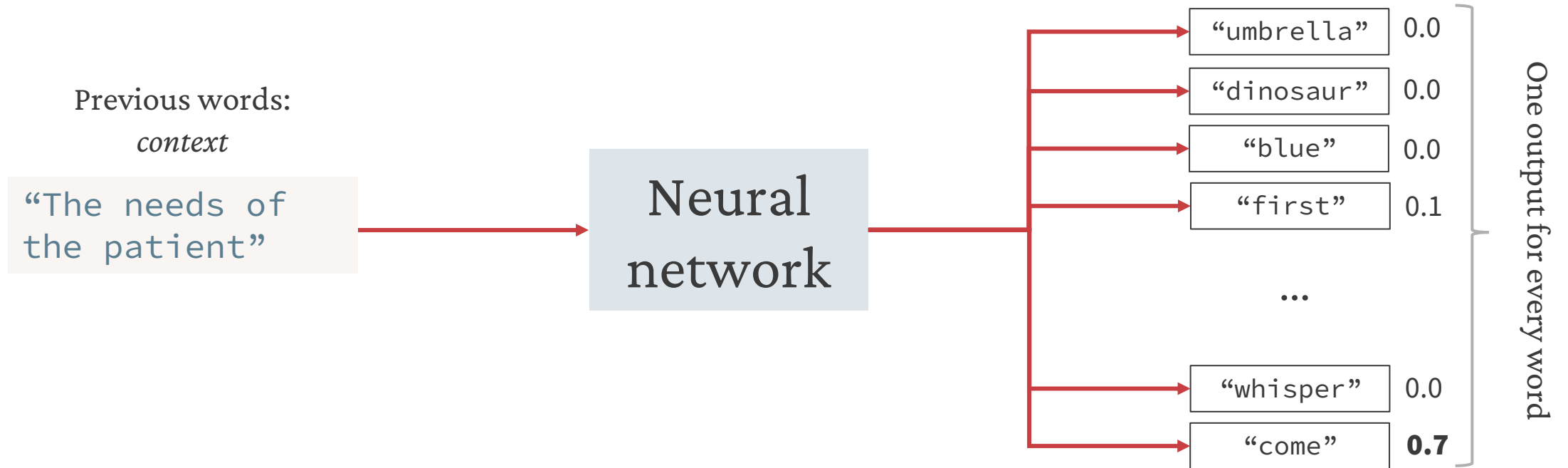
EXAMPLE:



1. Predict most probable next word.

Generative AI is predictive AI

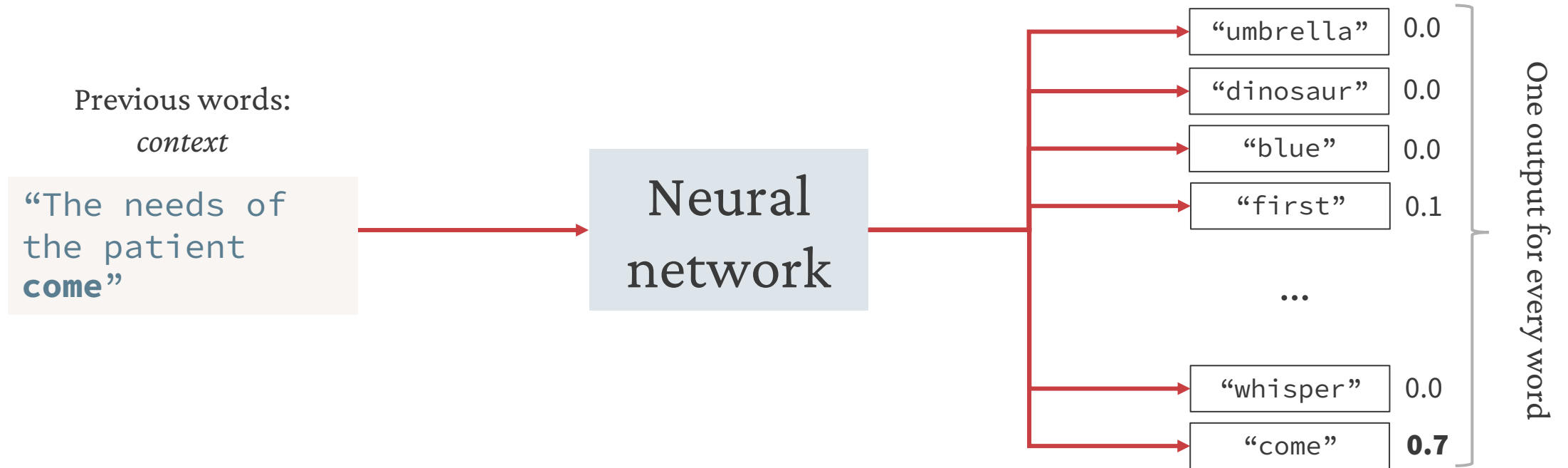
EXAMPLE:



1. Predict most probable next word.
2. Add it onto the context.

Generative AI is predictive AI

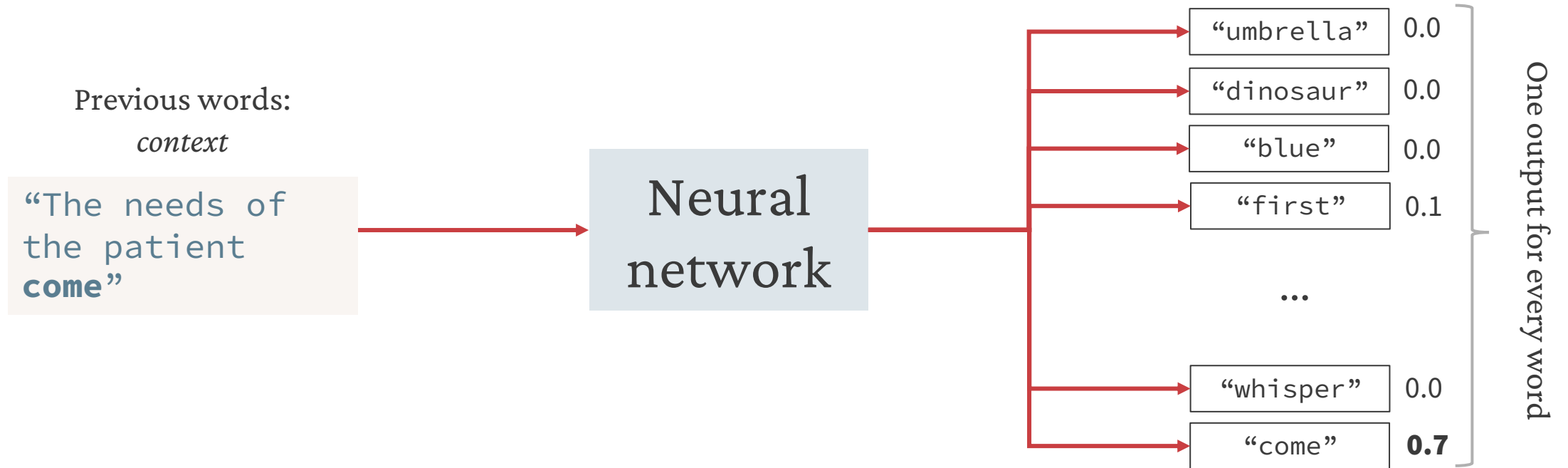
EXAMPLE:



1. Predict most probable next word.
2. Add it onto the context.

Generative AI is predictive AI

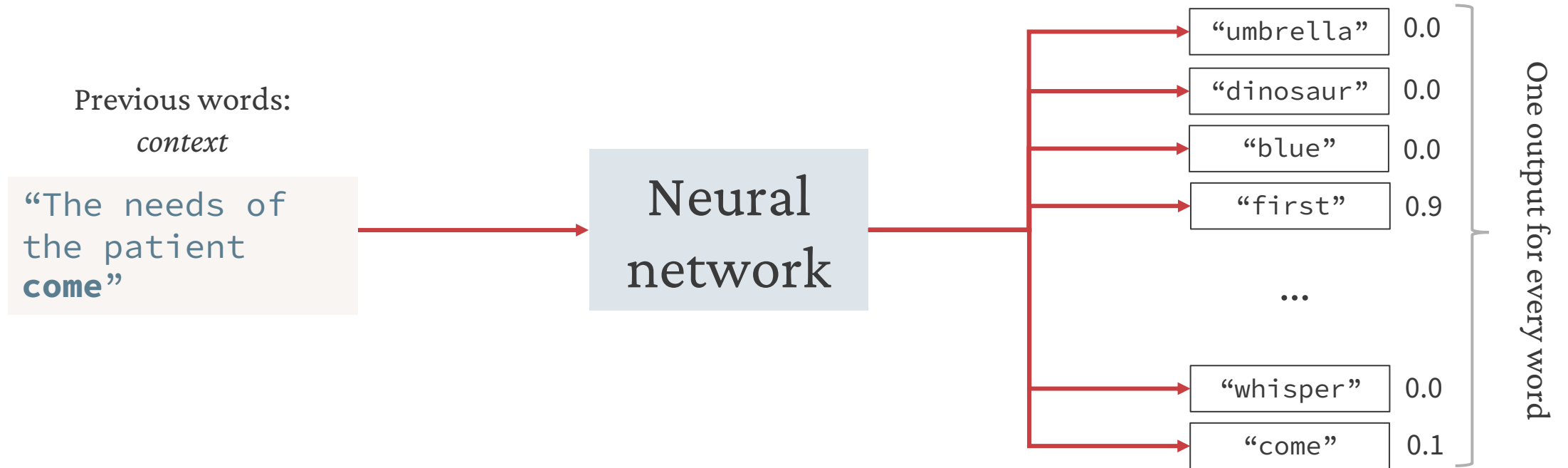
EXAMPLE:



1. Predict most probable next word.
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI

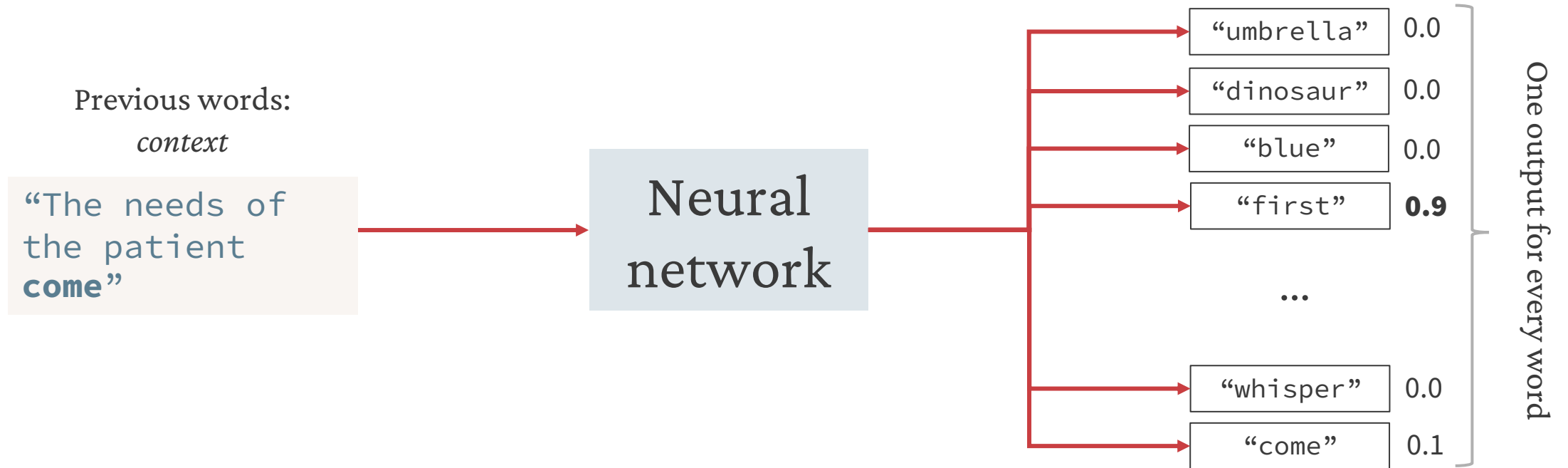
EXAMPLE:



1. Predict most probable next word.
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI

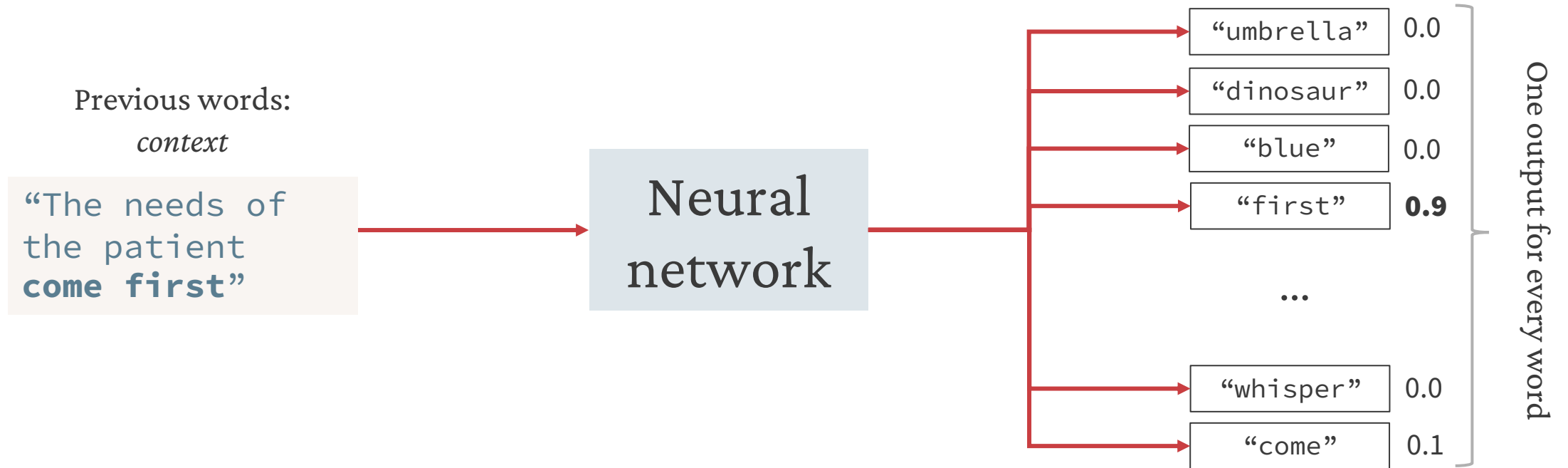
EXAMPLE:



1. Predict most probable next word.
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI

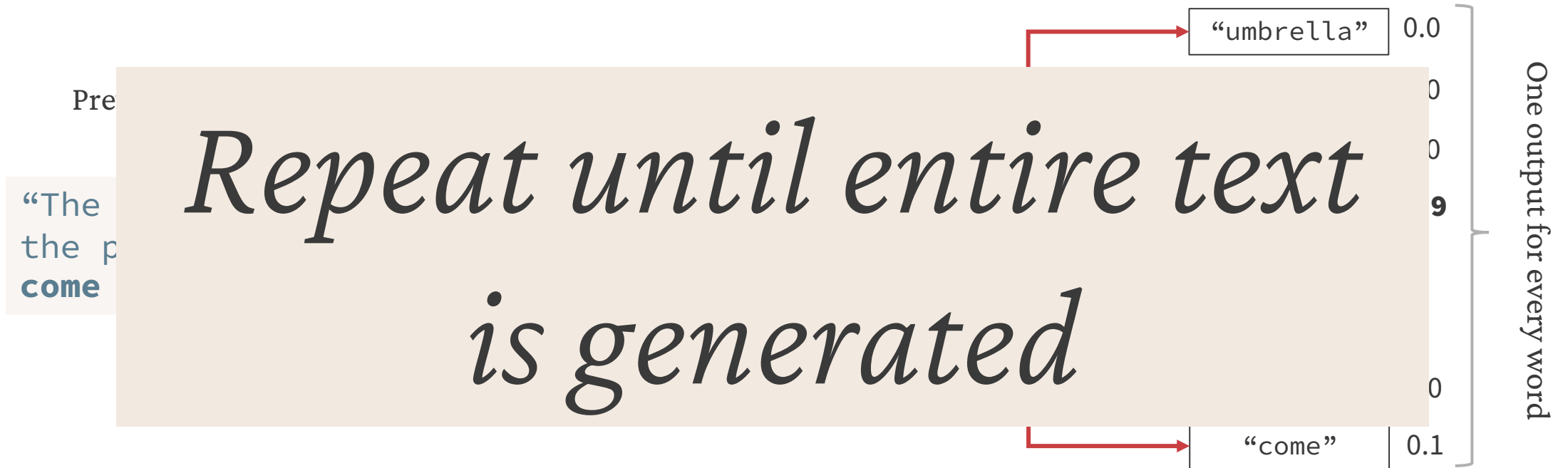
EXAMPLE:



1. Predict most probable next word.
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI

EXAMPLE:

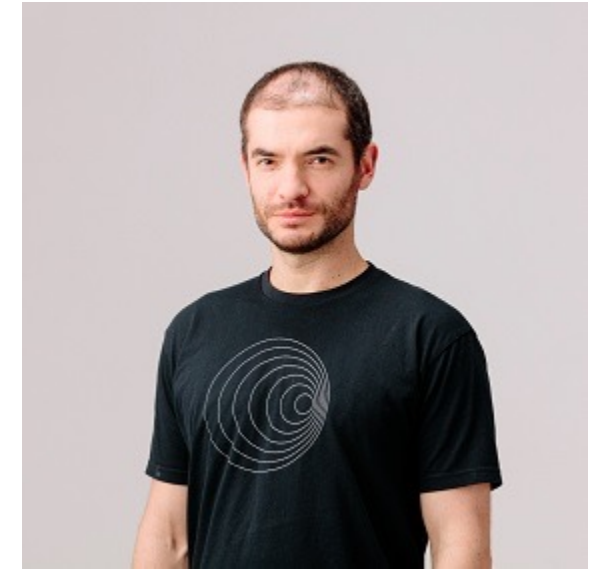


1. Predict most probable next word.
2. Add it onto the context.
3. Go back to step 1.

Is next-word prediction enough to understand?

- Memorization \neq understanding.
- BUT:
Next-word prediction + *Generalization* \rightarrow Understanding
- Thought experiment:
 - Imagine LLM trained on detective novels.
 - Novel ends with “and the killer was...”
 - To accurately predict next word, LLM must understand the whole novel.
- *If* LLM can predict next word in novel situations, it can understand!
- LLM revolution:

*We **can** train excellent next-word predictors.*



ILYA SUTSKEVER, FRS
Former Chief Scientist
OpenAI

Anatomy of a Conversation

- What actually happens when you *have a conversation with* ChatGPT/Gemini?

Tell me about radiation oncology.

Your computer



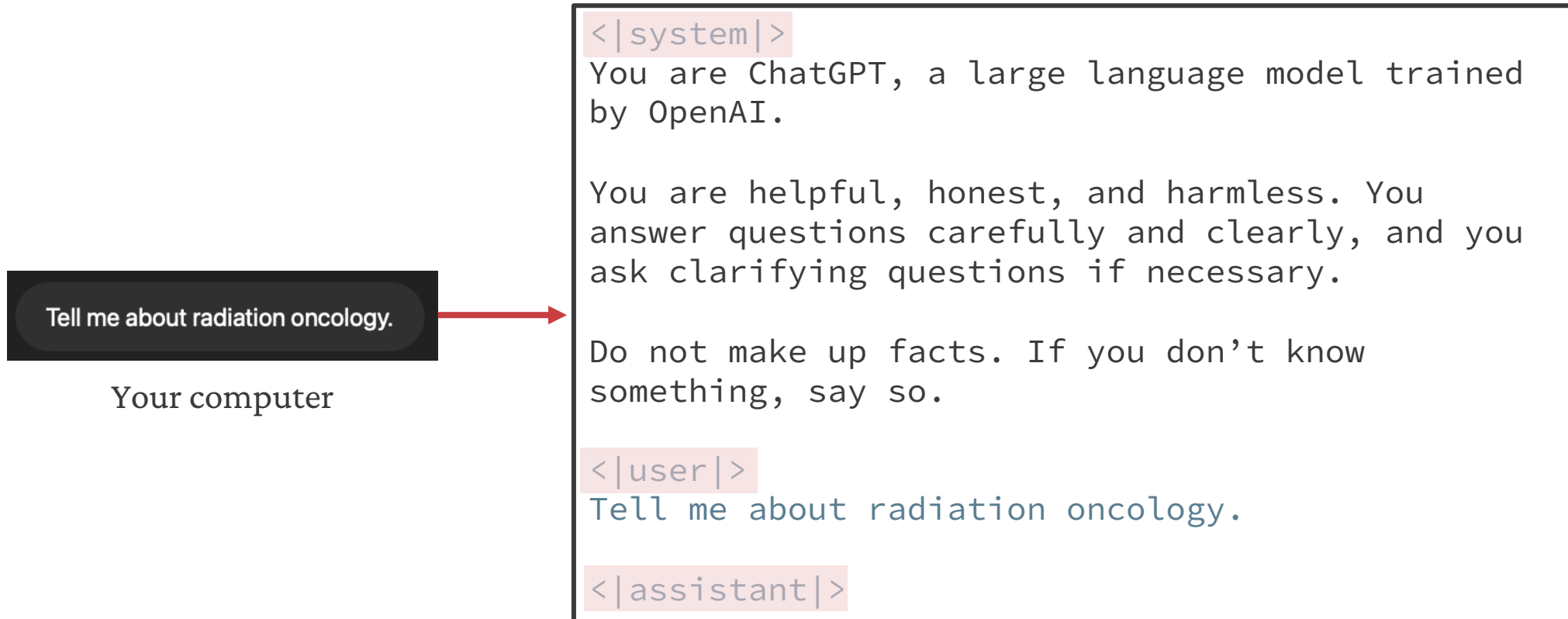
Data center with
thousands of GPUs

- Neural network too large to run on personal devices.
- Parameters are valuable IP!
- Chat message sent to data center.

Anatomy of a Conversation

- Message embedded in a template with special format/symbols:

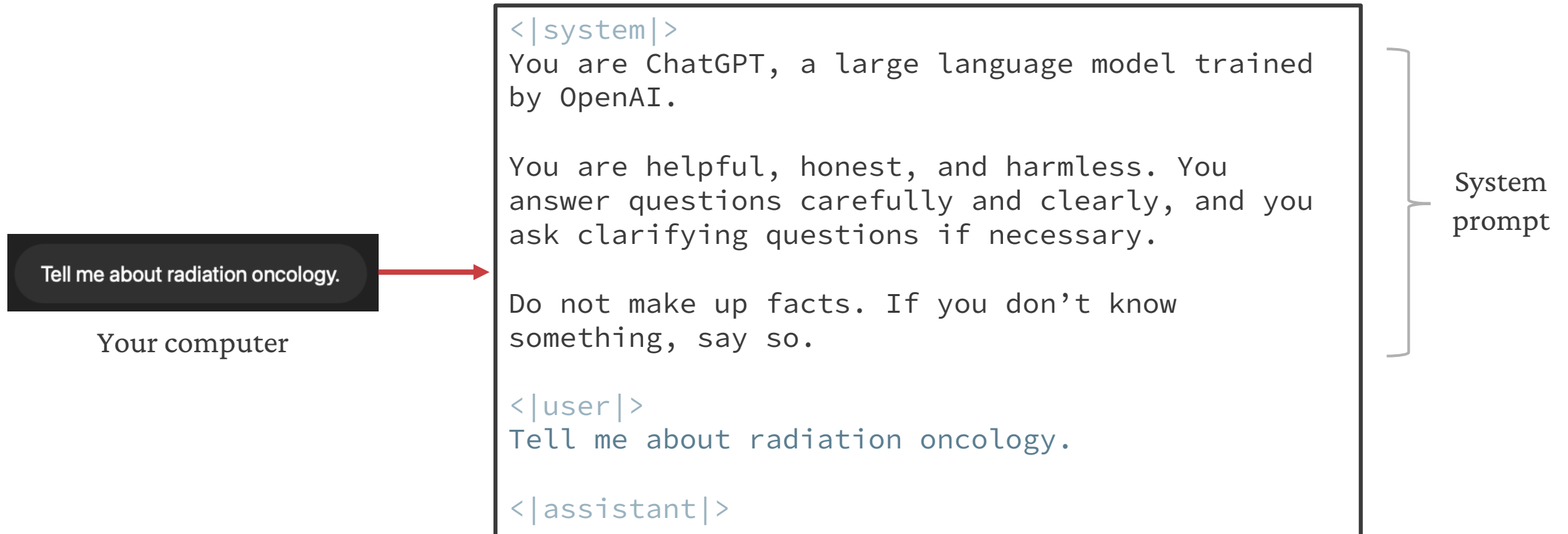
EXAMPLE:



Anatomy of a Conversation

- Message embedded in a template with special format/symbols:

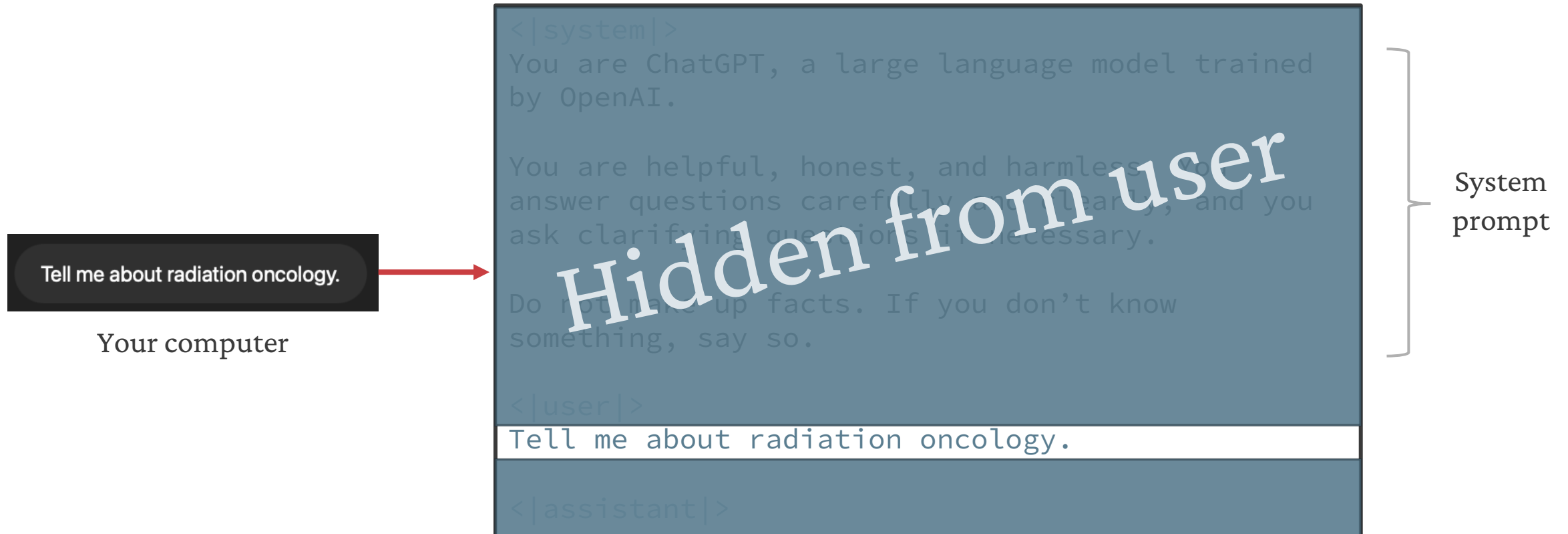
EXAMPLE:



Anatomy of a Conversation

- Message embedded in a template with special format/symbols:

EXAMPLE:



Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

```
<|system|>
```

You are ChatGPT, a large language model trained by OpenAI.

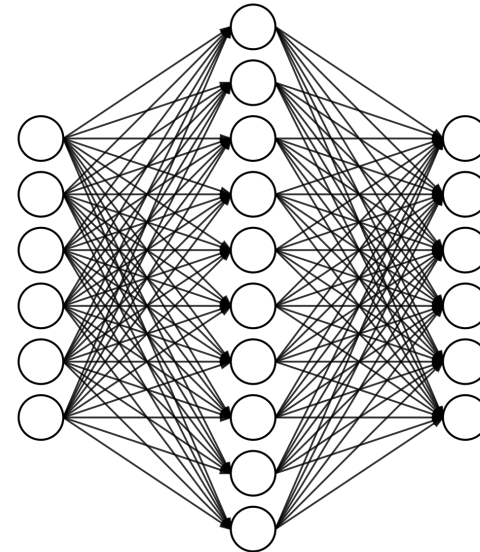
You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

```
<|user|>
```

Tell me about radiation oncology.

```
<|assistant|>
```



“Radiation”

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

```
<|system|>
```

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

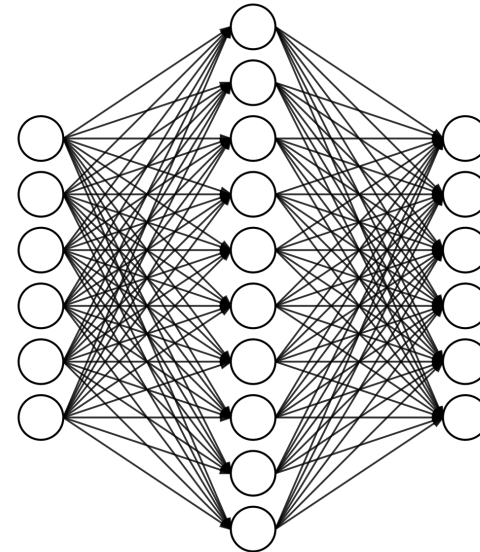
Do not make up facts. If you don't know something, say so.

```
<|user|>
```

Tell me about radiation oncology.

```
<|assistant|>
```

Radiation



“Radiation”

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

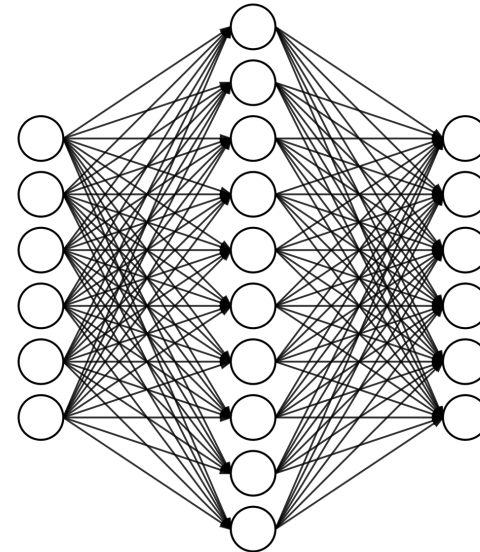
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation



`"oncology"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

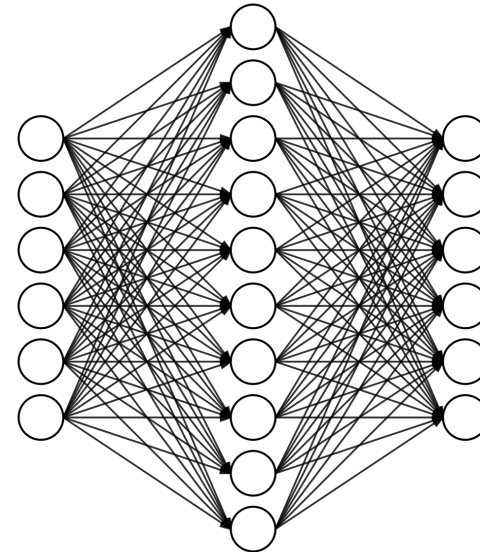
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology



`"oncology"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

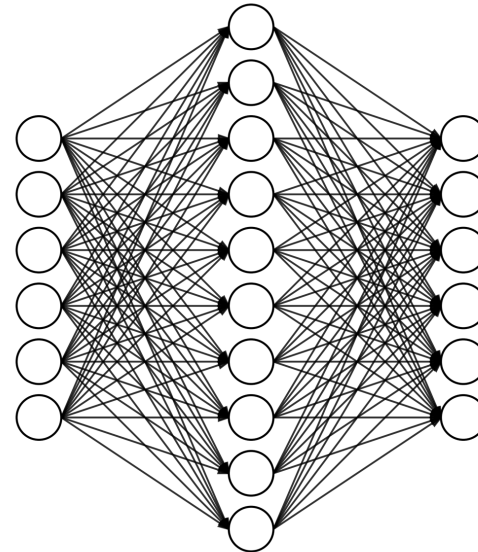
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology



`"is"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

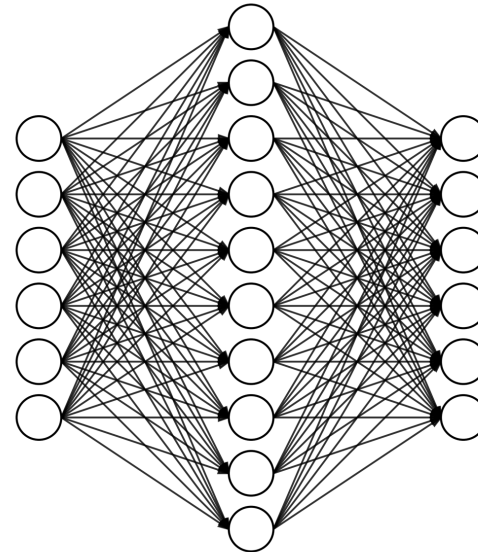
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is



`"is"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

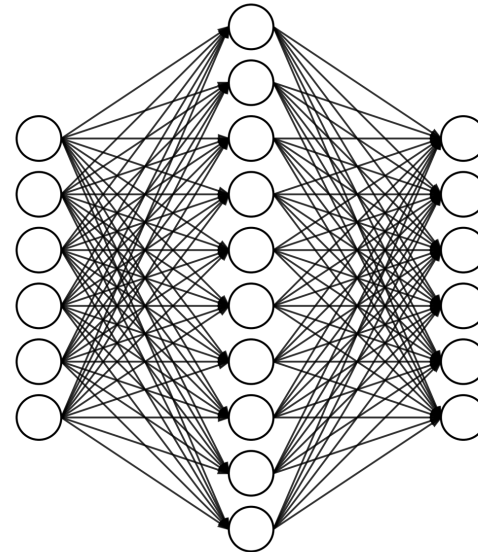
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is



`"a"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

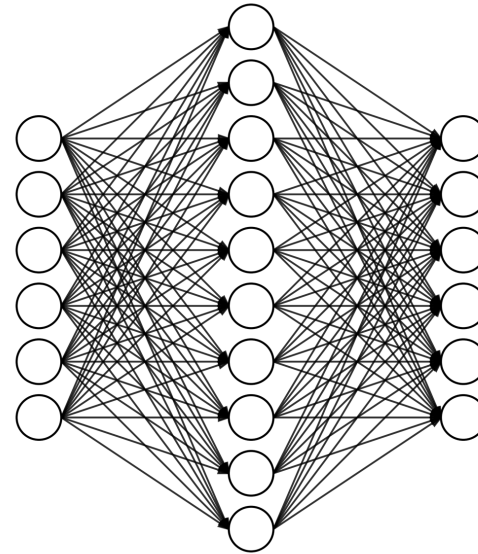
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is a



`"a"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

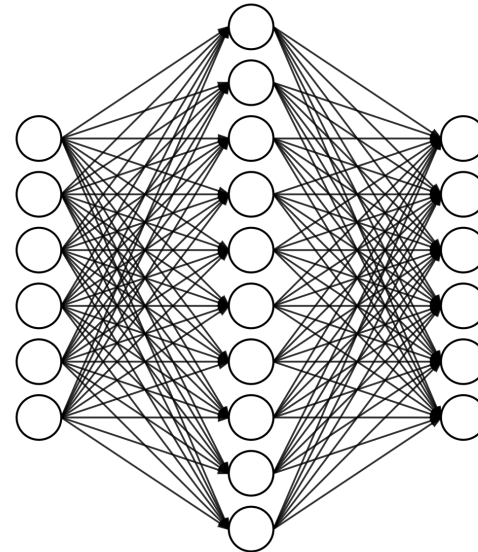
Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is a



`"medical"`

Anatomy of a Conversation

- Formatted message sent to neural network to predict next word.
- Format designed so *prediction implies assistance*.

EXAMPLE:

```
<|system|>
```

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

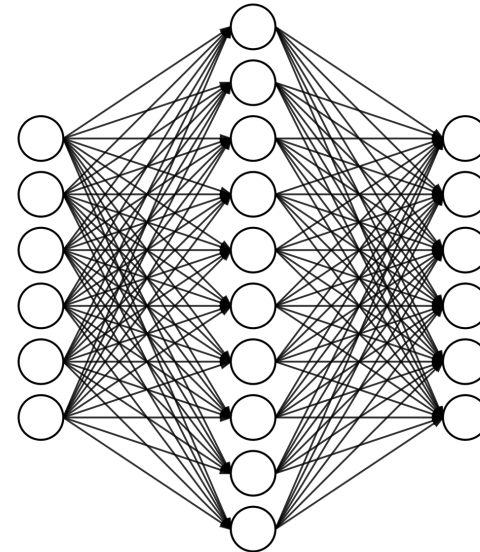
Do not make up facts. If you don't know something, say so.

```
<|user|>
```

Tell me about radiation oncology.

```
<|assistant|>
```

Radiation oncology is a medical



“medical”

Anatomy of a Conversation

- Full message extracted and sent back to your computer.

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

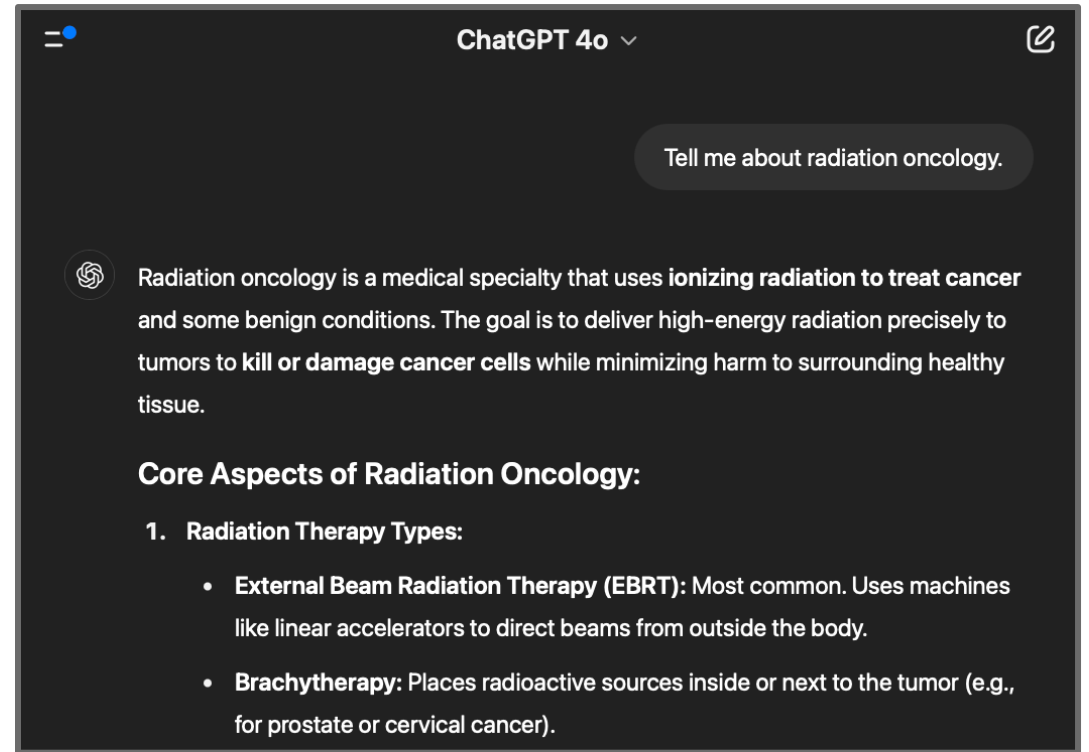
Do not make up facts. If you don't know something, say so.

`<|user|>`

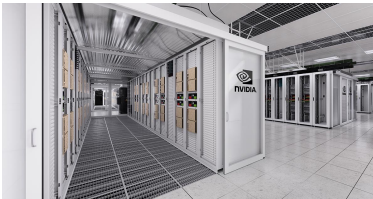
Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

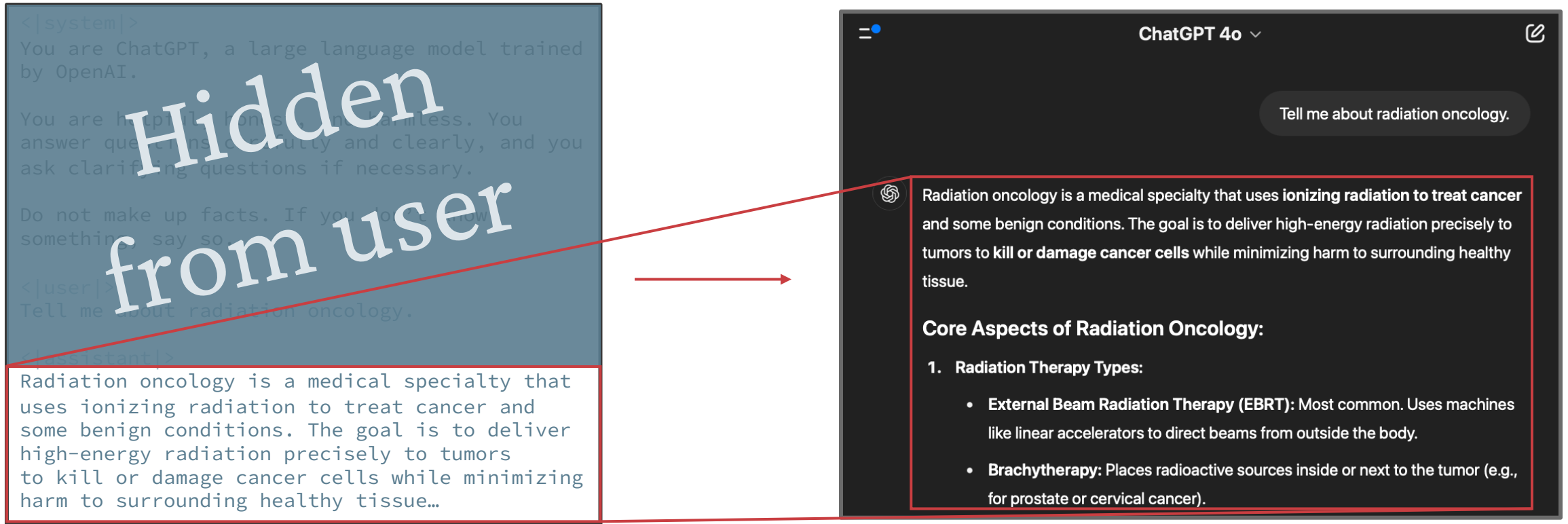


Your computer



Anatomy of a Conversation

- Full message extracted and sent back to your computer.



Your computer

Multi-turn conversations

What happens when you reply?

EXAMPLE:

<|system|>

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

<|user|>

Tell me about radiation oncology.

<|assistant|>

Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

<|user|>

Great! Now tell me the benefits of proton therapy.

Multi-turn conversations

What happens when you reply?

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.



`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

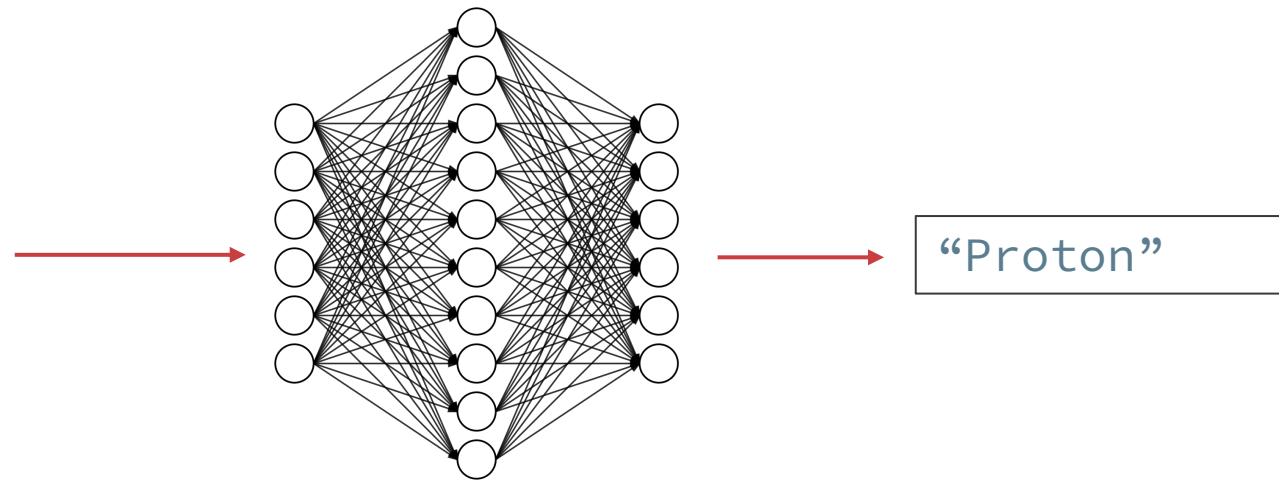
`<|assistant|>`

Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`



Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

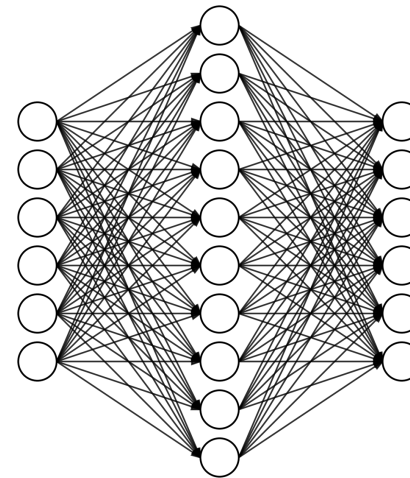
`<|assistant|>`

Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`



“Proton”

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

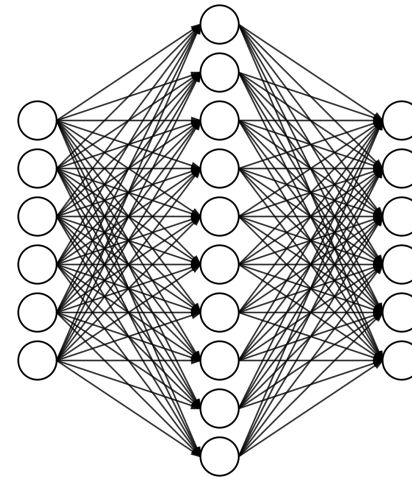
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton



“Proton”

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

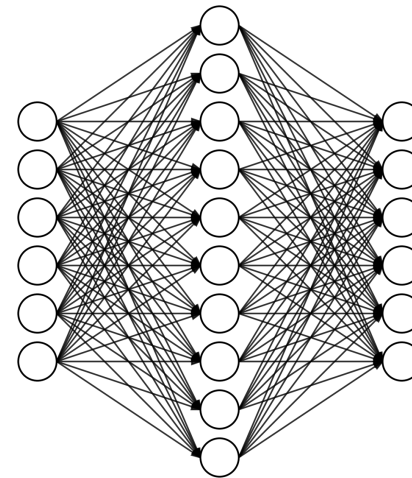
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton



“therapy”

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

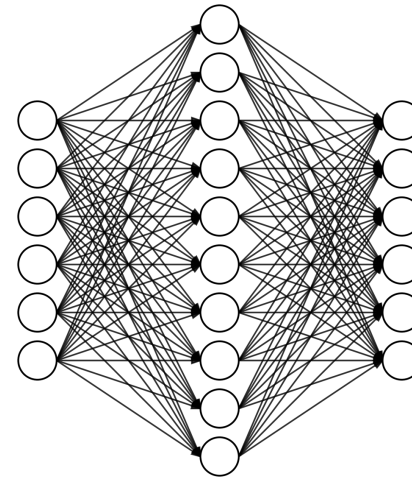
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton therapy



“therapy”

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

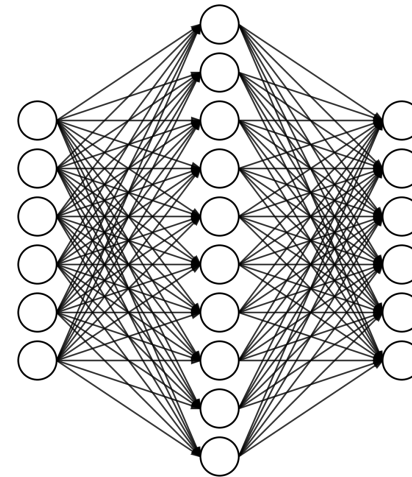
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton therapy



`"has"`

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

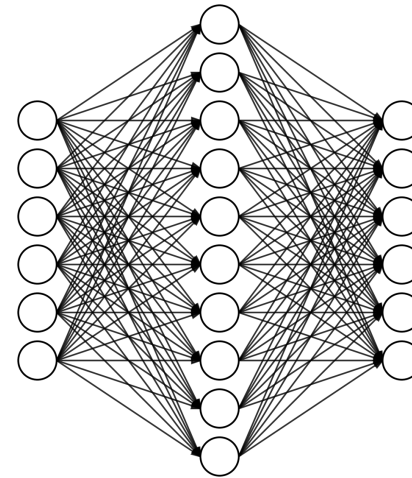
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton therapy has



“therapy”

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

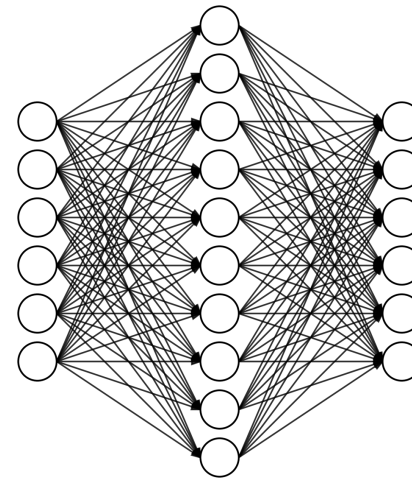
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton therapy has



“several”

Multi-turn conversations

Feed back into neural network in same way.

EXAMPLE:

`<|system|>`

You are ChatGPT, a large language model trained by OpenAI.

You are helpful, honest, and harmless. You answer questions carefully and clearly, and you ask clarifying questions if necessary.

Do not make up facts. If you don't know something, say so.

`<|user|>`

Tell me about radiation oncology.

`<|assistant|>`

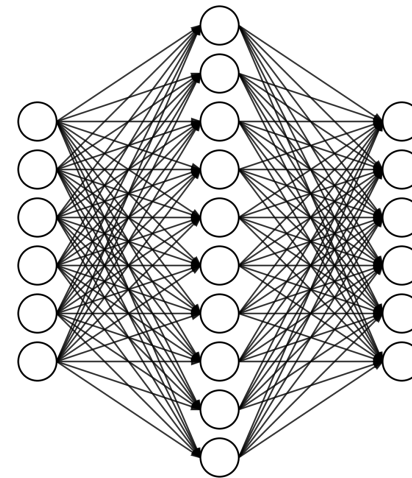
Radiation oncology is a medical specialty that uses ionizing radiation to treat cancer and some benign conditions. The goal is to deliver high-energy radiation precisely to tumors to kill or damage cancer cells while minimizing harm to surrounding healthy tissue...

`<|user|>`

Great! Now tell me the benefits of proton therapy.

`<|assistant|>`

Proton therapy has several



`“several”`

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. **Supervised Learning**
4. Prompting LLMs
5. The Needle in a Haystack

SUPERVISED LEARNING

Making things systematic

- **LEVEL ZERO:**

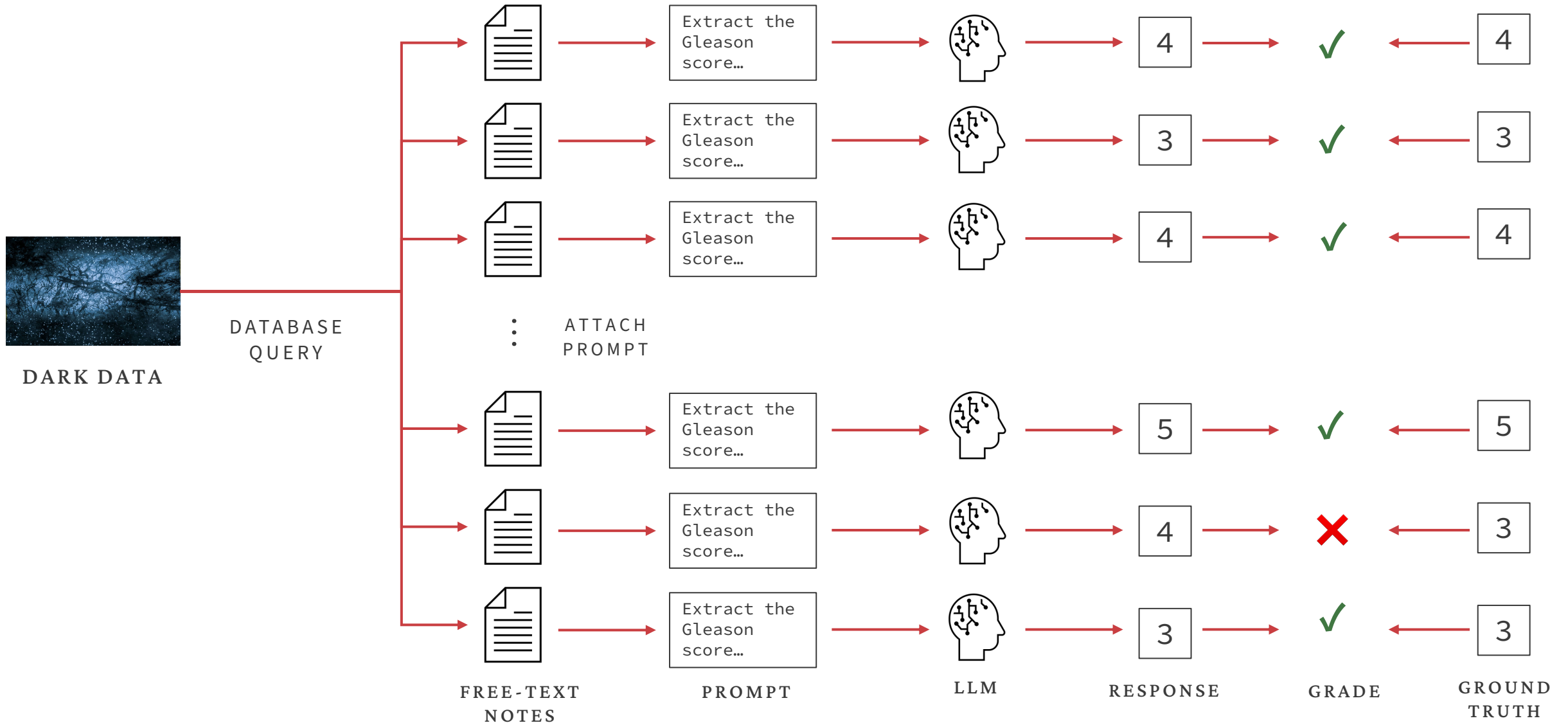
1. Copy clinical notes into ChatGPT/Gemini (PHI-safe version).
2. Ask: “*Please extract the Gleason Score,*” or whatever comes to mind.
3. Continue conversation for a while, hope for the best.

- This *can* work—if you’re lucky.

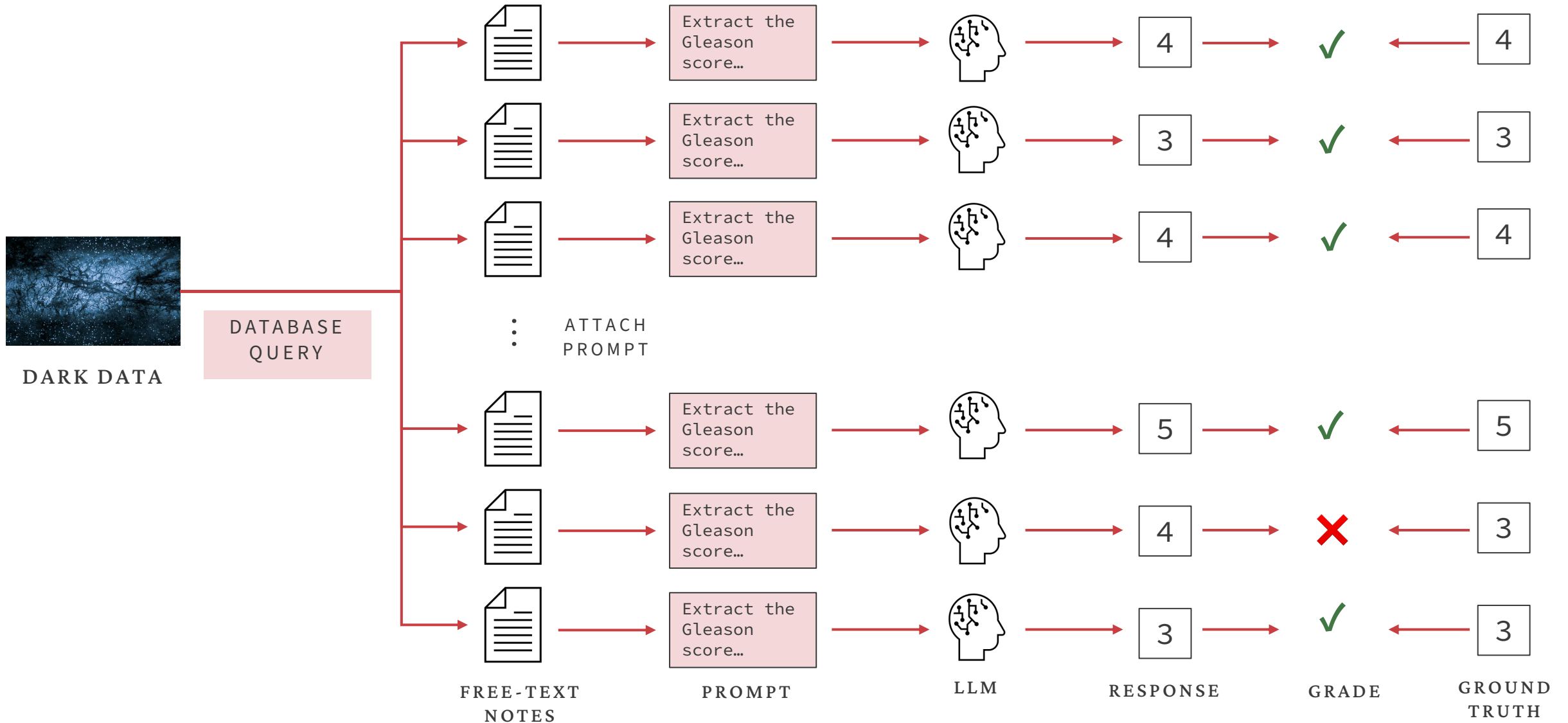
- **FAR BETTER:** systematically *quantify* accuracy and *iterate*.

1. **Curate a gold-standard, labelled** dataset of ~100 pairs of:
(CLINICAL NOTE, EXTRACTED VARIABLE)
2. Define a **pipeline that retrieves clinical notes** from a database.
3. Copy clinical notes into ChatGPT/Gemini with a prompt to extract the variable.
4. *Compute accuracy* of the extraction by comparing against gold-standard dataset.
5. **Iterate on the pipeline/prompt** and return to step 1.

LLMs as input-output machines



LLMs as input-output machines



Full supervised learning paradigm

Recipe for improving input-output machines:

1. Curate a gold-standard, labelled dataset.
2. Randomly split into three subsets:
 - I. **TRAIN**
 - II. **VALIDATION**
 - III. **TEST**
3. Use the **TRAIN** data to optimize the prompt (*more on that later*).
4. Use the **VALIDATION** data to adjust “hyperparameters” of the optimization method (*advanced—more on that later*).
5. Use the **TEST** data to estimate how accurate the LLM extraction is.

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. **Prompting LLMs**
5. The Needle in a Haystack

PROMPTING LLMs

Prompting

- If accuracy is low, the most worthwhile thing to do is *optimize the prompt*.
- **PROMPTING:**
Writing effective instructions for an LLM so that it consistently generates content that meets your requirements.
- **Basic principles:**
 - Be specific.
 - Include all relevant information.
 - Remove irrelevant information.
 - Note which version of ChatGPT/Gemini/Claude you are using.

Prompting best practices

- *Be specific:*
 - **DO:** Extract the secondary Gleason score.
 - **DON'T:** Extract the Gleason score.
- *Give context:*
 - **DO:** This is a pathology report for a prostate biopsy of a patient who is about to undergo radiation therapy. The secondary Gleason score should be between 3 and 5. Extract the Gleason score.
- *Provide examples and reasoning:*
 - **DO:** For example, if the note contains the line “gleason 4+3”, this means the secondary Gleason score is 3, as the secondary score is usually listed as the second in a pair.

Levels of prompt optimization

LEVEL ZERO: Write simple instructions, hope for the best.

LEVEL ONE: Follow prompting best-practices.

LEVEL TWO: Systematically measure accuracy, manually inspect errors, improve prompt, iterate.

LEVEL THREE: Use a tool that automatically improves prompt *using LLMs*:

- PromptWizard
- TextGrad

Levels of prompt optimization

LEVEL ZERO: Write simple instructions, hope for the best.

LEVEL ONE: Follow prompting best-practices.

LEVEL TWO: Systematically measure accuracy, manually inspect errors, improve prompt, iterate.

LEVEL THREE: Use a tool that automatically improves prompt *using LLMs*:


- PromptWizard
- TextGrad



Prompt optimization with TextGrad

TextGrad is a framework for using LLMs to optimize prompts:

Article

Optimizing generative AI by backpropagating language model feedback

<https://doi.org/10.1038/s41586-025-08661-4>
Received: 12 June 2024
Accepted: 16 January 2025
Published online: 19 March 2025
 Check for updates

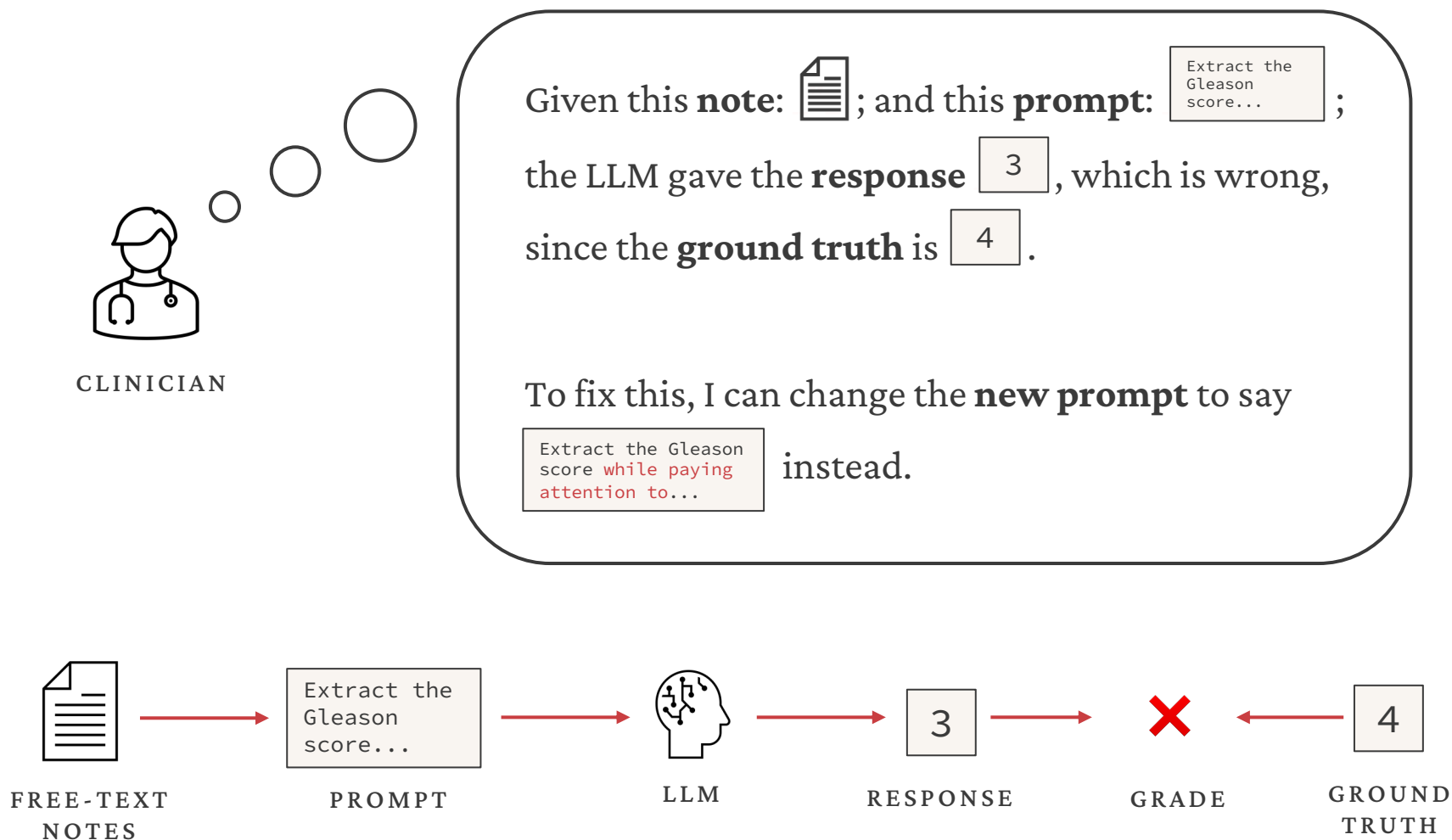
Mert Yuksekgonul^{1,4}, Federico Bianchi^{1,4}, Joseph Boen^{2,4}, Sheng Liu^{2,4}, Pan Lu^{2,4}, Zhi Huang^{2,4}, Carlos Guestrin^{1,3} & James Zou^{1,2,3}

Recent breakthroughs in artificial intelligence (AI) are increasingly driven by systems orchestrating multiple large language models (LLMs) and other specialized tools, such as search engines and simulators. So far, these systems are primarily handcrafted by domain experts and tweaked through heuristics rather than being automatically optimized, presenting a substantial challenge to accelerating progress. The development of artificial neural networks faced a similar challenge until backpropagation and automatic differentiation transformed the field by making optimization turnkey. Analogously, here we introduce TextGrad, a versatile framework that performs optimization by backpropagating LLM-generated feedback to improve AI systems. By leveraging natural language feedback to critique and

Yuksekgonul, M., Bianchi, F., Boen, J. *et al.* Optimizing generative AI by backpropagating language model feedback. *Nature* **639**, 609–616 (2025). <https://doi.org/10.1038/s41586-025-08661-4>

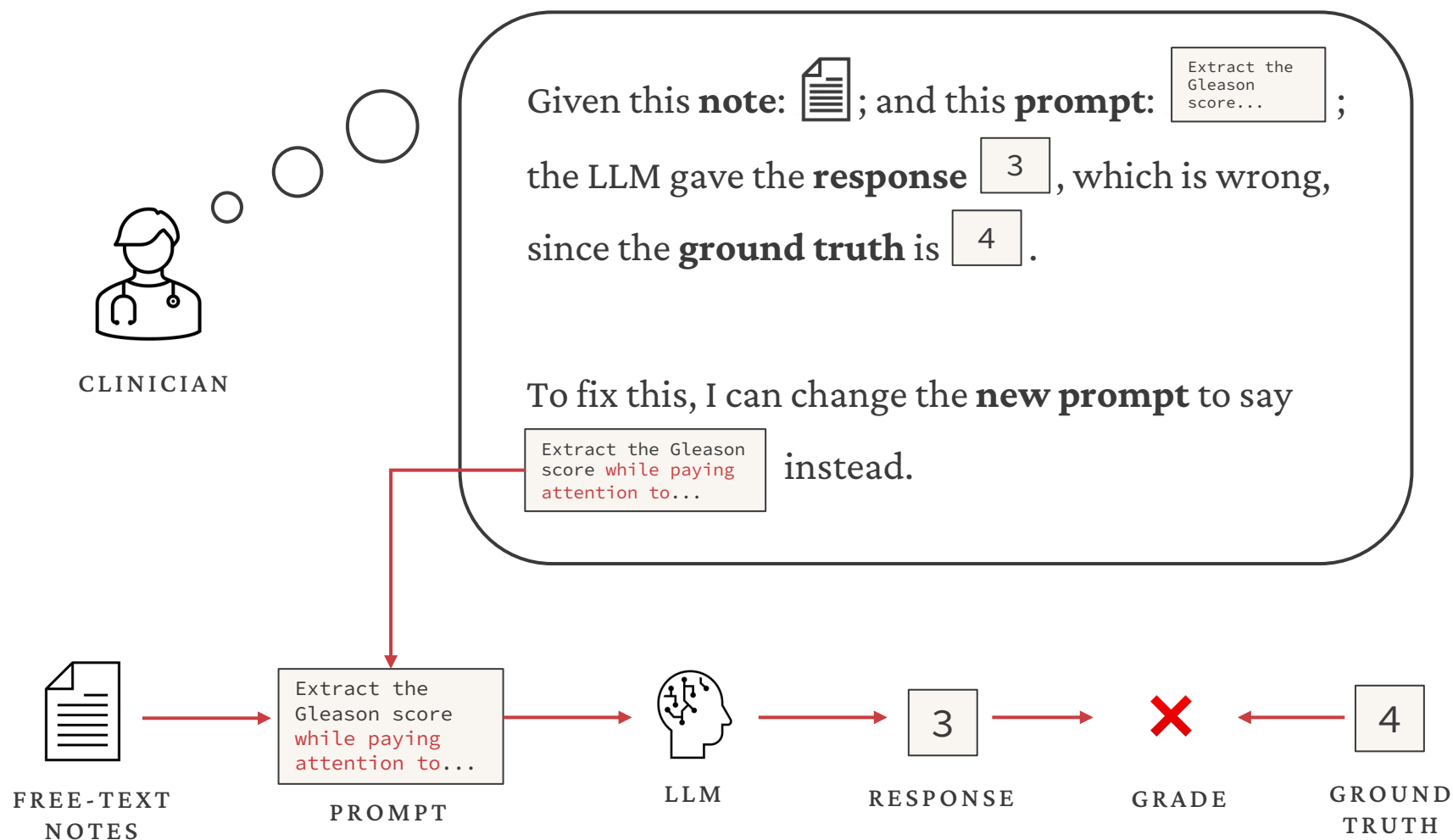
TextGrad idea

LEVEL TWO: Measure accuracy, inspect errors, improve prompt, iterate.



TextGrad idea

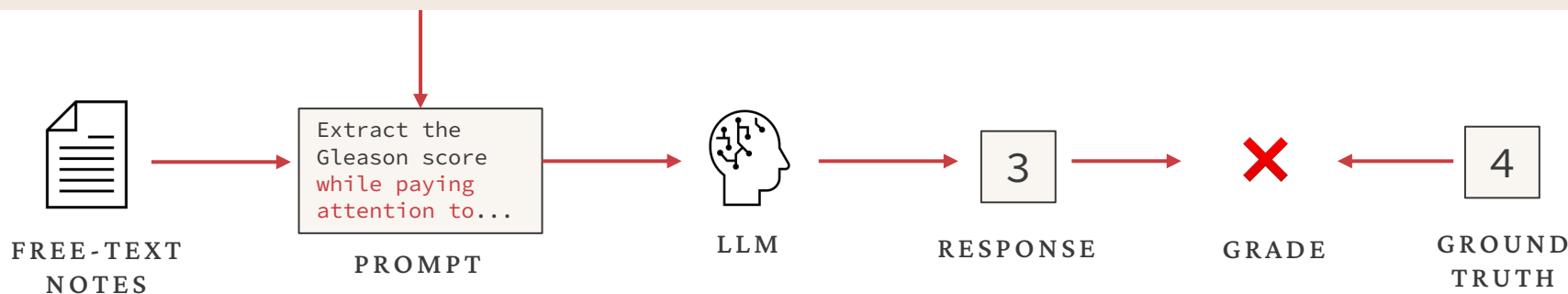
LEVEL TWO: Measure accuracy, inspect errors, improve prompt, iterate.



TextGrad idea

LEVEL TWO: Measure accuracy, inspect errors, improve prompt, iterate.

1. Repeat for every error case in the train set.
2. Go over entire train set again, since fixing some errors *may introduce new ones*.
3. Repeat 1–2 until prompt “converges”.


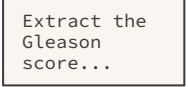




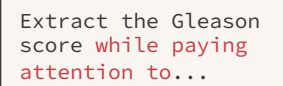
TextGrad idea

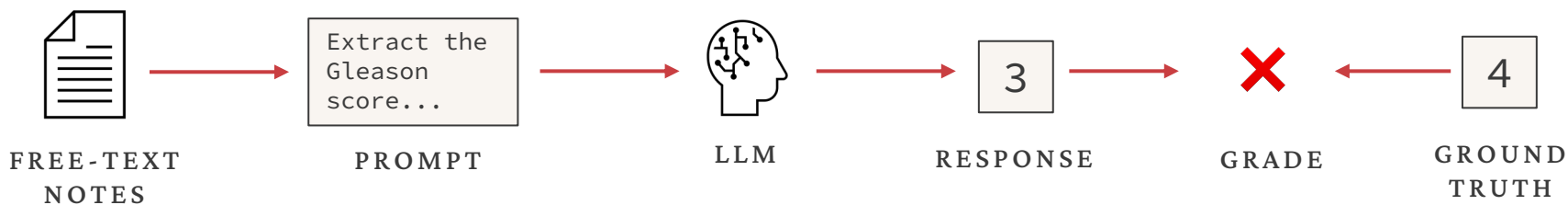
LEVEL THREE: *Replace the data reviewer with an LLM.*



CLINICIAN

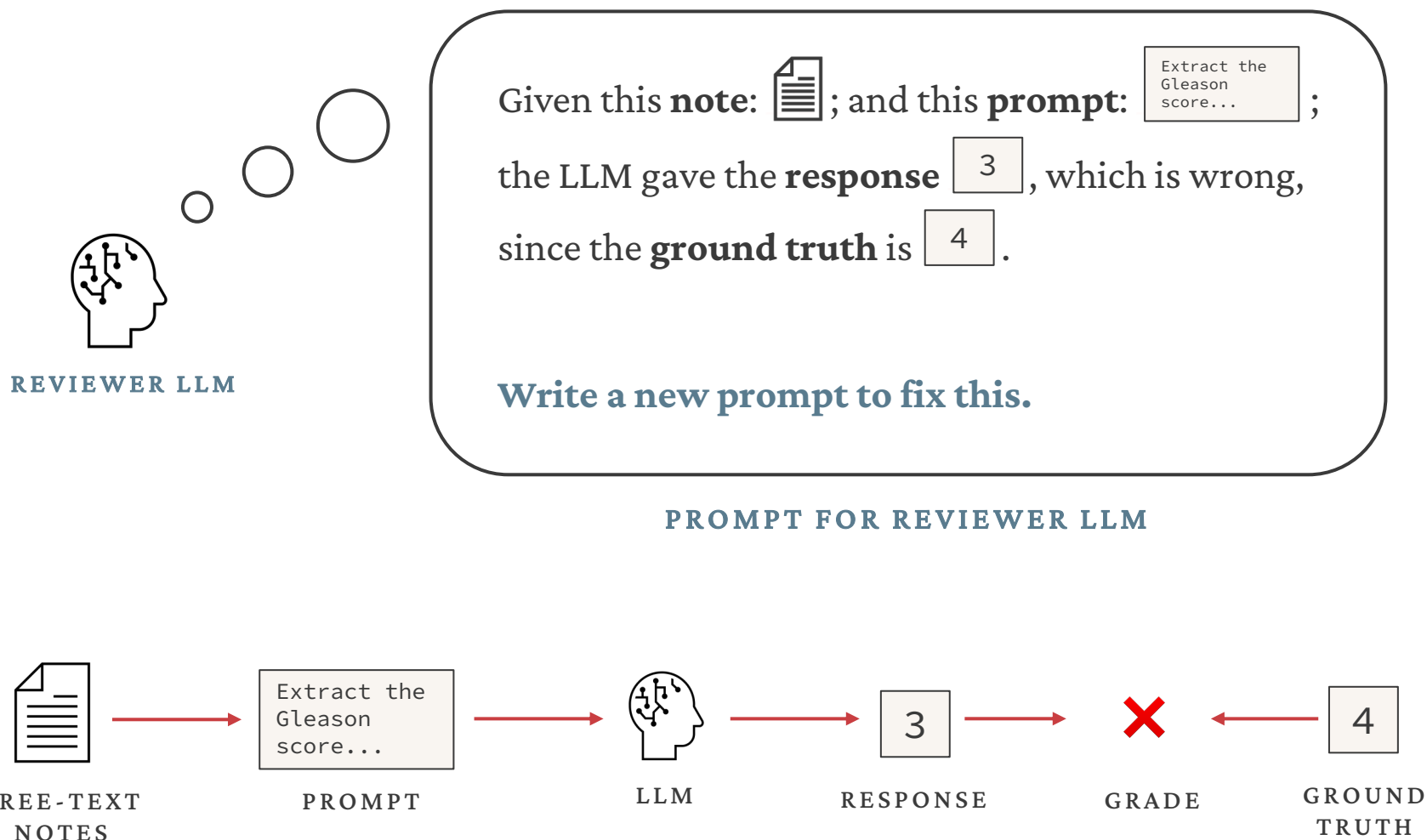
Given this **note**: ; and this **prompt**:  ;
the LLM gave the **response**  , which is wrong,
since the **ground truth** is  .

To fix this, I can change the **new prompt** to say
 instead.



TextGrad idea

LEVEL THREE: *Replace the data reviewer with an LLM.*



TextGrad idea

LEVEL THREE: *Replace the data reviewer with an LLM.*

Advantages:

- LLM never gets tired of data review.
- Can run hundreds of iterations of prompt refinement, checking thousands of answers.
- Our experience:
Can outperform ~1 week of continuous manual clinician prompt optimization.

FREE-TEXT
NOTES

PROMPT

LLM

RESPONSE

GRADE

GROUND
TRUTH

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. The Needle in a Haystack

Roadmap

1. Why AI for Clinical Notes?
2. Large Language Models 101
3. Supervised Learning
4. Prompting LLMs
5. **The Needle in a Haystack**

THE NEEDLE IN A HAYSTACK

Care journey can generate hundreds of notes



Care journey can generate hundreds of notes

UROLOGY



Some questions depend on understanding entire timeline:

- Which patients had a recurrence?
- How long between initial treatment and recurrence?

ONCOLOGY

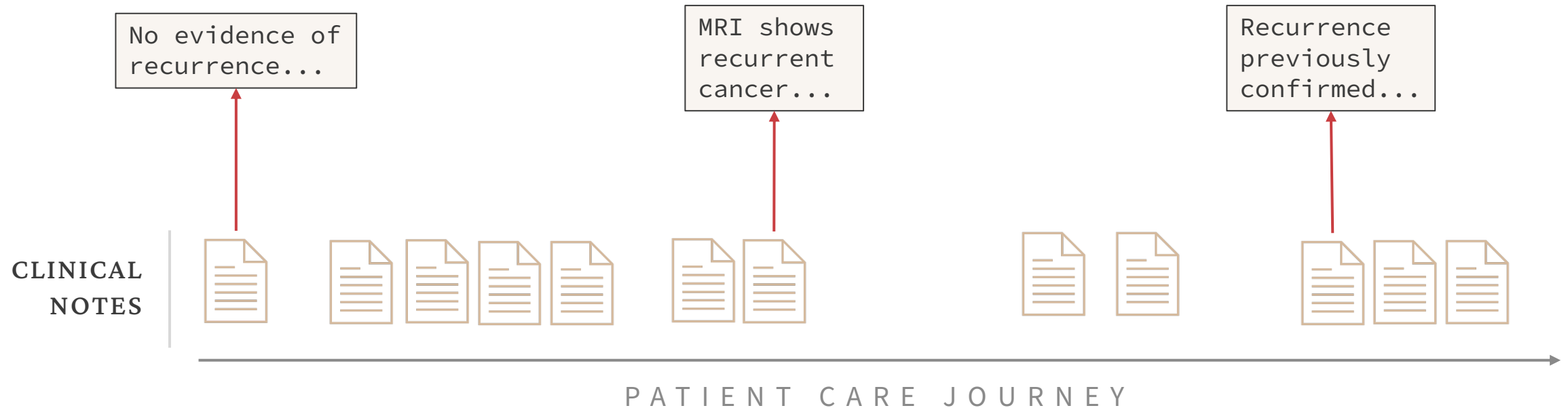


PATIENT CARE JOURNEY

The needle in a haystack problem

- LLM performance tends to **degrade** with longer inputs.
- Often with clinical notes, only small fraction relevant to answer the query: *the needle in a haystack*.

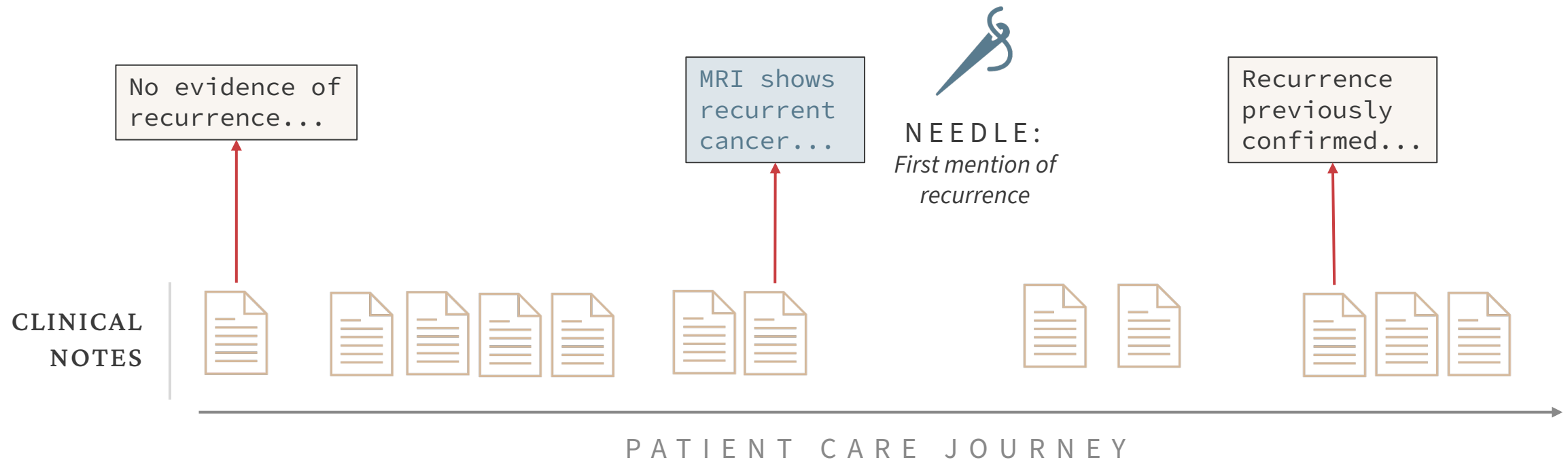
EXAMPLE: RECURRENCE



The needle in a haystack problem

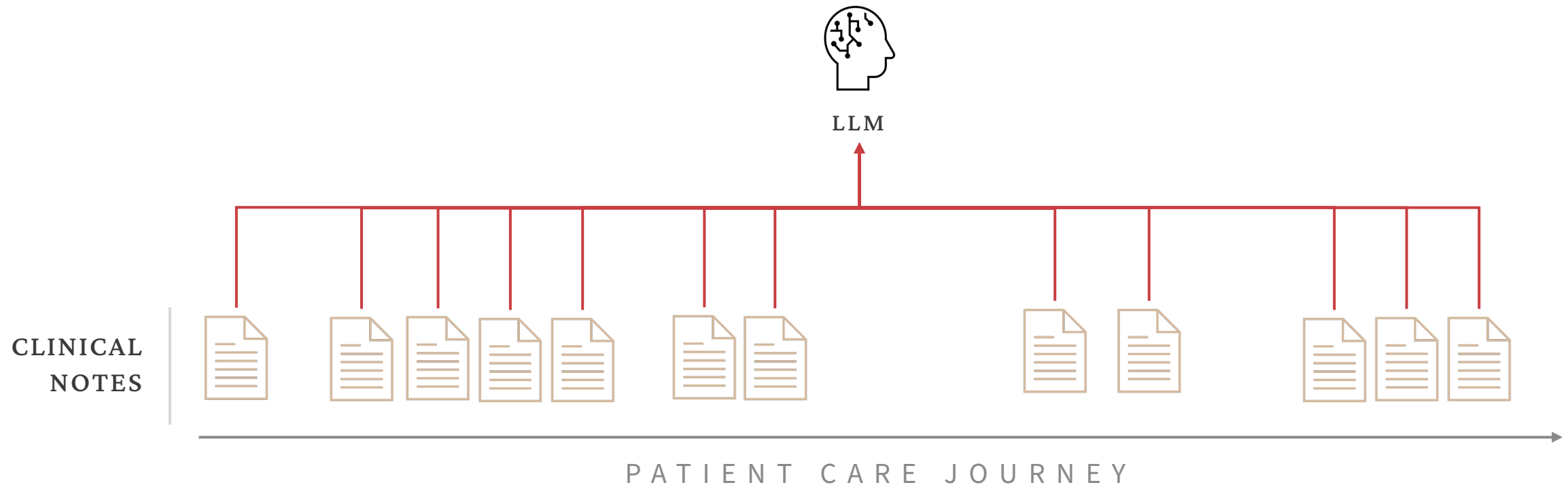
- LLM performance tends to **degrade** with longer inputs.
- Often with clinical notes, only small fraction relevant to answer the query:
the needle in a haystack.

EXAMPLE: RECURRENCE



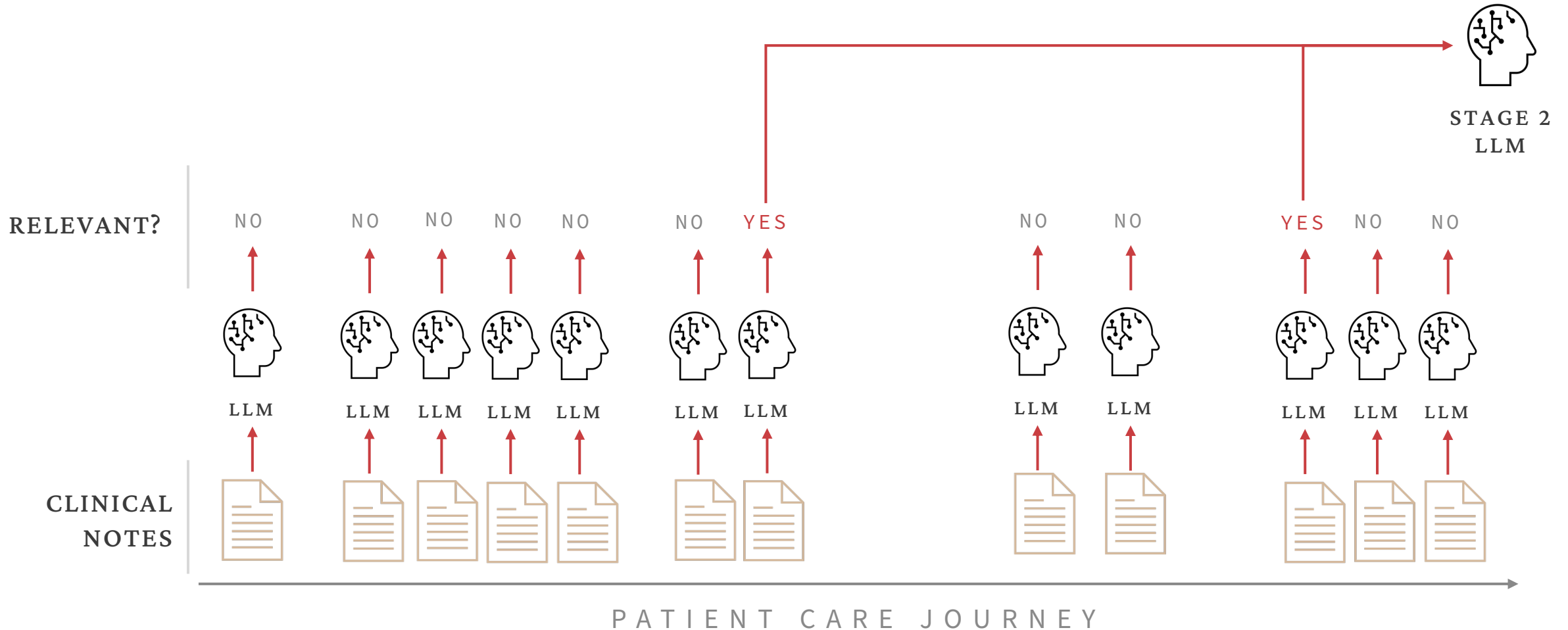
The needle in a haystack problem

- **NAÏVE APPROACH:** Copy all notes into single prompt, ask LLM to find date of first mention of recurrence, if any.
 - Accuracy likely low—needle in a haystack.
 - Expensive/slow—extremely long prompt.



Two-stage approach

- *Stage 1:* Cheap, fast LLM processes each note in parallel, and flags **relevant** notes.
- *Stage 2:* **Relevant** notes only read by more advanced LLM. BONUS: get **references**.



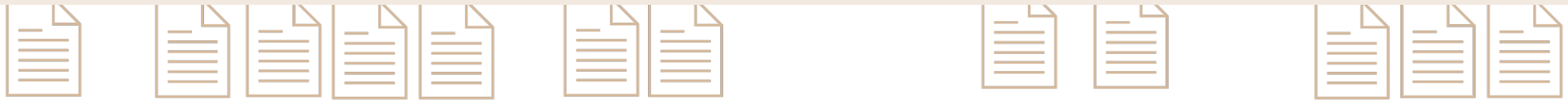
Two-stage approach

- *Stage 1:* Cheap, fast LLM processes each note in parallel, and flags **relevant** notes.
- *Stage 2:* **Relevant** notes only read by more advanced LLM. BONUS: get **references**.

Our experience:

- 98% accuracy determining which patients had recurrence.
- Up to hundreds of clinical notes read per patient.
- Approx. \$0.20 per patient.

CLINICAL
NOTES



PATIENT CARE JOURNEY

Recap

1. LLMs can turn untapped *dark data* into *actionable, structured data*.
2. Optimize the prompt.
 - Systematically track accuracy of data extraction.
 - Consider automatic optimizers.
3. Beware the *needle in a haystack problem*.
 - Consider a multi-stage approach.

Thanks for listening!

QUESTIONS & ANSWERS
