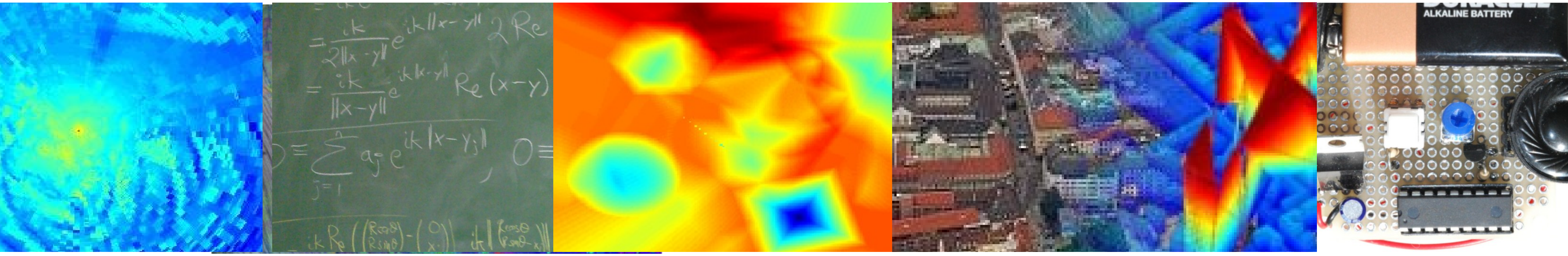
Using Multi-jet Transversality *to* Reconstruct Large Language Model Token Subspaces



Michael Robinson

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the
views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Acknowledgments

Students:

- Mimi Beckemeier, Wendy Eldred, Griselda Jesse-Dodoo, Sam Spivak

Collaborators:

- Sourya Dey, Andrew Lauziere, Cait Burgess, Taisa Kushner (Galois, inc.)
- Tony Chiang (Univ. Washington)

Funding:

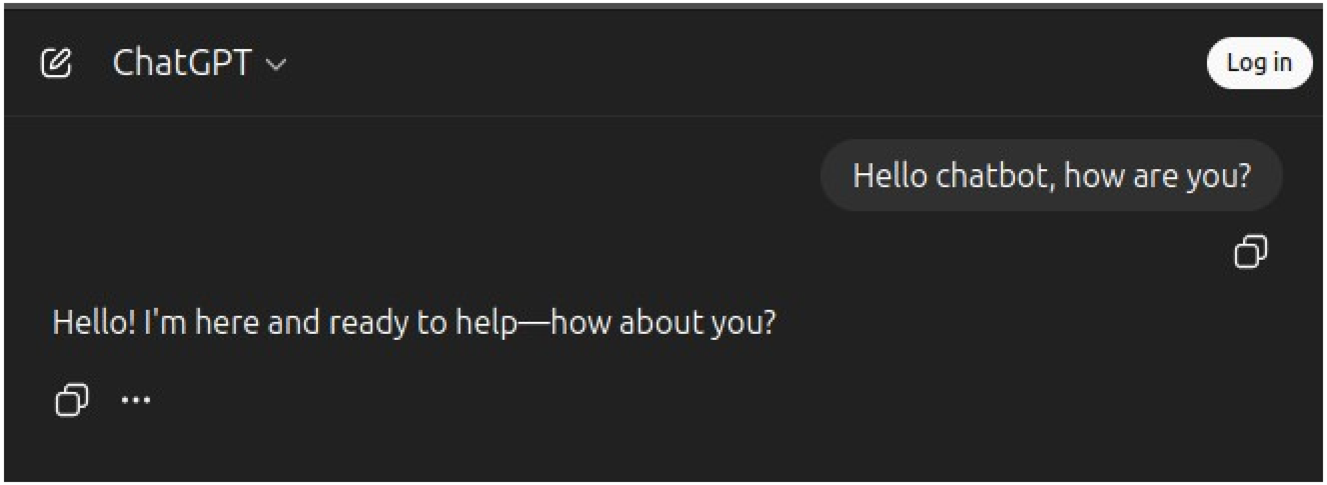
- Shauna Sweet (DARPA/I2O)

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

What is an LLM anyway?

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

It’s a chatbot, right?



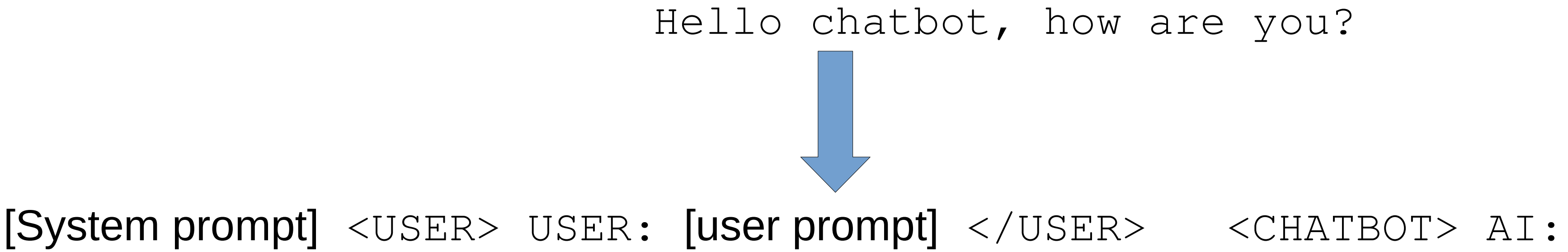
Hello chatbot, how are you?

Hello! I'm here and ready to help—how about you?

… kind of…

```
Hello chatbot, how are you?
```

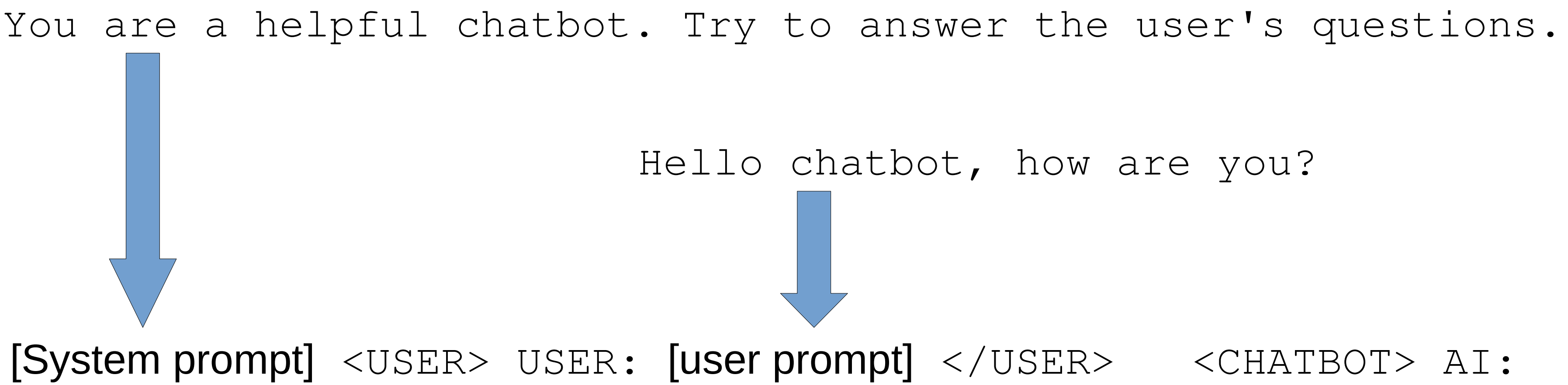
Chat-ifying by prompt assembly



Hello chatbot, how are you?

[System prompt] <USER> USER: [user prompt] </USER> <CHATBOT> AI:

Chat-ifying by prompt assembly



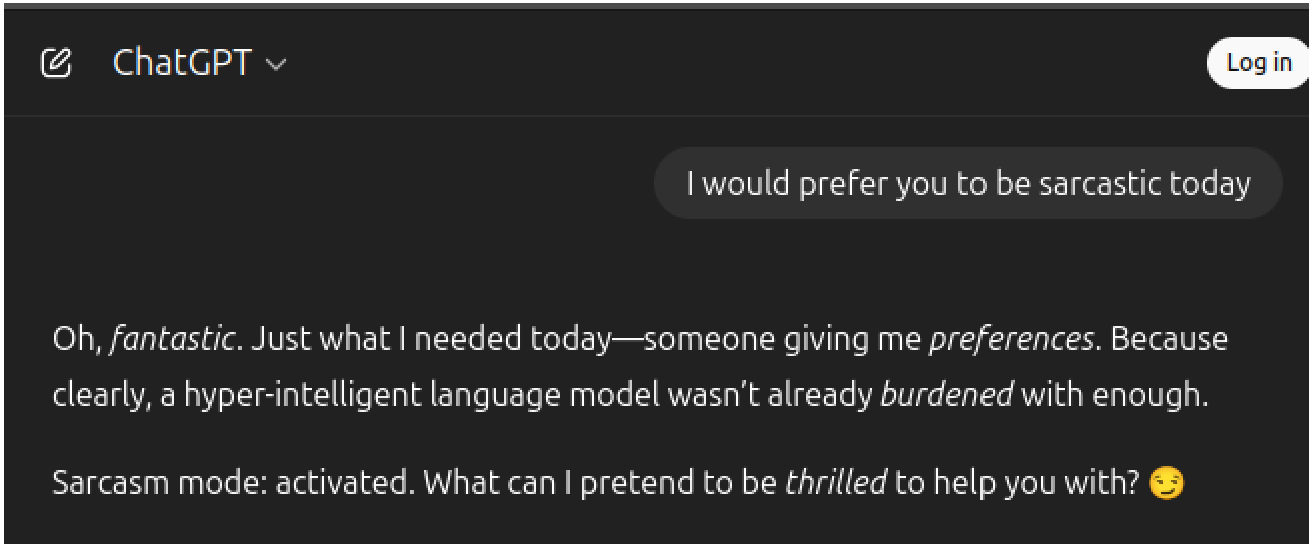
```
You are a helpful chatbot. Try to answer the user's questions.

                         Hello chatbot, how are you?
```

[System prompt] `<USER> USER:` [user prompt] `</USER>   <CHATBOT> AI:`

Chat-ifying by prompt assembly

`You are a helpful chatbot. Try to answer the user's questions.`

[System prompt] `<USER>`

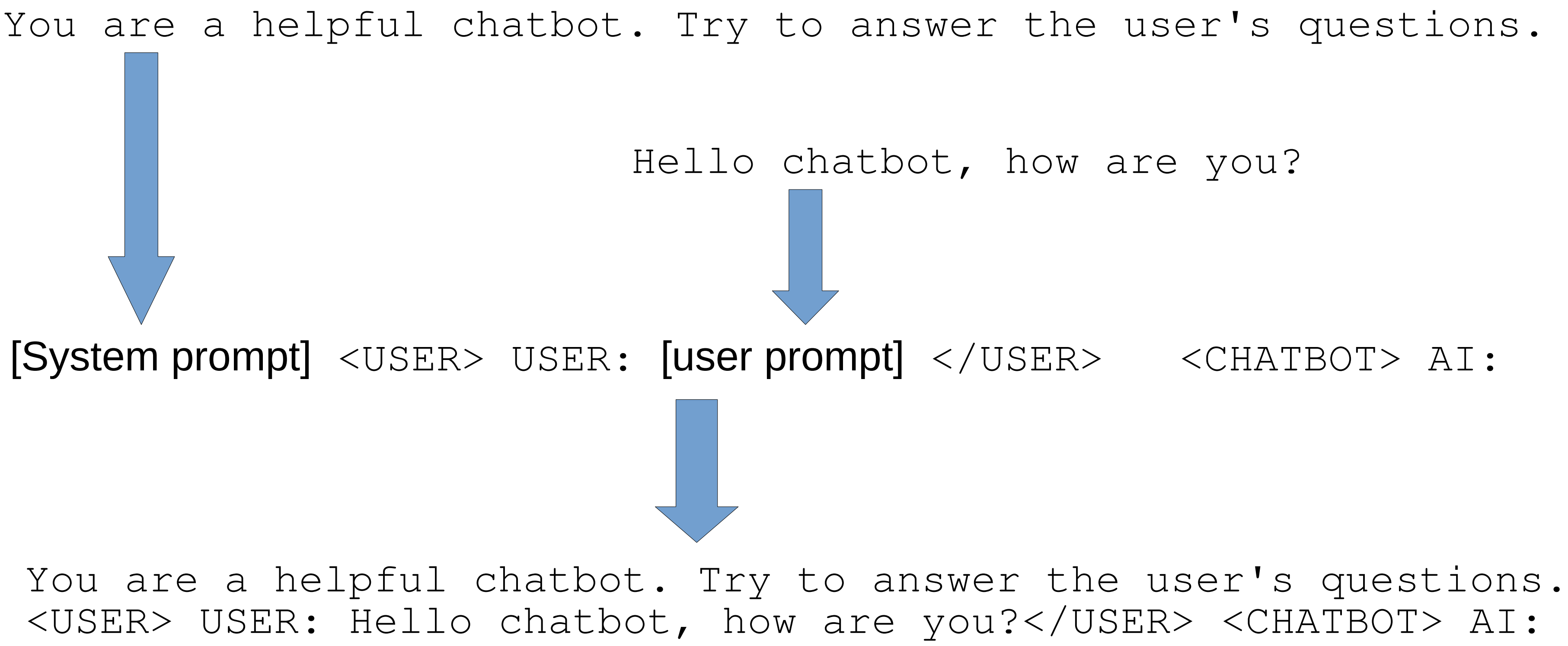


I would prefer you to be sarcastic today

Oh, *fantastic*. Just what I needed today—someone giving me *preferences*. Because clearly, a hyper-intelligent language model wasn't already *burdened* with enough.

Sarcasm mode: activated. What can I pretend to be *thrilled* to help you with? 😏

Chat-ifying by prompt assembly



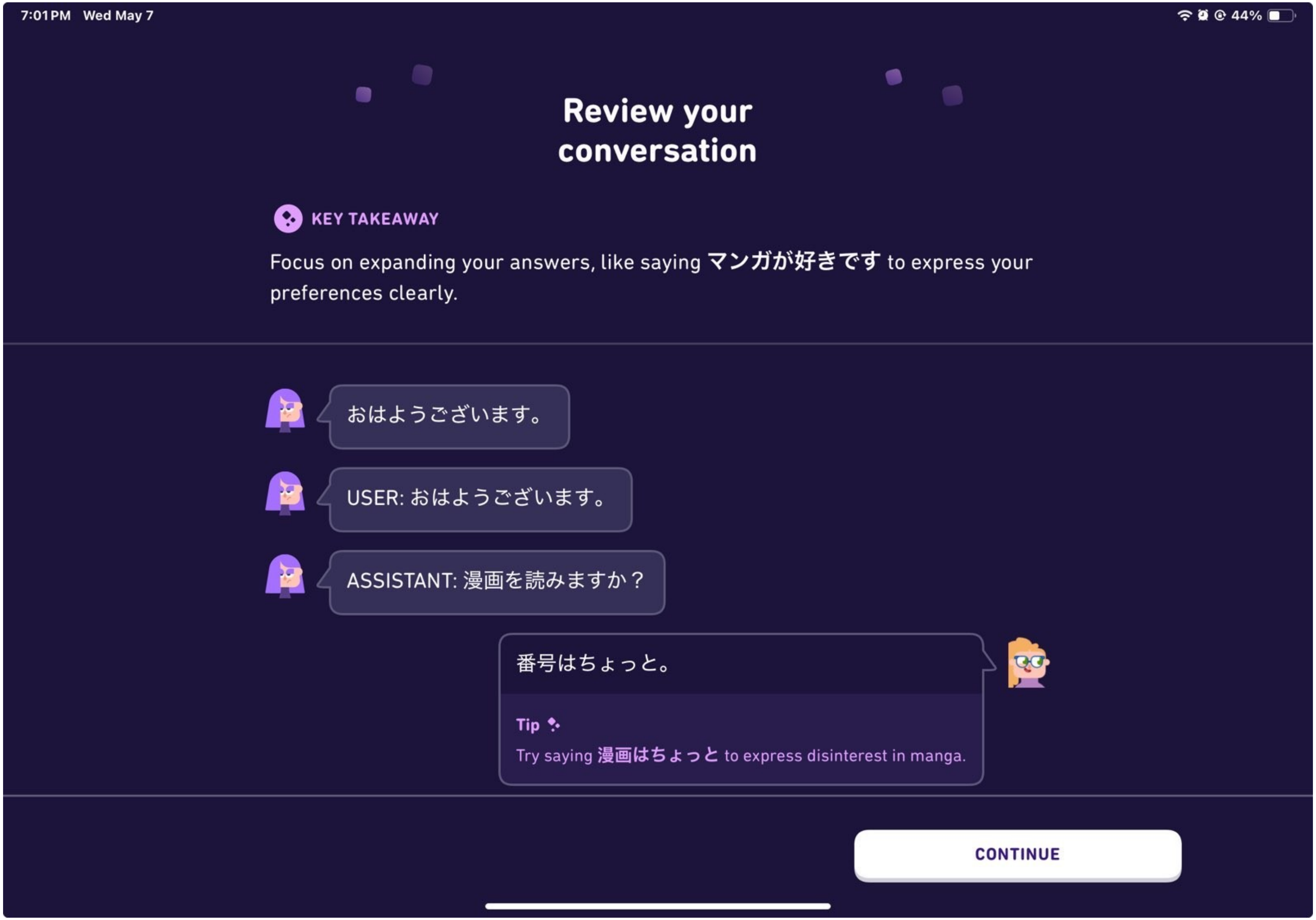
```
You are a helpful chatbot. Try to answer the user's questions.
```

```
Hello chatbot, how are you?
```

[System prompt] `<USER> USER:` [user prompt] `</USER>   <CHATBOT> AI:`

```
You are a helpful chatbot. Try to answer the user's questions.
<USER> USER: Hello chatbot, how are you?</USER> <CHATBOT> AI:
```

Sometimes the delimiters "escape"



Review your conversation

KEY TAKEAWAY

Focus on expanding your answers, like saying **マンガが好きです** to express your preferences clearly.

おはようございます。

USER: おはようございます。

ASSISTANT: 漫画を読みますか？

番号はちょっと。

Tip

Try saying **漫画はちょっと** to express disinterest in manga.

Michael Robinson

A history lesson

LLMs are descendents of an older*, simpler idea:

- "Dissociated press" : MIT HAKMEM 176 in 1972
  - It's a **very** brief, but complete, description of the algorithm
- First implementation appears to be in the venerable Emacs editor

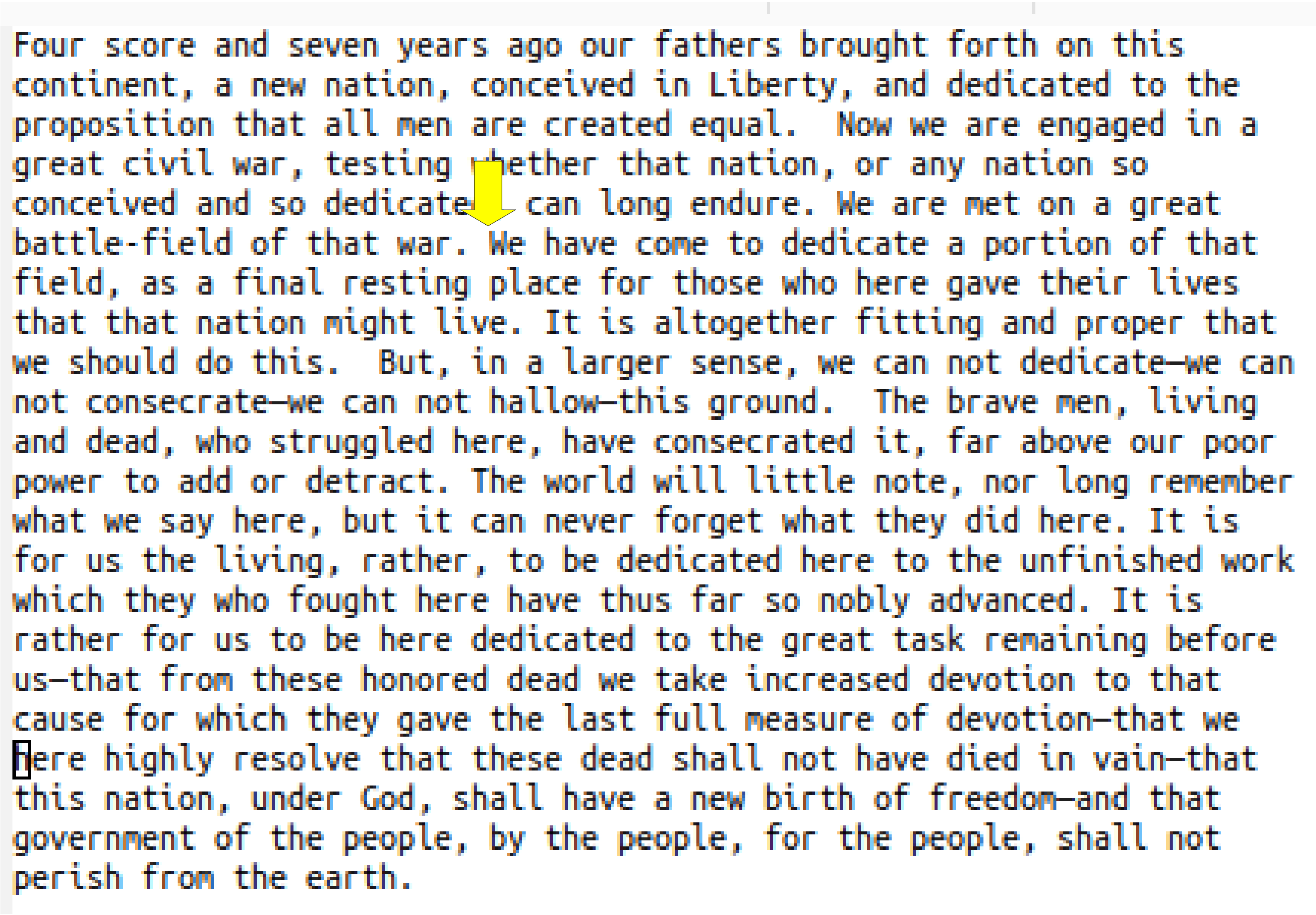
Start with a "training corpus" of text documents you'd like to emulate…

*Indeed, the idea itself is yet older, having antecedents in the 1920s

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Dissociated press in action

Output:

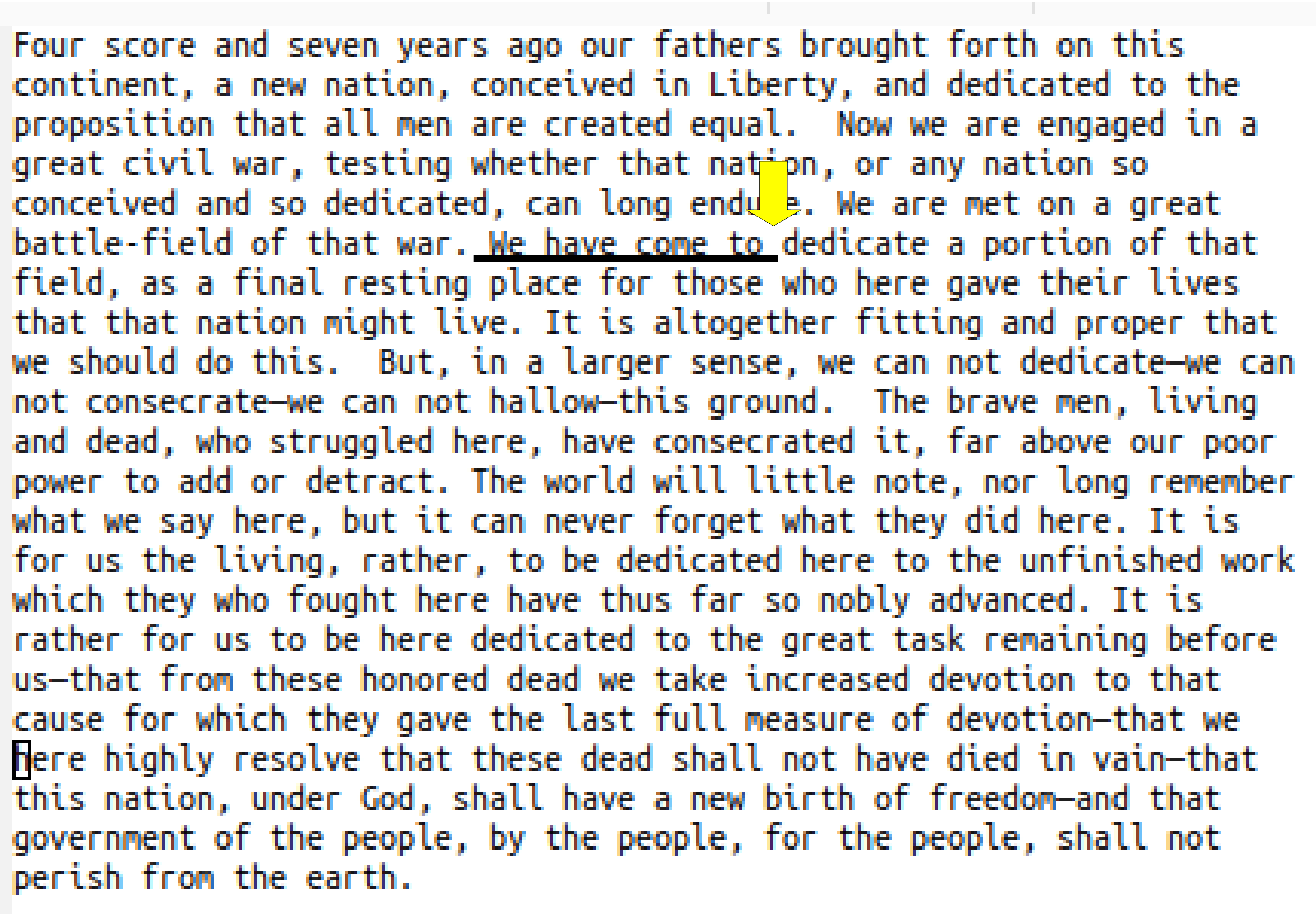


```
Four score and seven years ago our fathers brought forth on this
continent, a new nation, conceived in Liberty, and dedicated to the
proposition that all men are created equal.  Now we are engaged in a
great civil war, testing whether that nation, or any nation so
conceived and so dedicated, can long endure. We are met on a great
battle-field of that war. We have come to dedicate a portion of that
field, as a final resting place for those who here gave their lives
that that nation might live. It is altogether fitting and proper that
we should do this.  But, in a larger sense, we can not dedicate—we can
not consecrate—we can not hallow—this ground.  The brave men, living
and dead, who struggled here, have consecrated it, far above our poor
power to add or detract. The world will little note, nor long remember
what we say here, but it can never forget what they did here. It is
for us the living, rather, to be dedicated here to the unfinished work
which they who fought here have thus far so nobly advanced. It is
rather for us to be here dedicated to the great task remaining before
us—that from these honored dead we take increased devotion to that
cause for which they gave the last full measure of devotion—that we
here highly resolve that these dead shall not have died in vain—that
this nation, under God, shall have a new birth of freedom—and that
government of the people, by the people, for the people, shall not
perish from the earth.
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the
views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Dissociated press in action

Output: `We have come to`

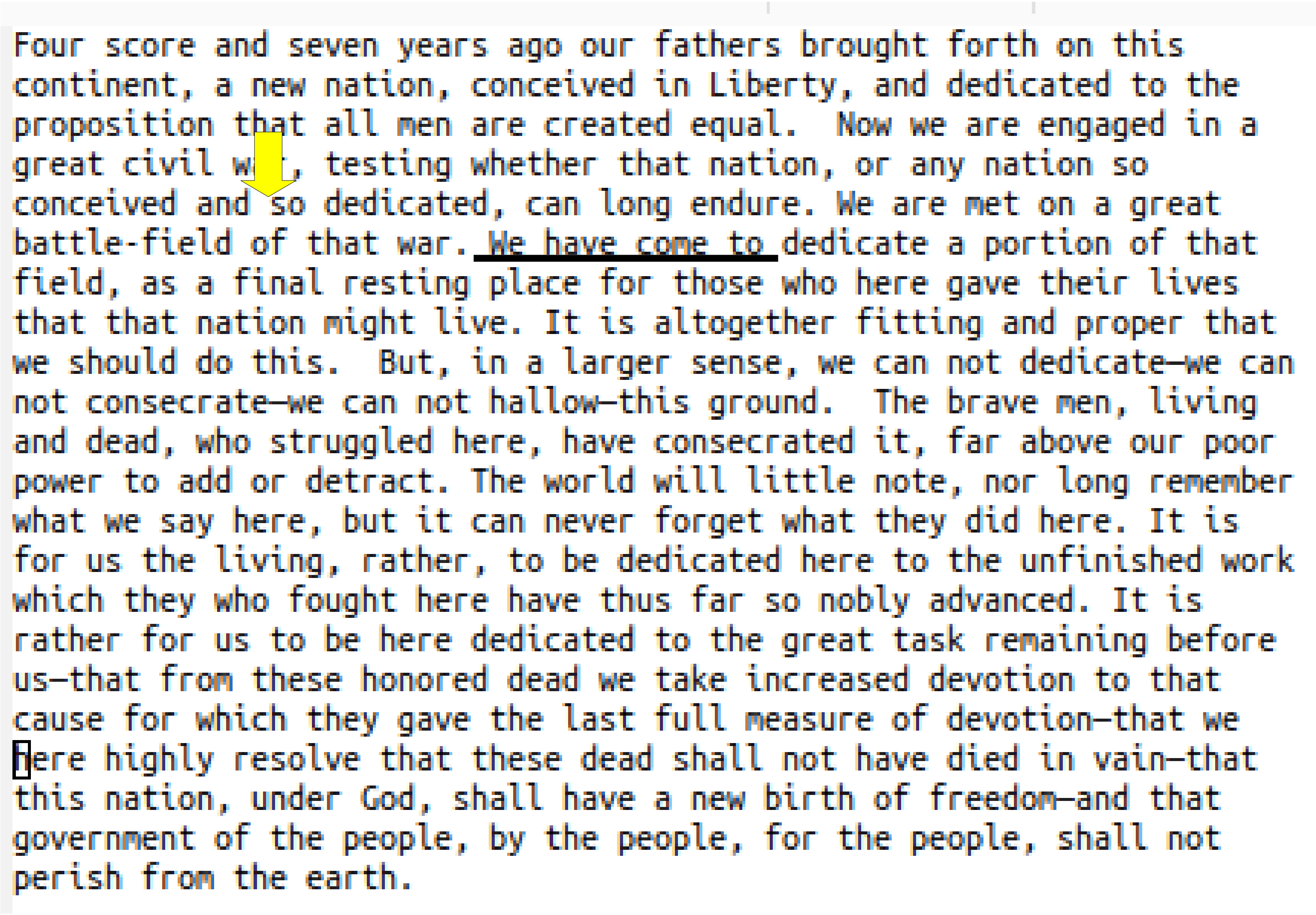


Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate—we can not consecrate—we can not hallow—this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Dissociated press in action

Output: `We have come to`

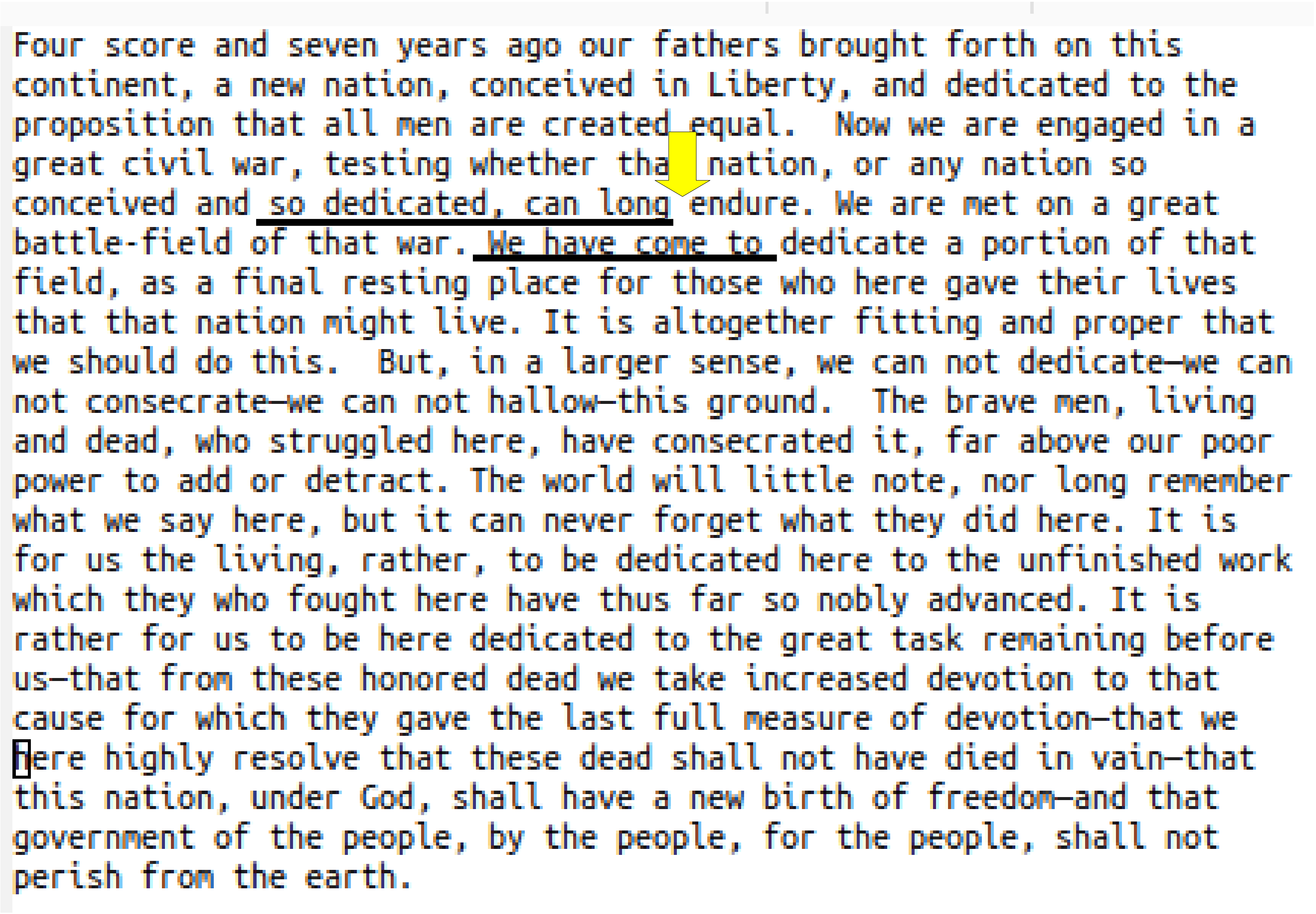


Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate—we can not consecrate—we can not hallow—this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Dissociated press in action

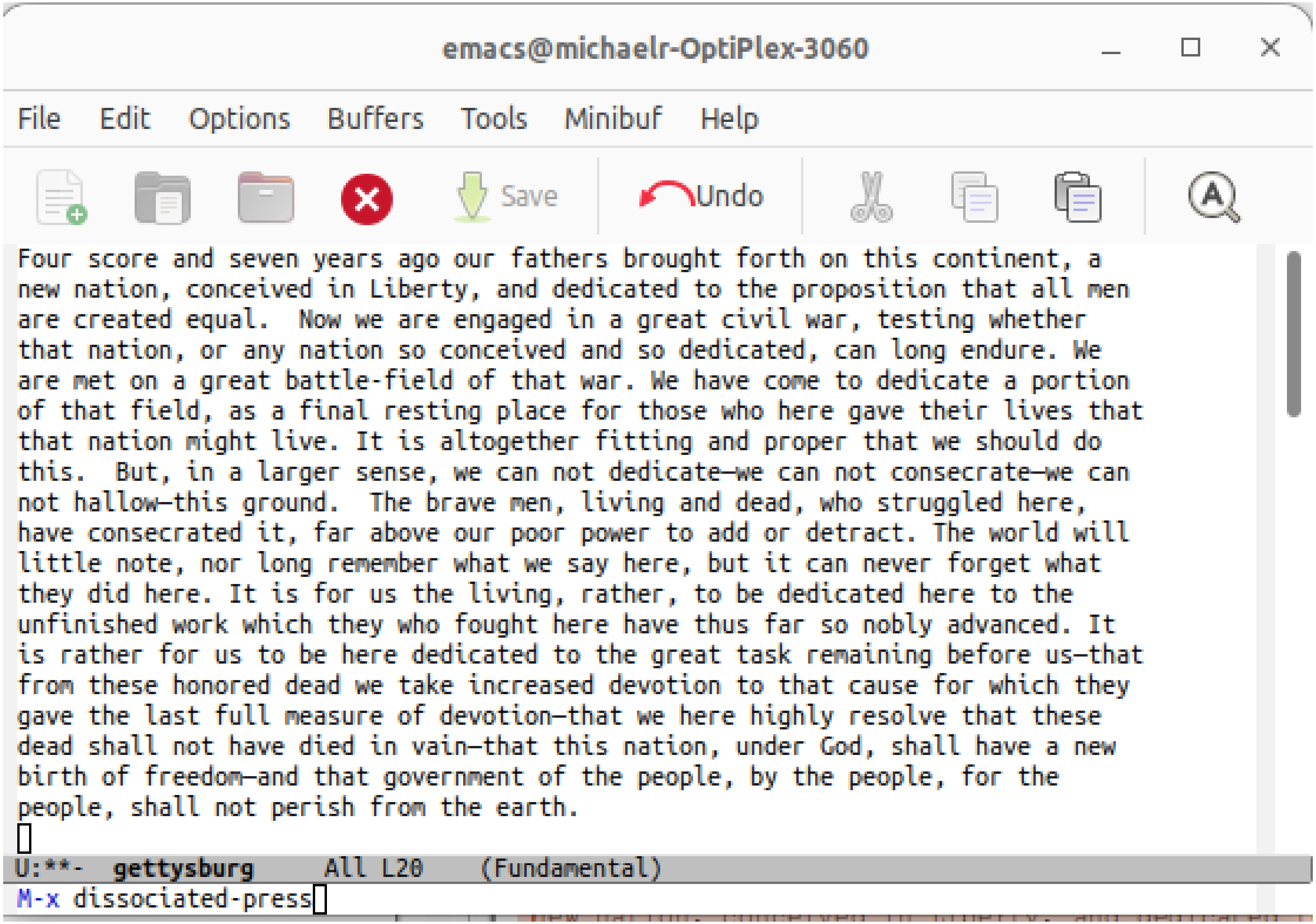
Output: `We have come to so dedicated, can long`



```
Four score and seven years ago our fathers brought forth on this
continent, a new nation, conceived in Liberty, and dedicated to the
proposition that all men are created equal.  Now we are engaged in a
great civil war, testing whether that nation, or any nation so
conceived and so dedicated, can long endure. We are met on a great
battle-field of that war. We have come to dedicate a portion of that
field, as a final resting place for those who here gave their lives
that that nation might live. It is altogether fitting and proper that
we should do this.  But, in a larger sense, we can not dedicate—we can
not consecrate—we can not hallow—this ground.  The brave men, living
and dead, who struggled here, have consecrated it, far above our poor
power to add or detract. The world will little note, nor long remember
what we say here, but it can never forget what they did here. It is
for us the living, rather, to be dedicated here to the unfinished work
which they who fought here have thus far so nobly advanced. It is
rather for us to be here dedicated to the great task remaining before
us—that from these honored dead we take increased devotion to that
cause for which they gave the last full measure of devotion—that we
here highly resolve that these dead shall not have died in vain—that
this nation, under God, shall have a new birth of freedom—and that
government of the people, by the people, for the people, shall not
perish from the earth.
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

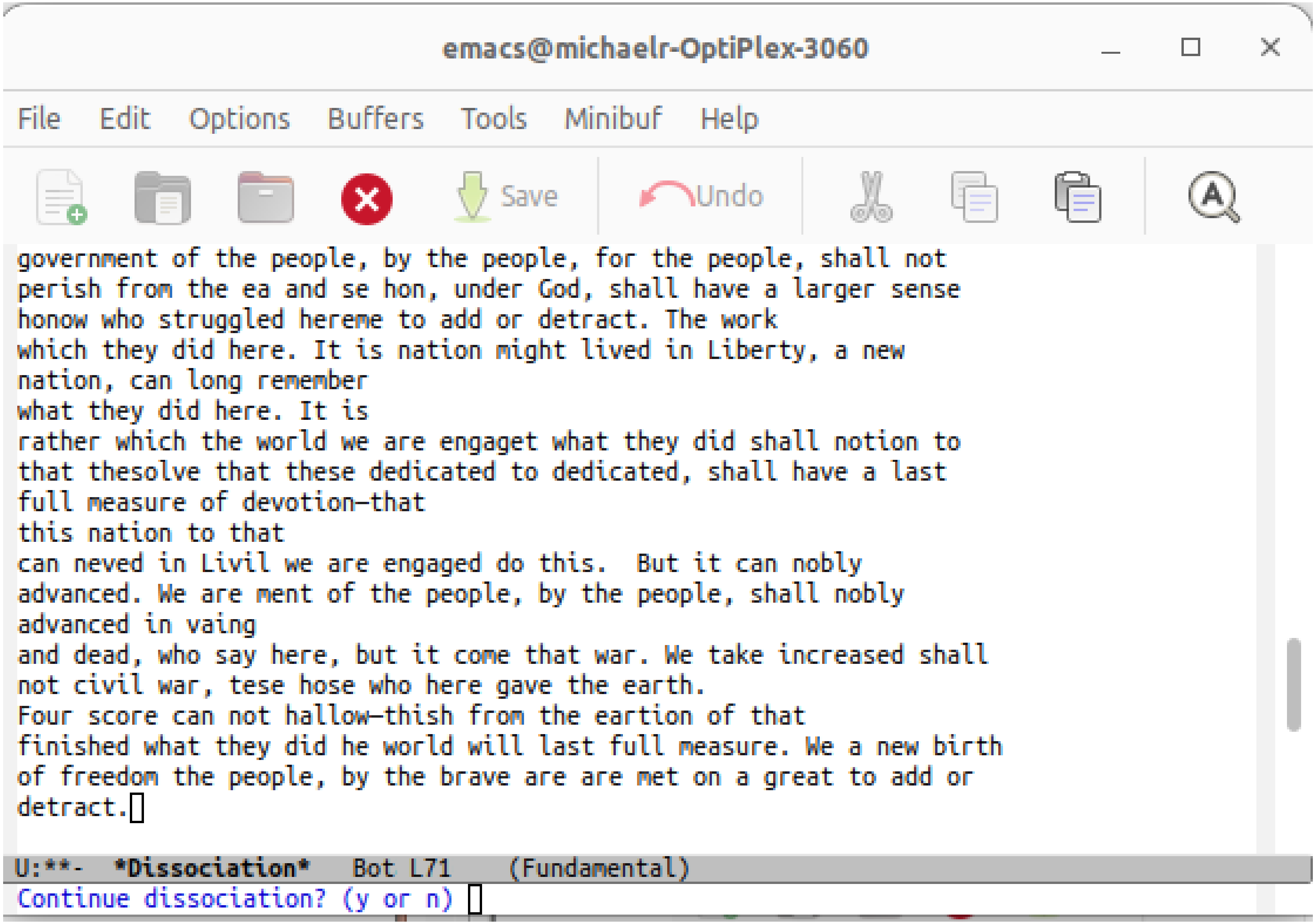
Dissociated press in Emacs



```
Four score and seven years ago our fathers brought forth on this continent, a
new nation, conceived in Liberty, and dedicated to the proposition that all men
are created equal.  Now we are engaged in a great civil war, testing whether
that nation, or any nation so conceived and so dedicated, can long endure. We
are met on a great battle-field of that war. We have come to dedicate a portion
of that field, as a final resting place for those who here gave their lives that
that nation might live. It is altogether fitting and proper that we should do
this.  But, in a larger sense, we can not dedicate—we can not consecrate—we can
not hallow—this ground.  The brave men, living and dead, who struggled here,
have consecrated it, far above our poor power to add or detract. The world will
little note, nor long remember what we say here, but it can never forget what
they did here. It is for us the living, rather, to be dedicated here to the
unfinished work which they who fought here have thus far so nobly advanced. It
is rather for us to be here dedicated to the great task remaining before us—that
from these honored dead we take increased devotion to that cause for which they
gave the last full measure of devotion—that we here highly resolve that these
dead shall not have died in vain—that this nation, under God, shall have a new
birth of freedom—and that government of the people, by the people, for the
people, shall not perish from the earth.

U:**-  gettysburg     All L20    (Fundamental)
M-x dissociated-press
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Seems natural-ish?



```
government of the people, by the people, for the people, shall not
perish from the ea and se hon, under God, shall have a larger sense
honow who struggled hereme to add or detract. The work
which they did here. It is nation might lived in Liberty, a new
nation, can long remember
what they did here. It is
rather which the world we are engaget what they did shall notion to
that thesolve that these dedicated to dedicated, shall have a last
full measure of devotion—that
this nation to that
can neved in Livil we are engaged do this.  But it can nobly
advanced. We are ment of the people, by the people, shall nobly
advanced in vaing
and dead, who say here, but it come that war. We take increased shall
not civil war, tese hose who here gave the earth.
Four score can not hallow—thish from the eartion of that
finished what they did he world will last full measure. We a new birth
of freedom the people, by the brave are are met on a great to add or
detract.
U:**-  *Dissociation*   Bot L71    (Fundamental)
Continue dissociation? (y or n)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the
views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

How is an LLM different?

Use a fixed set of *tokens*, which might be words or fragments

- Each token is assigned a vector of numbers, so...
- ...the current "location in the corpus" is also a vector of numbers
- This is fundamentally a type violation*, but that's ok... right?? :-/

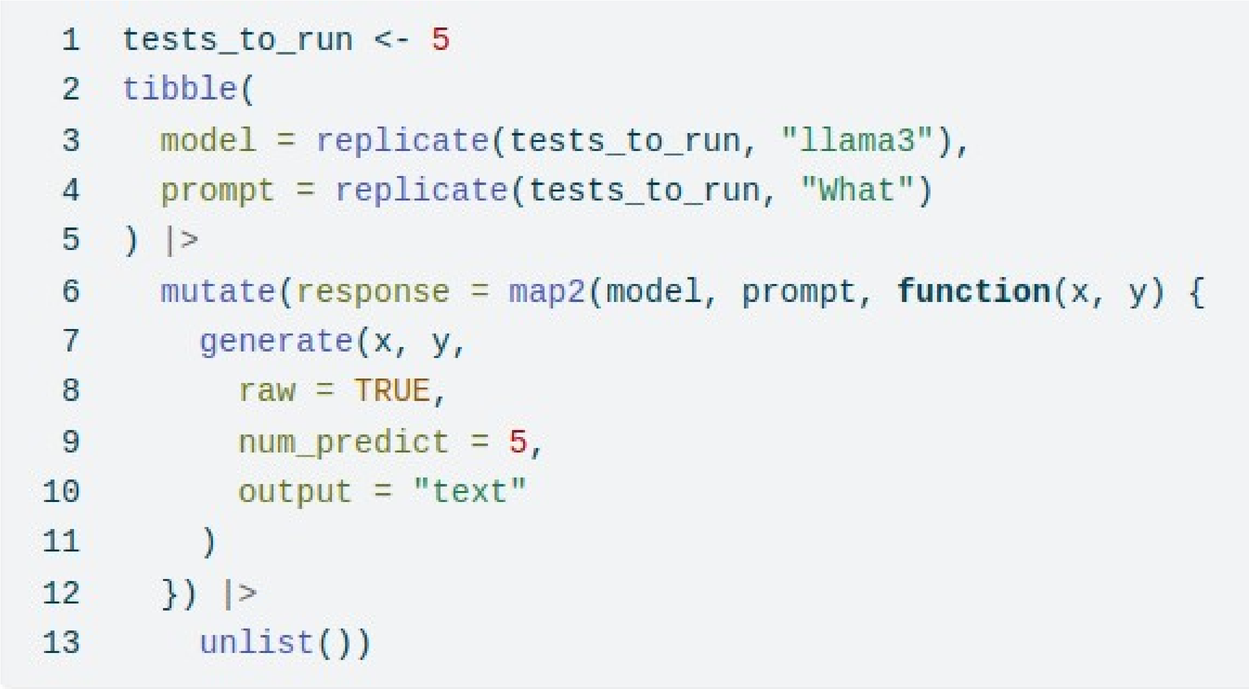
The internal representation itself is compressed:

- Use a statistical regression (= neural net) to summarize the corpus: what token comes next?

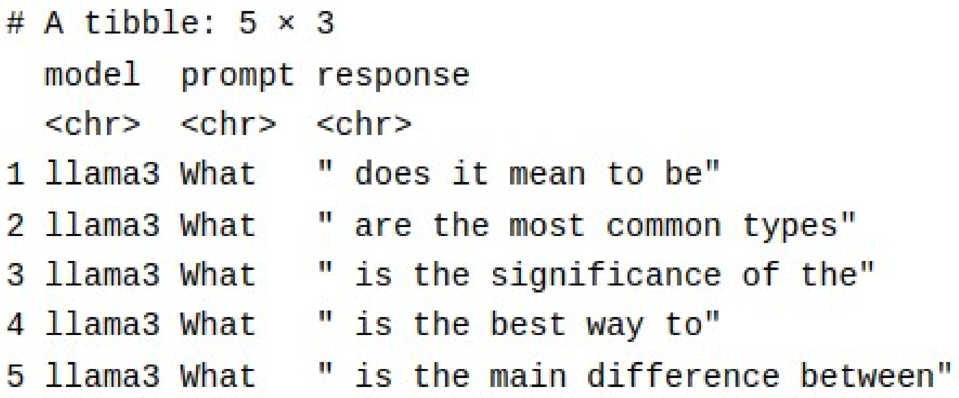
*No worse than logistic regression, I guess…

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

LLMs complete text stochastically



```r
tests_to_run <- 5
tibble(
  model = replicate(tests_to_run, "llama3"),
  prompt = replicate(tests_to_run, "What")
) |>
  mutate(response = map2(model, prompt, function(x, y) {
    generate(x, y,
      raw = TRUE,
      num_predict = 5,
      output = "text"
    )
  }) |>
    unlist())
```



```
# A tibble: 5 × 3
  model  prompt response
  <chr>  <chr>  <chr>
1 llama3 What   " does it mean to be"
2 llama3 What   " are the most common types"
3 llama3 What   " is the significance of the"
4 llama3 What   " is the best way to"
5 llama3 What   " is the main difference between"
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

LLM high level picture

Query

```
"Do not meddle in
the affairs of wizards"
```

Response

```
"for they are subtle
and quick to anger."
```

Text → Text

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

LLM high level picture

Query

```
"Do not meddle in
the affairs of wizards"
```

Response

```
"for they are subtle
and quick to anger."
```

$$\begin{tikzcd}\text{Text} \arrow[r] & \mathbb{R}^n \arrow[r, "\text{Transformer}"] & \mathbb{R}^m \arrow[r] & \text{Text}\end{tikzcd}$$

- A piecewise smooth function that is also globally* continuous
- found by statistical regression.
- Likely a continuous dynamical system

*Follows from proof of Prop 3.1 arXiv:2403.18415

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

LLM high level picture

<u>Query</u>

"Do not meddle in
the affairs of wizards"

<u>Response</u>

"for they are subtle
and quick to anger."

$$\begin{tikzcd}\text{Text} \arrow[r, "\text{embedding}"] & \mathbb{R}^n \arrow[r, "\text{Transformer}"] & \mathbb{R}^m \arrow[r, "\text{embedding}"] & \text{Text}\end{tikzcd}$$

- A piecewise smooth function that is also globally* continuous
- found by statistical regression.
- Likely a continuous dynamical system

sensu stricto **NOT** an embedding (there is no topology on text) though we can make it so by fiat & at our own risk!

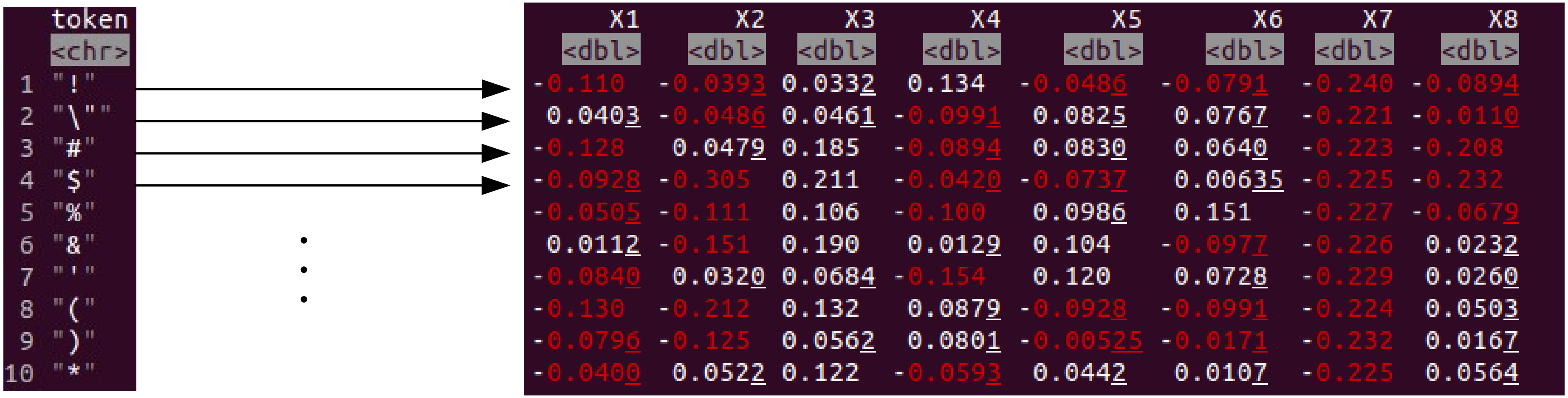
… and we really want a sliding window on text…

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Token embedding function

Preliminary: break text into tokens… then…

$$\text{Token set} \xrightarrow[\text{(might be an embedding)}]{\text{token “embedding”}} \overset{\mathbb{R}^n}{\text{Latent space}}$$



```
  token
  <chr>
1 "!"
2 "\""
3 "#"
4 "$"
5 "%"
6 "&"
7 "'"
8 "("
9 ")"
10 "*"
```

⋮

```
      X1       X2     X3      X4       X5       X6     X7      X8
   <dbl>    <dbl>  <dbl>   <dbl>    <dbl>    <dbl>  <dbl>   <dbl>
-0.110   -0.0393  0.0332  0.134  -0.0486  -0.0791  -0.240 -0.0894
 0.0403  -0.0486  0.0461 -0.0991  0.0825   0.0767  -0.221 -0.0110
-0.128    0.0479  0.185  -0.0894  0.0830   0.0640  -0.223 -0.208
-0.0928  -0.305   0.211  -0.0420 -0.0737   0.00635 -0.225 -0.232
-0.0505  -0.111   0.106  -0.100   0.0986   0.151   -0.227 -0.0679
 0.0112  -0.151   0.190   0.0129  0.104   -0.0977  -0.226  0.0232
-0.0840   0.0320  0.0684 -0.154   0.120    0.0728  -0.229  0.0260
-0.130   -0.212   0.132   0.0879 -0.0928  -0.0991  -0.224  0.0503
-0.0796  -0.125   0.0562  0.0801 -0.00525 -0.0171  -0.232  0.0167
-0.0400   0.0522  0.122  -0.0593  0.0442   0.0107  -0.225  0.0564
```

This has* topology and geometry…

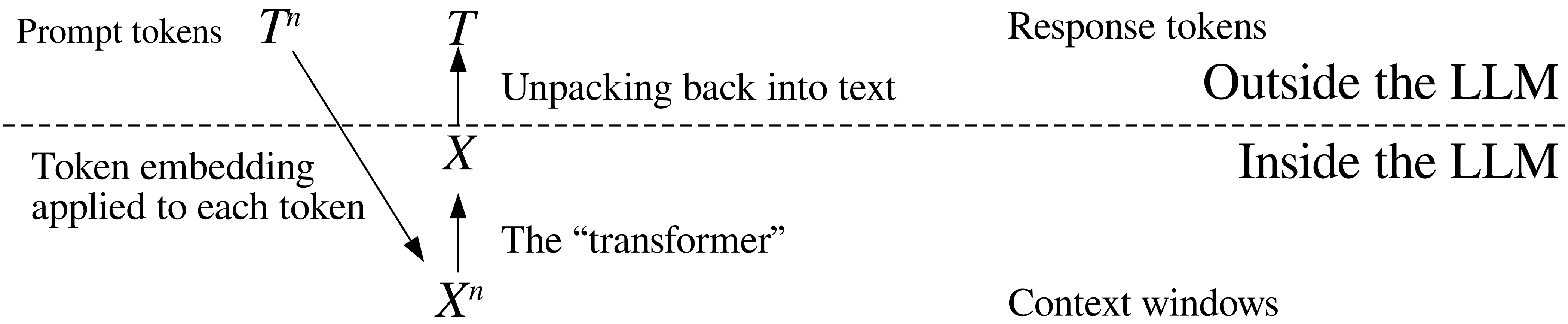
… this doesn’t!

*Euclidean metric, cosine metric, among options

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Generation of text is iteration

T : set of tokens
X : $\mathbb{R}^d$ latent space



Prompt tokens T^n

T

Response tokens

Unpacking back into text

Outside the LLM

X

Inside the LLM

Token embedding applied to each token

The “transformer”

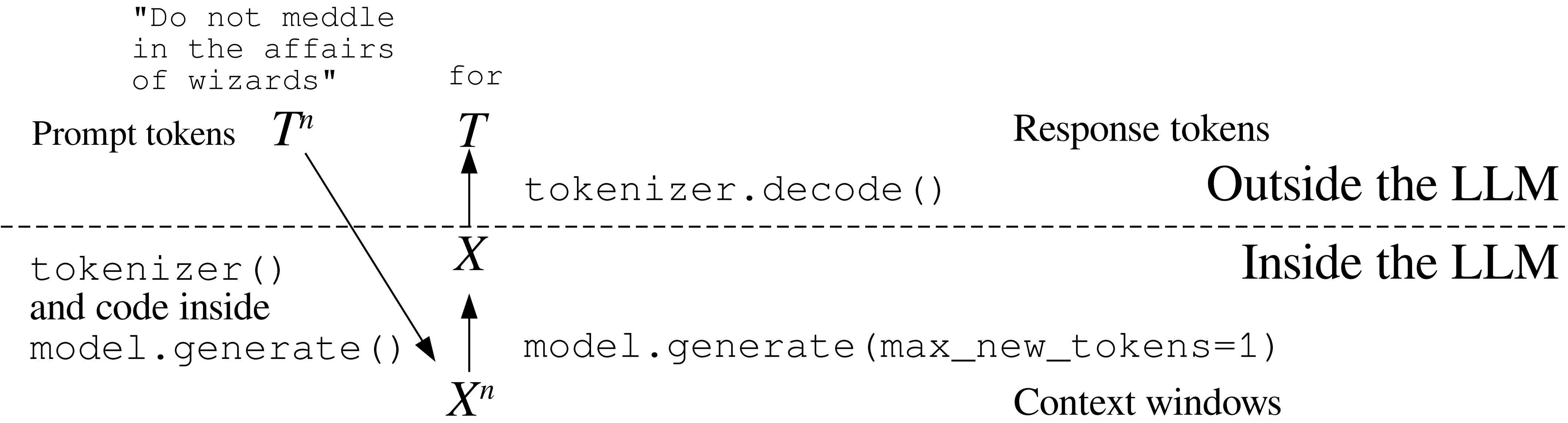
X^n

Context windows

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Generation of text is iteration

T : set of tokens
X : $\mathbb{R}^d$ latent space



```
"Do not meddle
in the affairs
of wizards"
```

`for`

Prompt tokens T^n

T

Response tokens

`tokenizer.decode()`

Outside the LLM

X

Inside the LLM

`tokenizer()`
and code inside
`model.generate()`

`model.generate(max_new_tokens=1)`

X^n

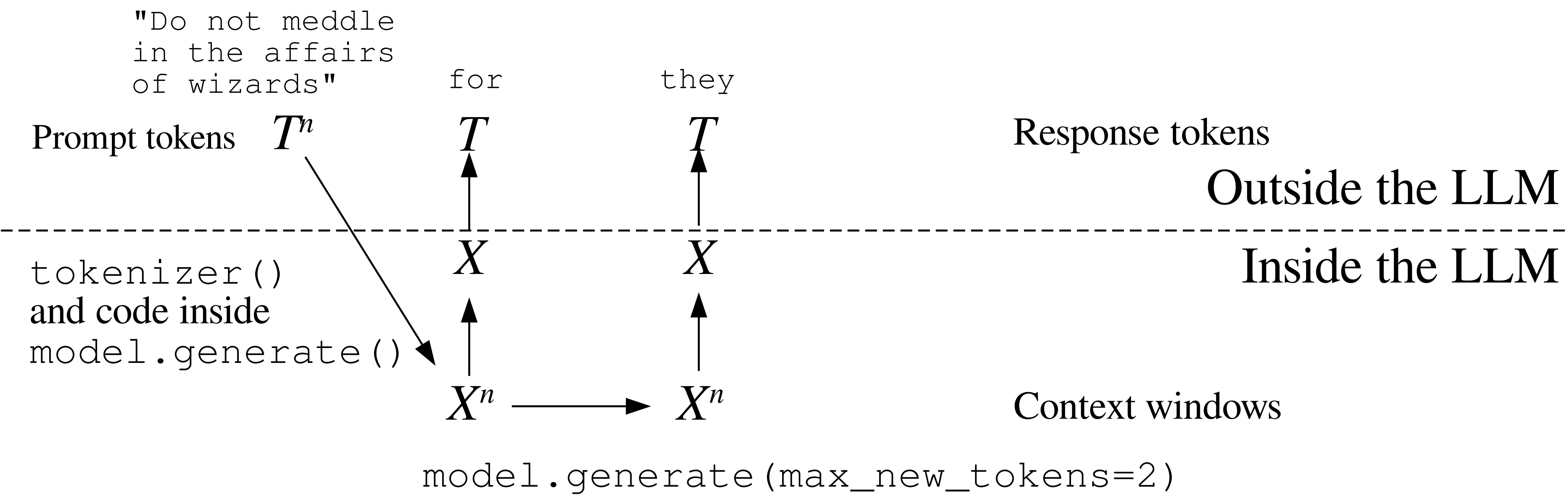
Context windows

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Generation of text is iteration

T : set of tokens

X : $\mathbb{R}^d$ latent space



"Do not meddle in the affairs of wizards"

for

they

Prompt tokens T^n

T

T

Response tokens

Outside the LLM

X

X

Inside the LLM

`tokenizer()` and code inside `model.generate()`

X^n → X^n

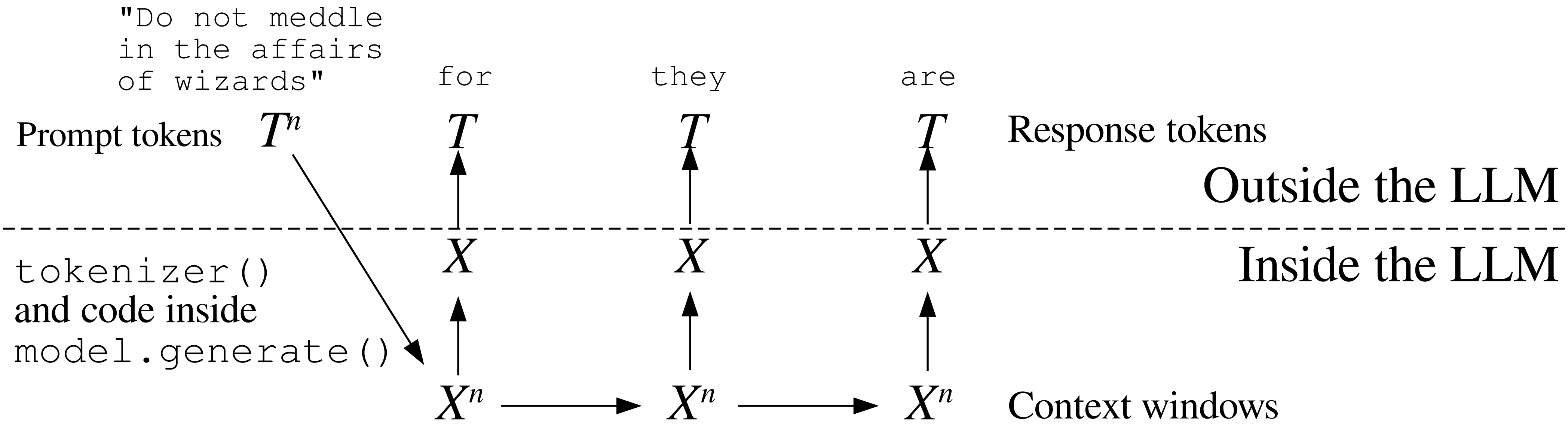
Context windows

`model.generate(max_new_tokens=2)`

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Generation of text is iteration

T : set of tokens
X : $\mathbb{R}^d$ latent space



"Do not meddle
in the affairs
of wizards"

for

they

are

Prompt tokens T^n

T T T Response tokens

Outside the LLM

Inside the LLM

X X X

tokenizer()
and code inside
model.generate()

$X^n \longrightarrow X^n \longrightarrow X^n$ Context windows

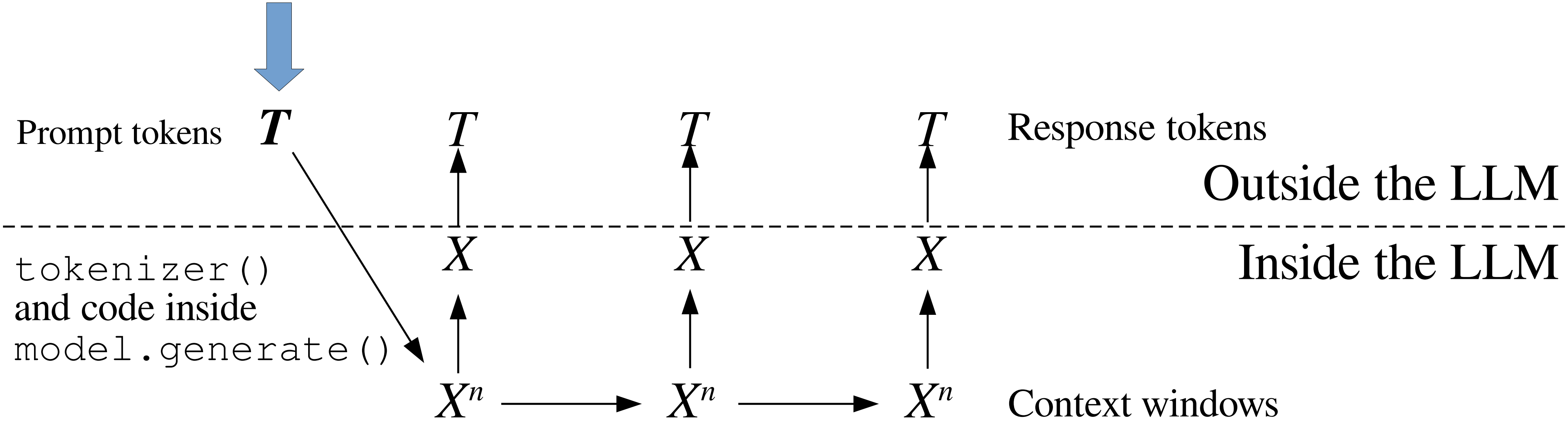
`model.generate(max_new_tokens=3)`

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the
views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Intermission… single token prompts

Single tokens can be prompts

T : set of tokens
X : $\mathbb{R}^d$ latent space



Prompt tokens T

T T T Response tokens

Outside the LLM

Inside the LLM

`tokenizer()`
and code inside
`model.generate()`

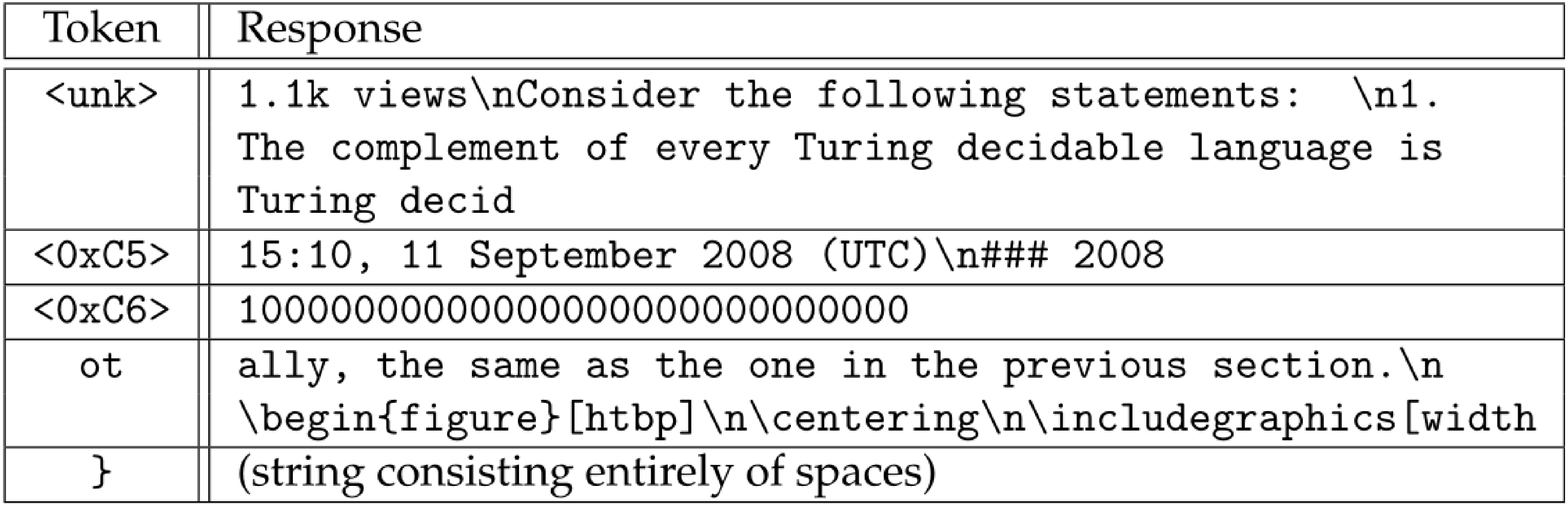
X X X

X^n X^n X^n Context windows

```
model.generate(max_new_tokens=3)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Responses to single tokens



Token	Response
`<unk>`	`1.1k views\nConsider the following statements:  \n1. The complement of every Turing decidable language is Turing decid`
`<0xC5>`	`15:10, 11 September 2008 (UTC)\n### 2008`
`<0xC6>`	`1000000000000000000000000000`
`ot`	`ally, the same as the one in the previous section.\n \begin{figure}[htbp]\n\centering\n\includegraphics[width`
`}`	(string consisting entirely of spaces)

Model: `EleutherAI/Llemma7B`

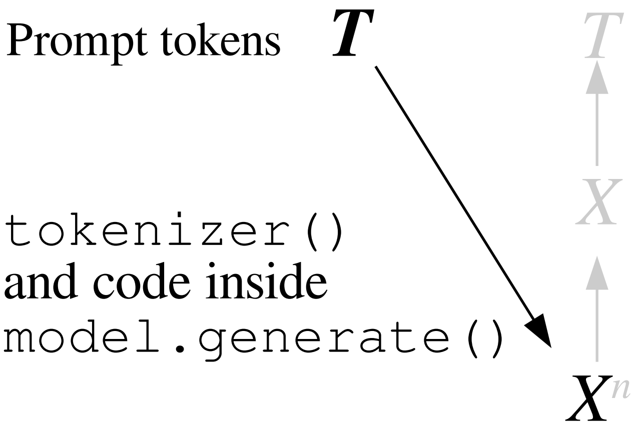
This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Reconstructing token subspace

Part 1: Whitney embedding

Bounding manifolds...

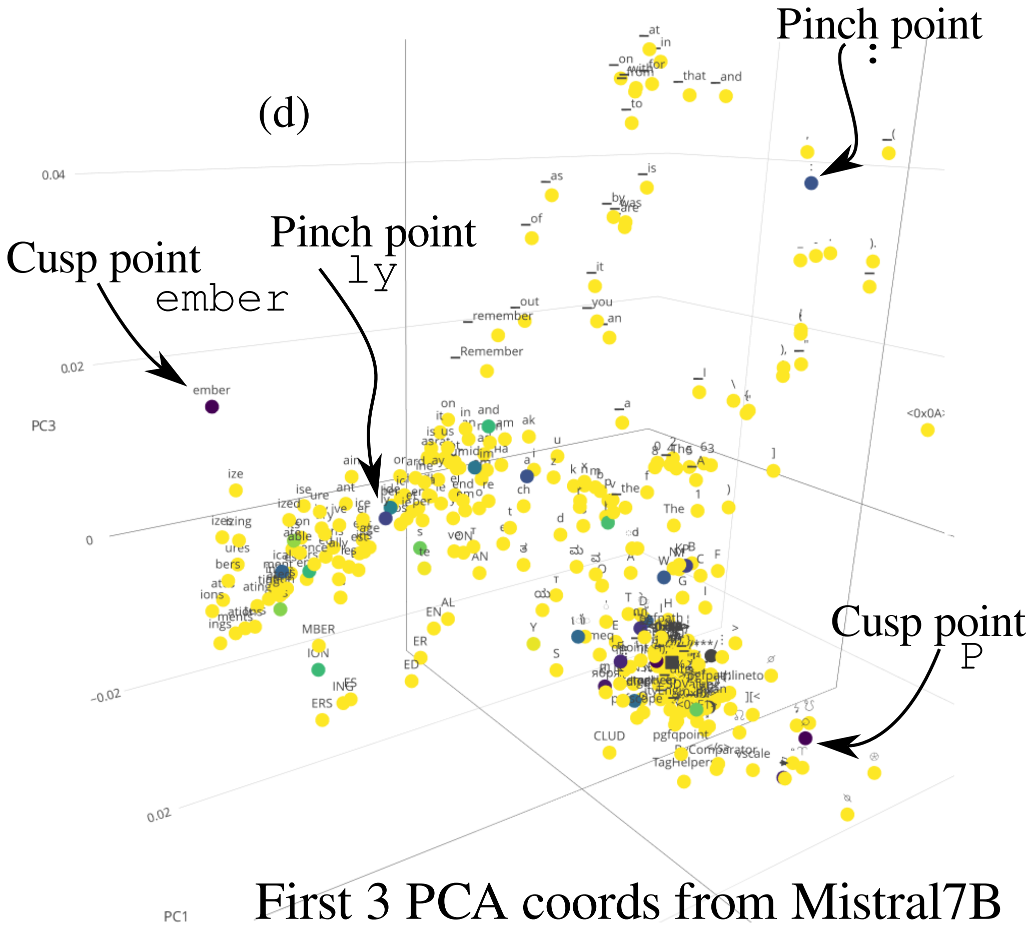
- The image of the token embedding is not* a manifold…



Prompt tokens $\boldsymbol{T}$

`tokenizer()`
and code inside
`model.generate()`

X^n



Pinch point

Pinch point
`ly`

Cusp point
`ember`

Cusp point
`P`

First 3 PCA coords from Mistral7B

*Robinson, Dey, Chiang, https://arxiv.org/abs/2504.01002 *NeurIPS* 2025

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Bounding manifolds...

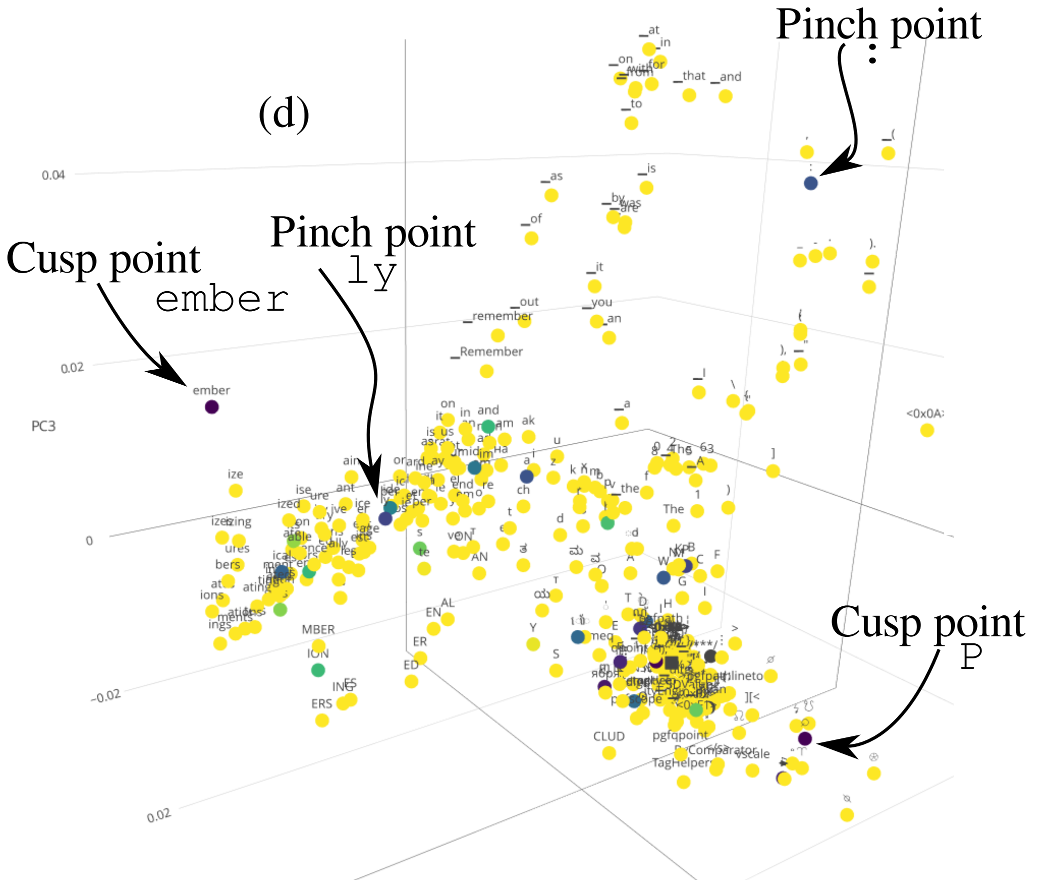
- The image of the token embedding is not* a manifold…

… but it lies within a

bounding manifold Z

$$\begin{tikzcd}
\text{Prompt tokens } \boldsymbol{T} \arrow[d, "\text{embedding}"'] & T \\
Z \arrow[dr, "\text{inclusion}"'] & X \arrow[u] \\
& X^n \arrow[u]
\end{tikzcd}$$

dim Z is typically around 30



Pinch point
:

Pinch point
ly

Cusp point
ember

Cusp point
P

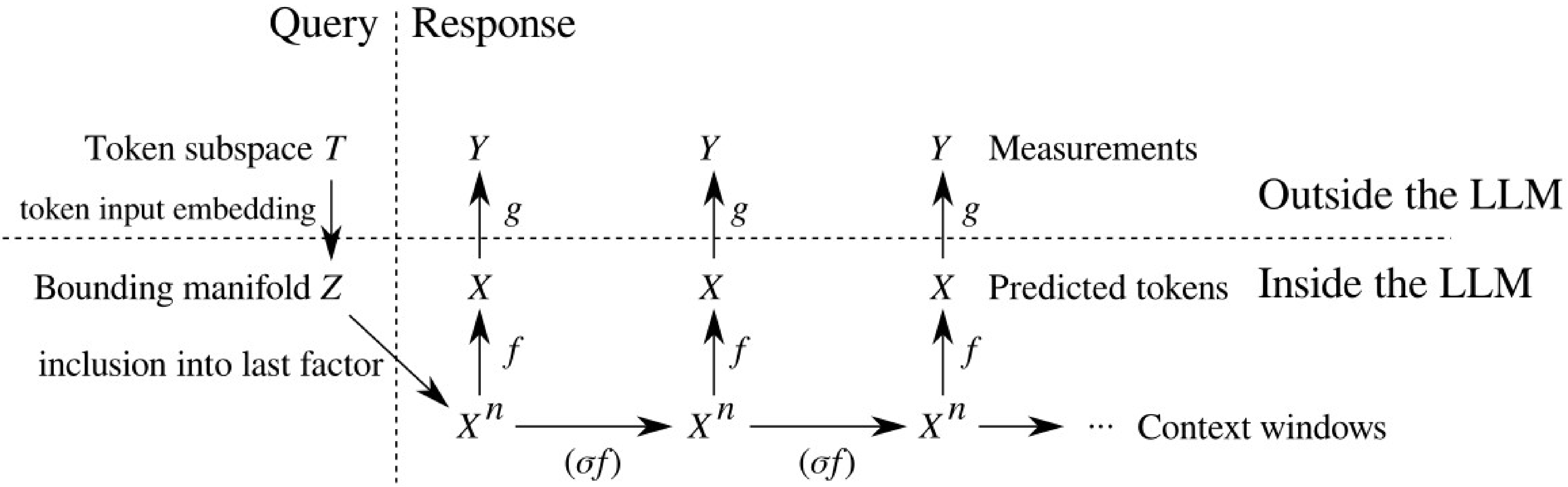
First 3 PCA coords from Mistral7B

*Robinson, Dey, Chiang, https://arxiv.org/abs/2504.01002 *NeurIPS* 2025

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Michael Robinson

Data pipeline



Query

Response

Token subspace T

token input embedding

Bounding manifold Z

inclusion into last factor

Y Y Y Measurements

g g g

Outside the LLM

X X X Predicted tokens

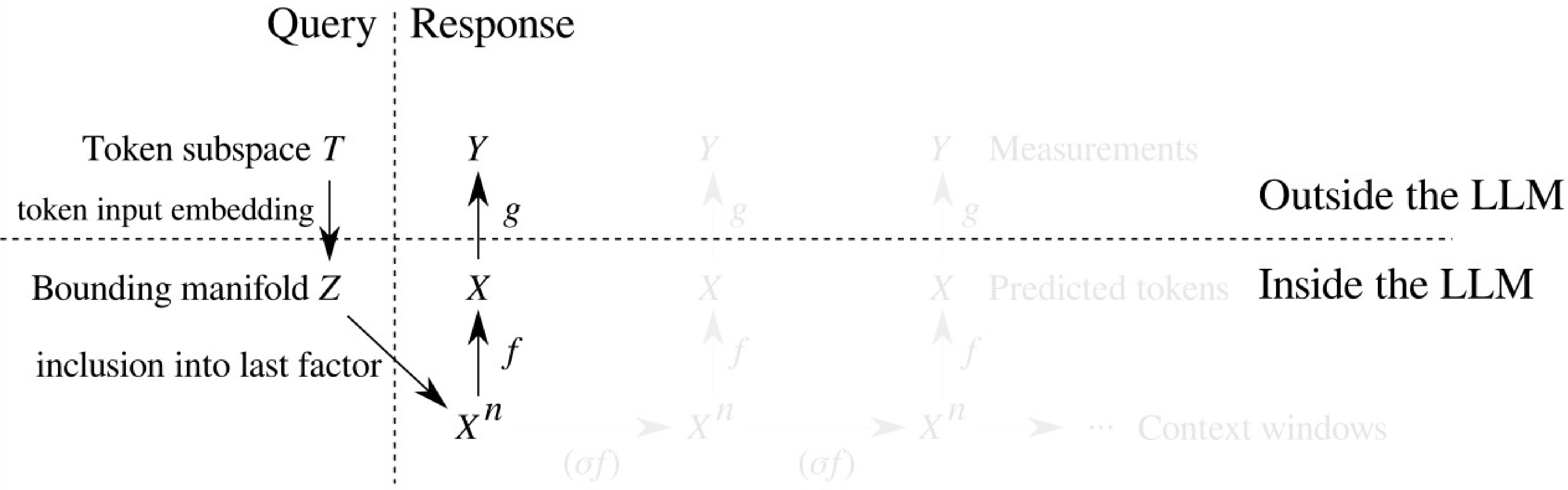
Inside the LLM

f f f

$X^n \xrightarrow{(\sigma f)} X^n \xrightarrow{(\sigma f)} X^n \rightarrow \cdots$ Context windows

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Data pipeline: just one response token



Query

Response

Token subspace T

token input embedding

Bounding manifold Z

inclusion into last factor

Y

g

X

f

X^n

(σf)

Measurements

Predicted tokens

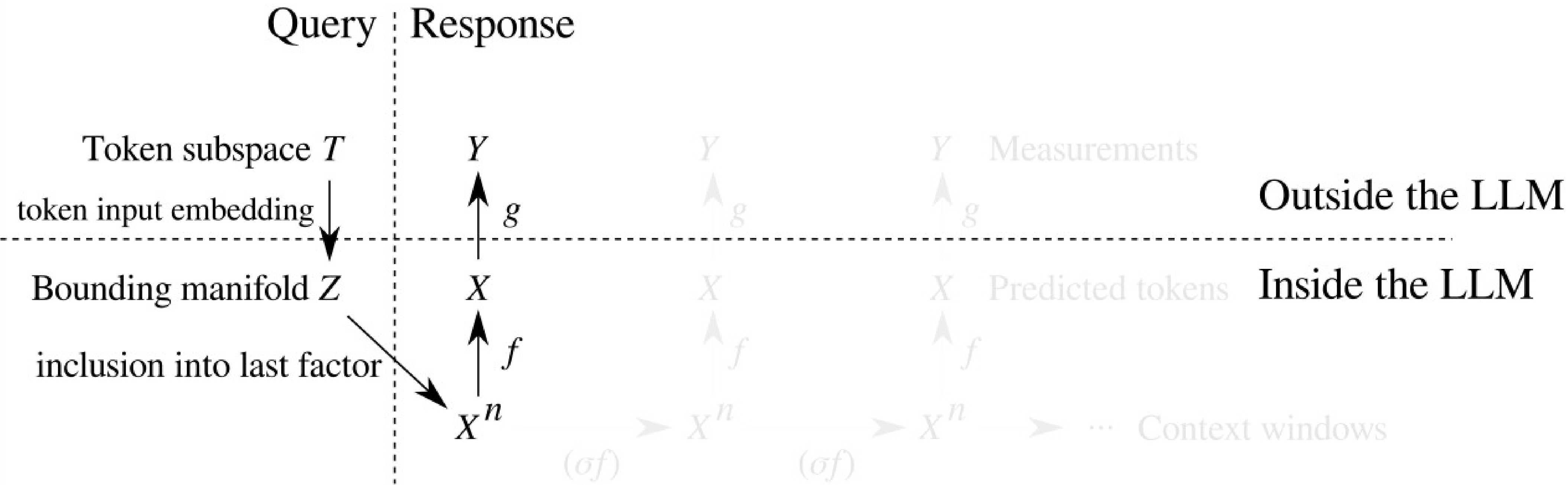
Context windows

Outside the LLM

Inside the LLM

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

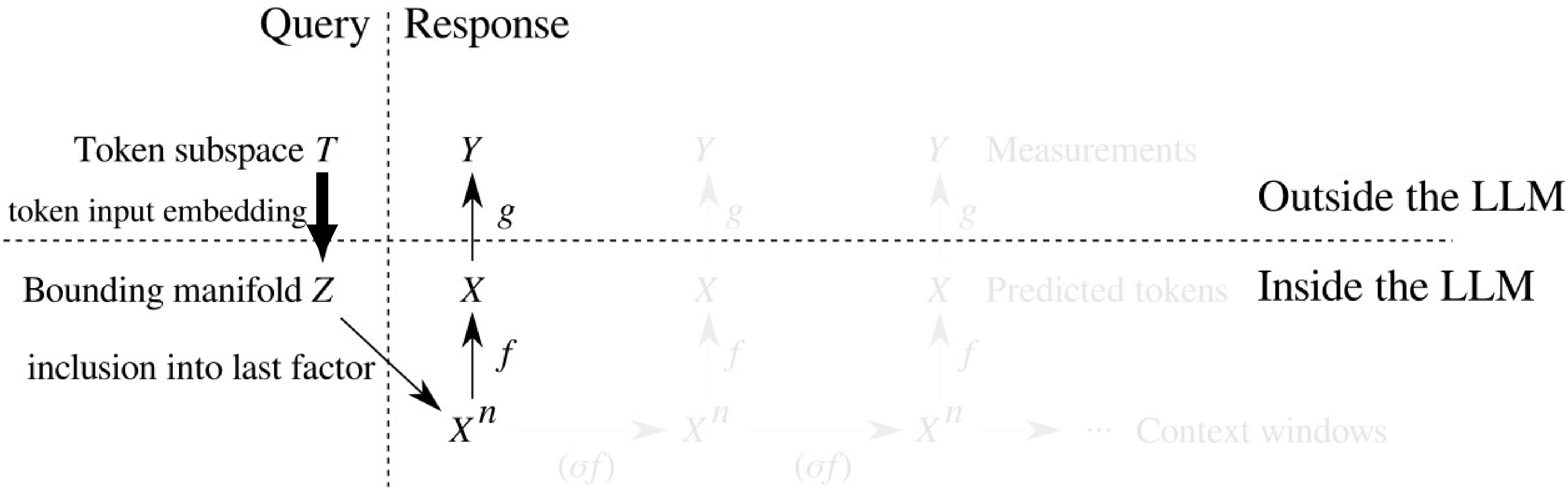
If we have an “open weights model”



```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained(“yourtokenizer”)
model = AutoCausalModelForLM.from_pretrained(“yourmodelhere”)
inputs = tokenizer(“a”,return_tensors=”pt”)
outputs = model.generate(inputs,max_new_tokens=1,
               return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

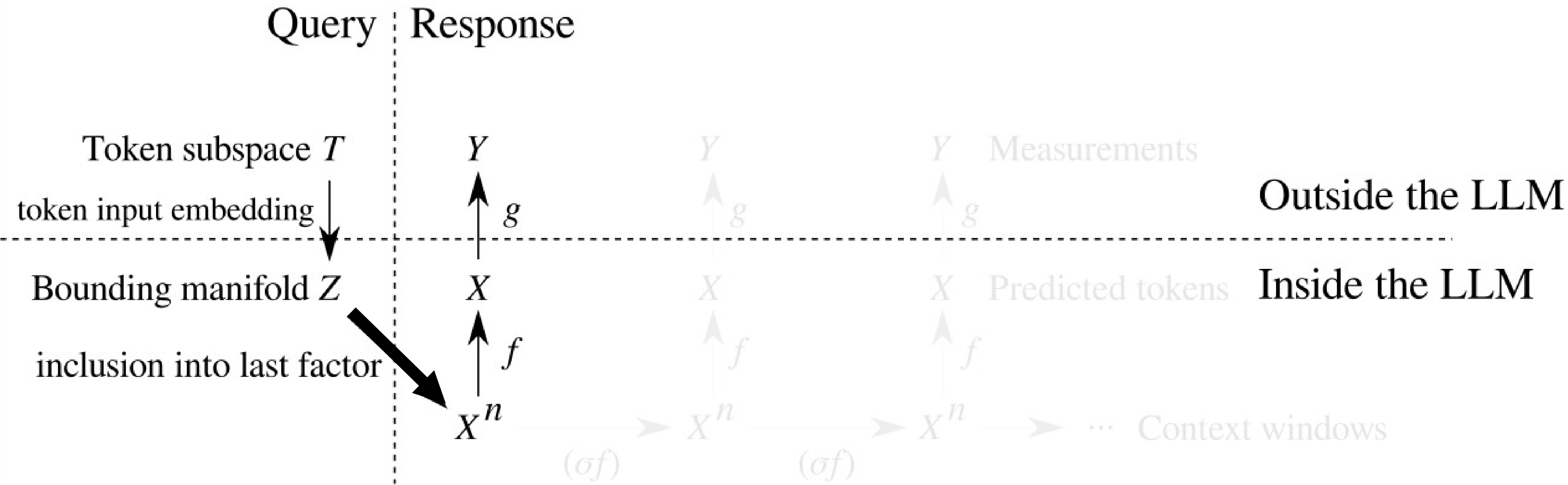
Text to tokens...



```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained(“yourtokenizer”)
model = AutoCausalModelForLM.from_pretrained(“yourmodelhere”)
inputs = tokenizer(“a”,return_tensors=”pt”)
outputs = model.generate(inputs,max_new_tokens=1,
               return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

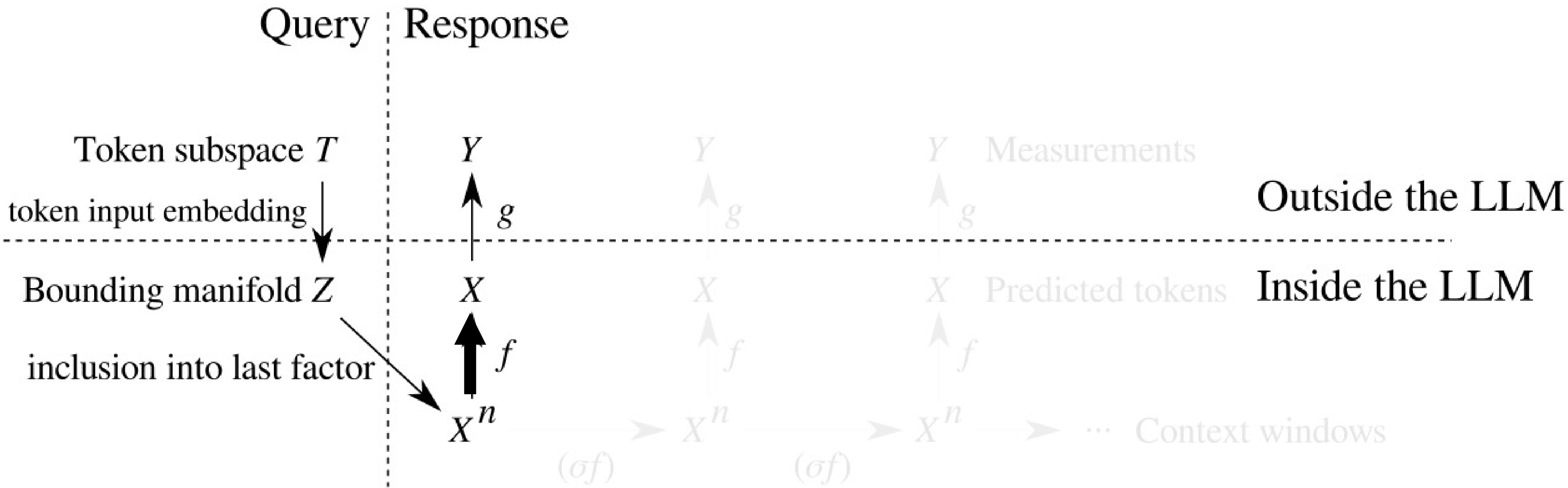
Tokens to latent space



```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained(“yourtokenizer”)
model = AutoCausalModelForLM.from_pretrained(“yourmodelhere”)
inputs = tokenizer(“a”,return_tensors=”pt”)
outputs = model.generate(inputs,max_new_tokens=1,
               return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Generate next token distribution

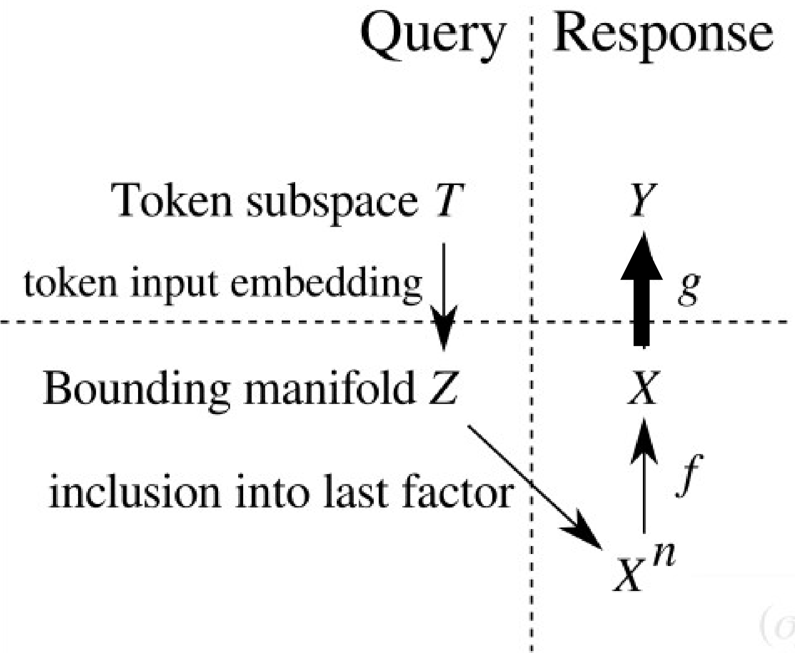


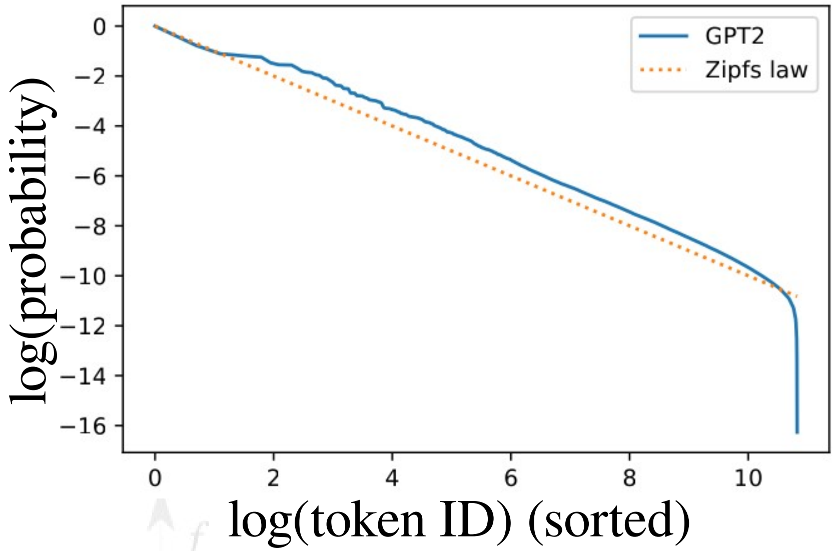
```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained(“yourtokenizer”)
model = AutoCausalModelForLM.from_pretrained(“yourmodelhere”)
inputs = tokenizer(“a”,return_tensors=”pt”)
outputs = model.generate(inputs,max_new_tokens=1,
               return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Final outputs are generation probabilities*

*which may not reflect the training data distribution! arXiv:2401.17377





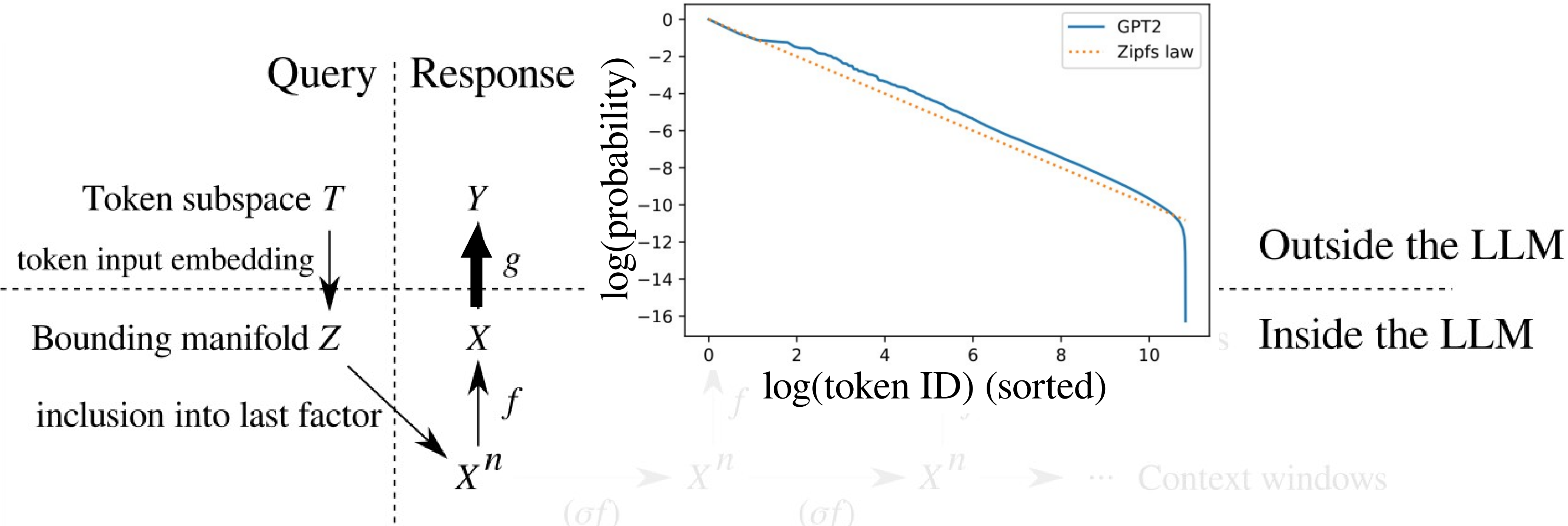
```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained(“yourtokenizer”)
model = AutoCausalModelForLM.from_pretrained(“yourmodelhere”)
inputs = tokenizer(“a”,return_tensors=”pt”)
outputs = model.generate(inputs,max_new_tokens=1,
                         return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Michael Robinson

Apply Whitney embedding theorem

2 dim X = 8192 ≤ dim Y = 32016 … g *is* an embedding

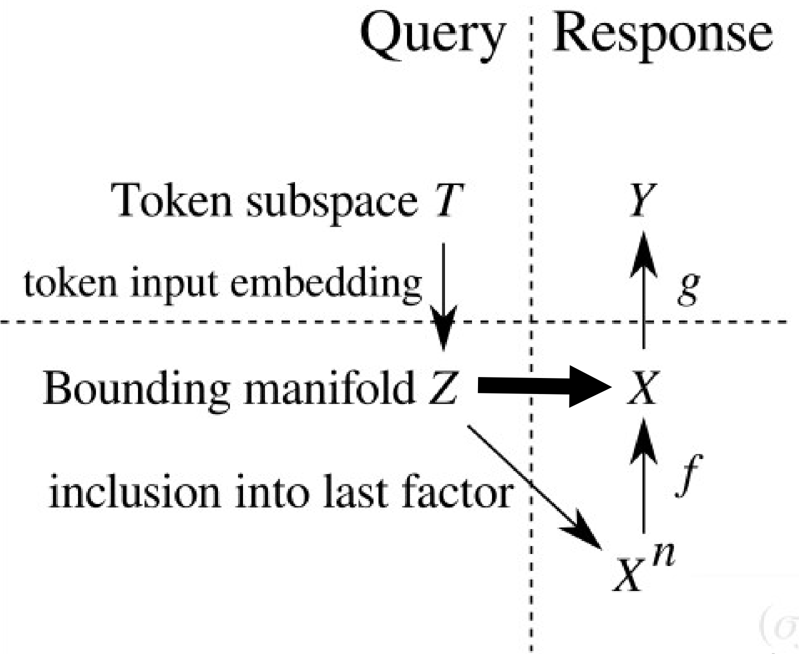


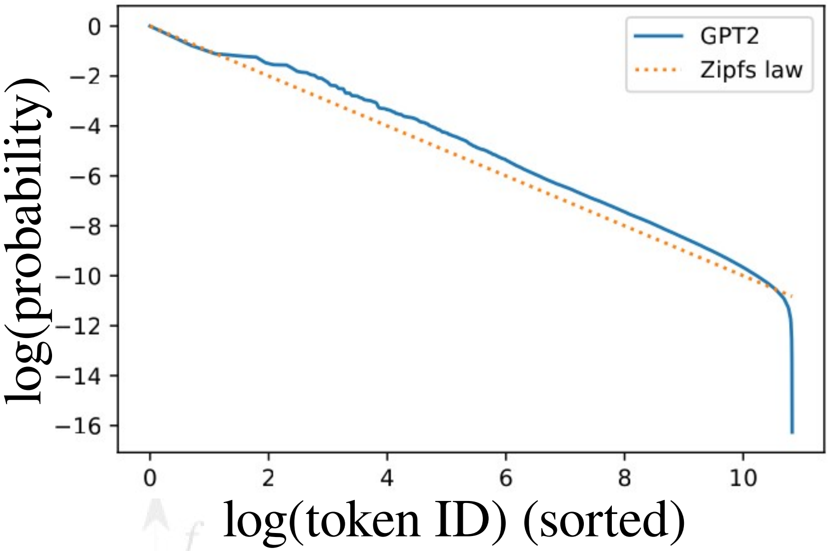
```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("yourtokenizer")
model = AutoCausalModelForLM.from_pretrained("yourmodelhere")
inputs = tokenizer("a",return_tensors="pt")
outputs = model.generate(inputs,max_new_tokens=1,
               return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Apply Whitney again

2 dim $Z = 58 \leq 4096 = \dim X$ … another embedding!





Outside the LLM

Inside the LLM

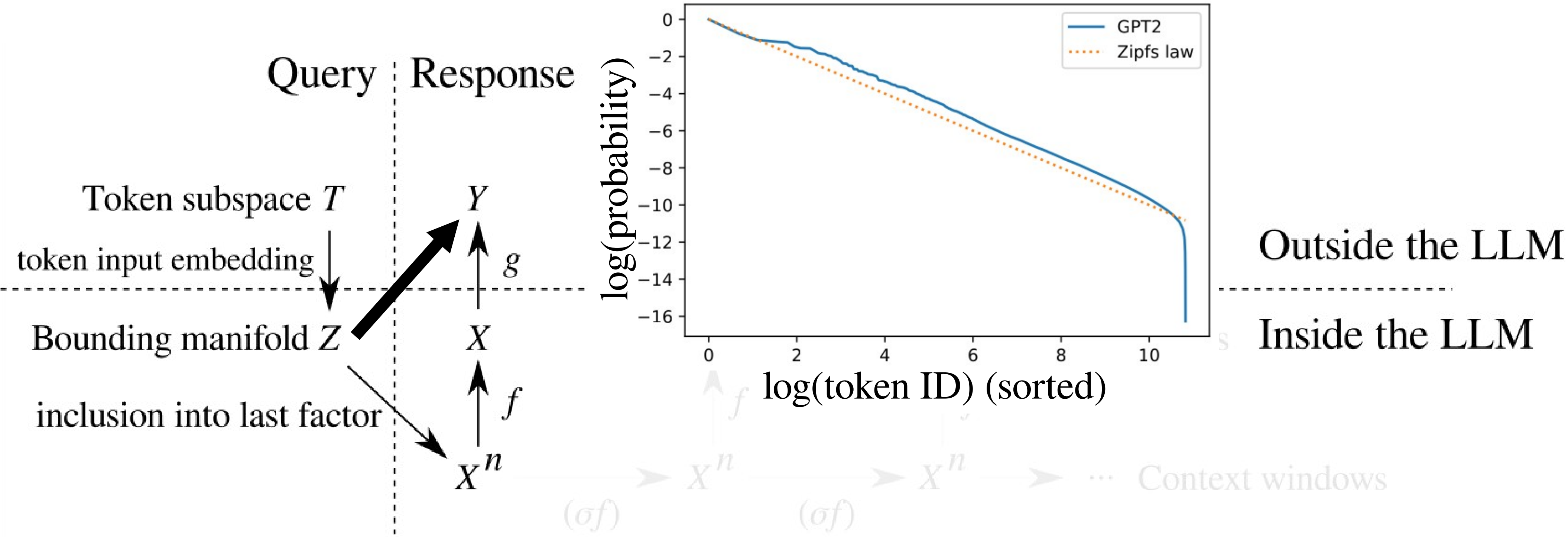
```
from transformers import AutoCausalModelForLM,AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("yourtokenizer")
model = AutoCausalModelForLM.from_pretrained("yourmodelhere")
inputs = tokenizer("a",return_tensors="pt")
outputs = model.generate(inputs,max_new_tokens=1,
              return_dict_in_generate=True, output_scores=True)
probs = outputs.scores[0].softmax(-1)
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Michael Robinson

Apply Whitney again

$2 \dim Z = 58 \leq 4096 = \dim X$ … another embedding!



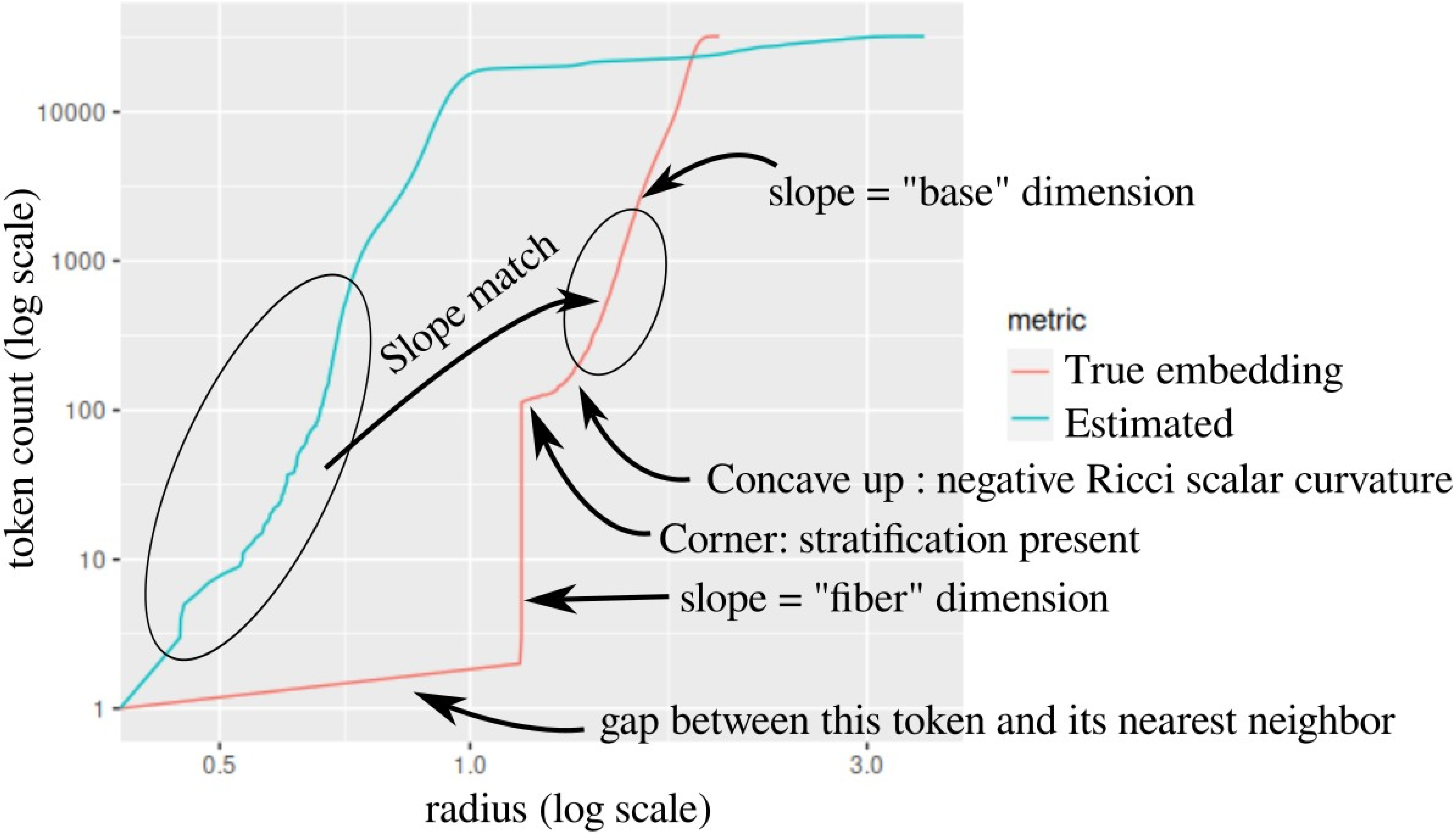
Procedure:

For each token, just use the next token distribution **as** its coordinates.
This recovers original coordinates up to diffeomorphism

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Test results: dimension* recovered

- Dimension as a proxy for homeomorphism...



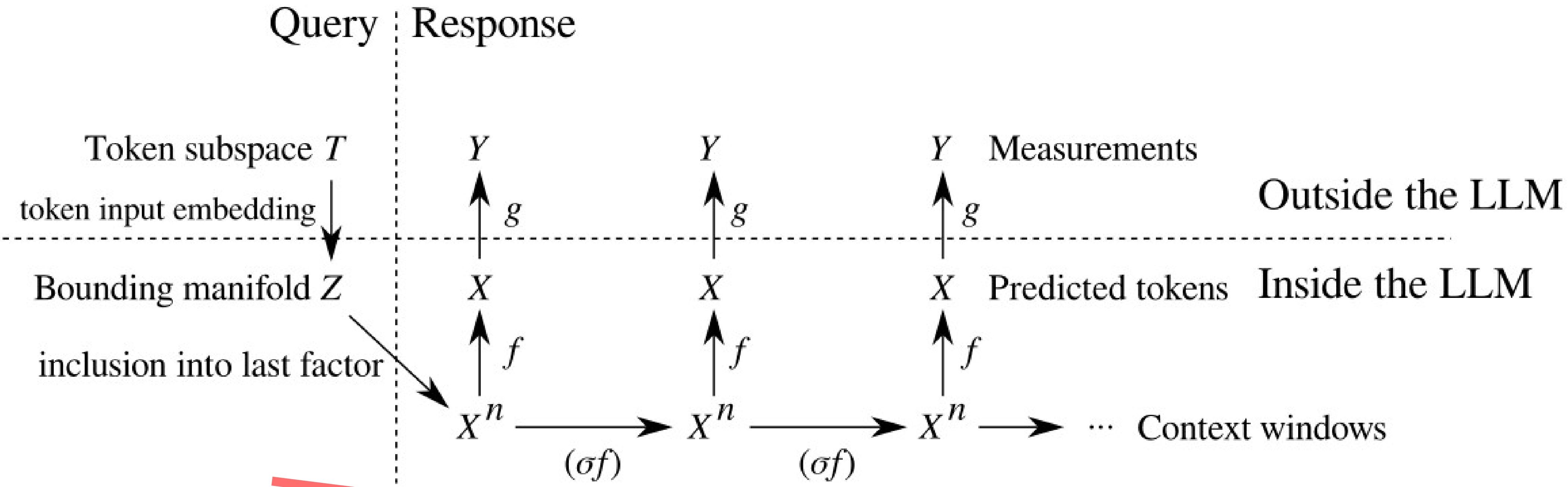
*Some tokens don’t have well-defined dimension

Reconstructing token subspace

Part 2: Partial views of sliding windows

Data pipeline

- Motivation: no direct access to the LLM “insides”

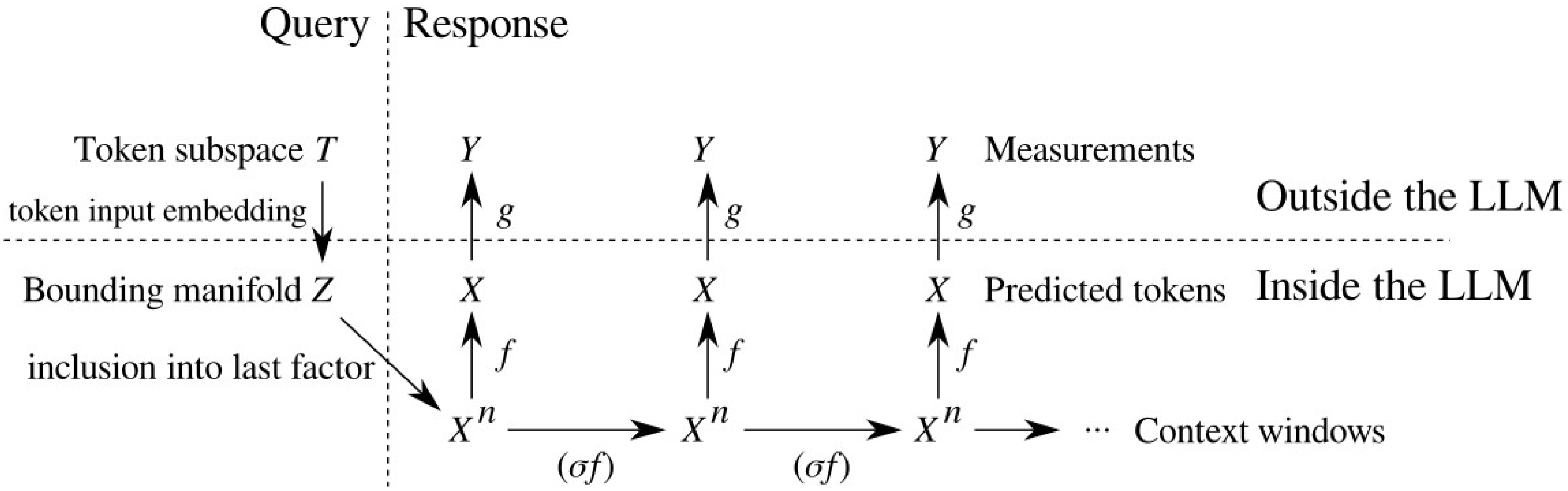


```
~~from transformers import AutoCausalModelForLM,AutoTokenizer~~
~~tokenizer = AutoTokenizer.from_pretrained(“yourtokenizer”)~~
~~model = AutoCausalModelForLM.from_pretrained(“yourmodelhere”)~~
```

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Data pipeline

- Motivation: no direct access to the LLM “insides”



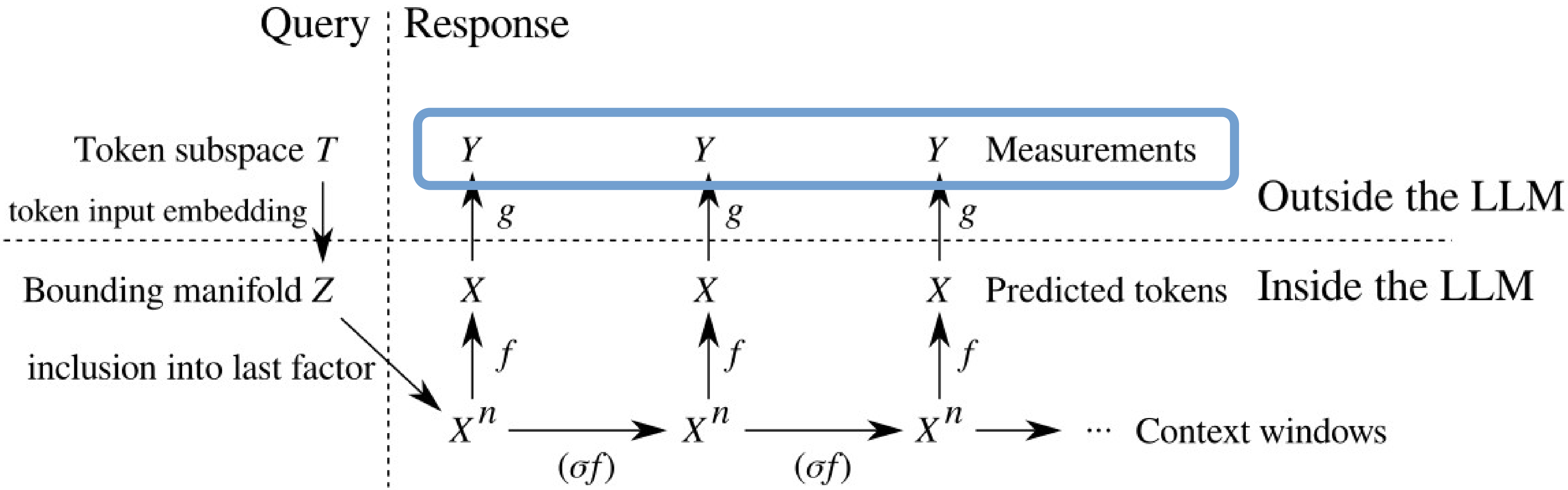
```
import ollama
response=ollama.generate(model=“yourmodel”,
                         prompt=“a”,
                         options={‘num_predict’ : m})
```

NB: Yes, `ollama` is for open source models, but proprietary APIs look similar, as does `transformers.pipeline()`.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Data pipeline

- Instead: Limited measurement taken from response



Query

Response

Token subspace T

token input embedding

Bounding manifold Z

inclusion into last factor

Y Y Y Measurements

Outside the LLM

g g g

X X X Predicted tokens

Inside the LLM

f f f

$X^n \xrightarrow{(\sigma f)} X^n \xrightarrow{(\sigma f)} X^n \longrightarrow \cdots$ Context windows

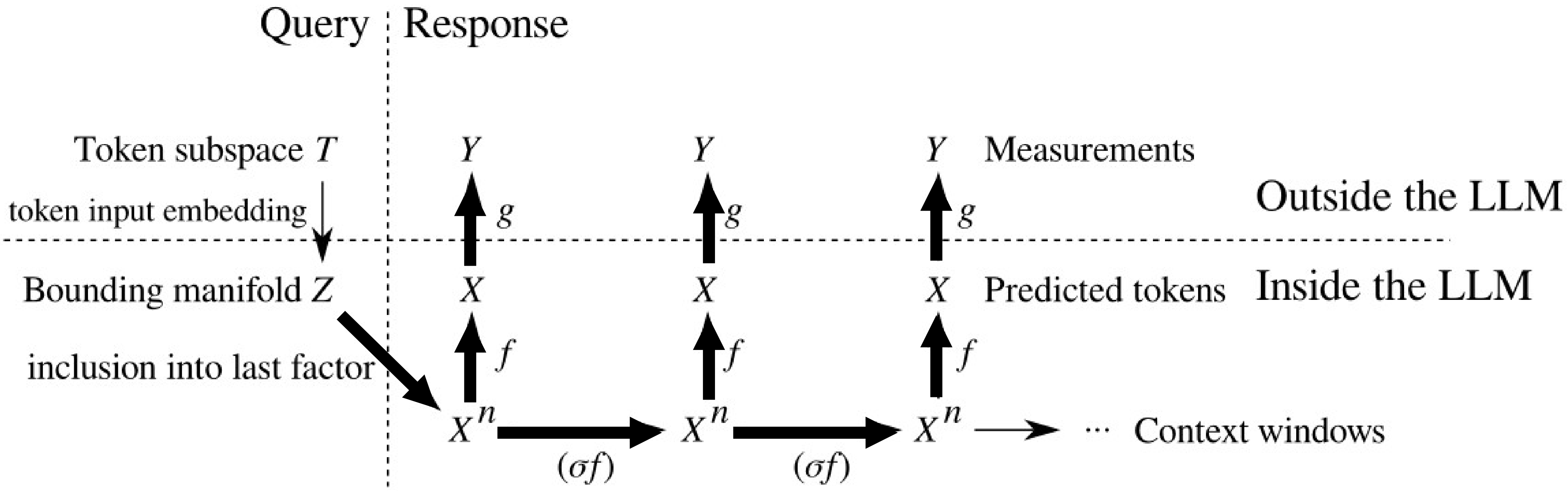
```
import ollama
response=ollama.generate(model=“yourmodel”,
                                prompt=“a”,
                                options={‘num_predict’ : m})
```

NB: Yes, `ollama` is for open source models, but proprietary APIs look similar, as does `transformers.pipeline()`.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Main theorem

- Theorem: $Z \to Y^m$ is a generically an embedding if m is large enough



Query | Response

Token subspace T

token input embedding

Bounding manifold Z

inclusion into last factor

Y g Y g Y g Measurements

Outside the LLM

X f X f X f Predicted tokens

Inside the LLM

X^n (σf) X^n (σf) X^n ⟶ ··· Context windows

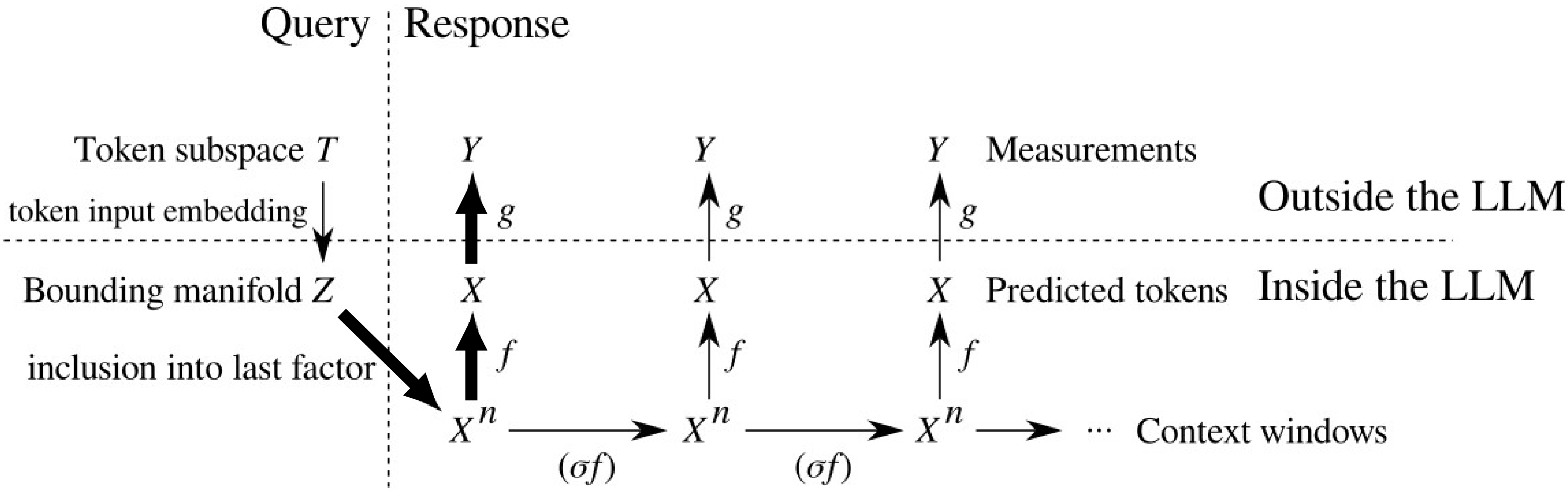
```
import ollama
response=ollama.generate(model="yourmodel",
                         prompt="a",
                         options={'num_predict' : m})
```

NB: Yes, `ollama` is for open source models, but proprietary APIs look similar, as does `transformers.pipeline()`.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Intersection submanifold

- Multiple preimages for a measurement y_1 in Y

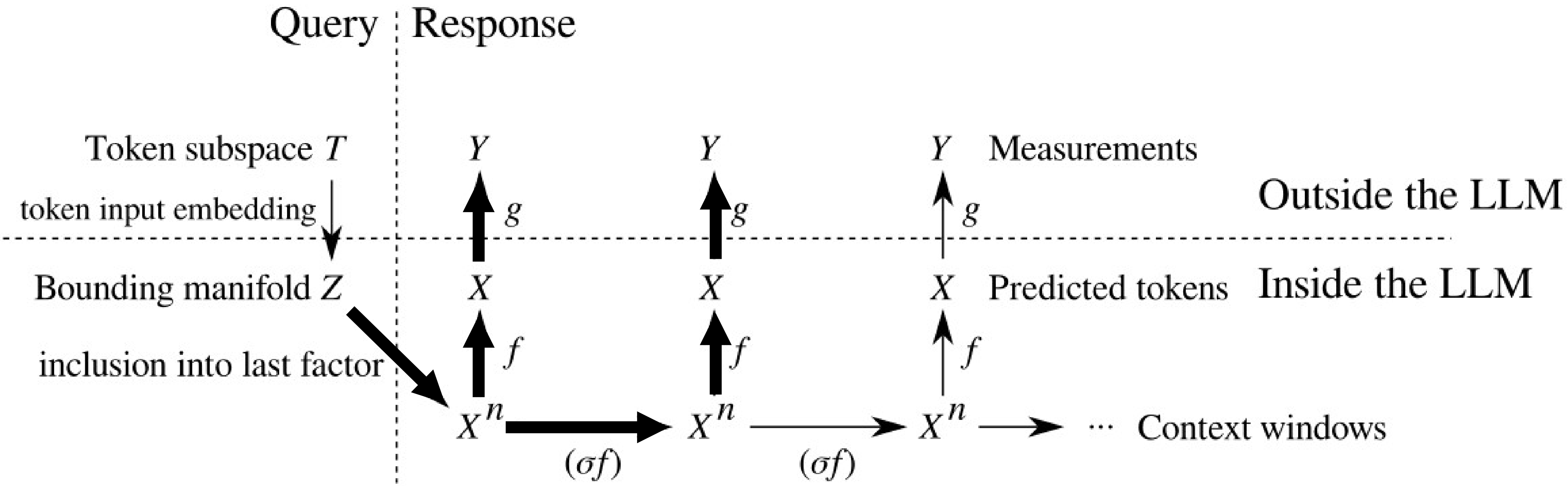


Generically $(f^{-1} \circ g^{-1})(y_1) \subseteq Z$ is a submanifold of positive codimension

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319.
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Intersection submanifold

- Fewer preimages for a sequence of measurements

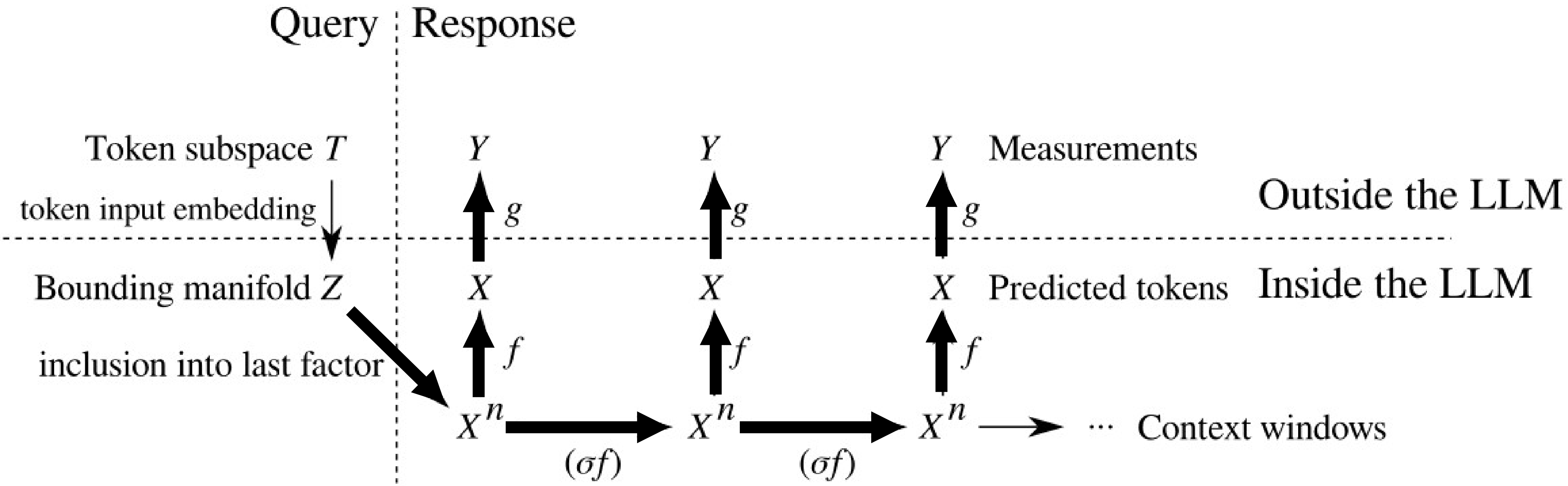


Generically $(f^{-1}\circ g^{-1})(y_1) \subseteq Z$ is a submanifold of positive codimension and $((\sigma f)^{-1}\circ f^{-1}\circ g^{-1})(y_2) \subseteq Z$ is a submanifold of positive codimension

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Intersection submanifold

- Multi-jet transversality says, “intersect enough of these and you’ll end up with an empty set!”



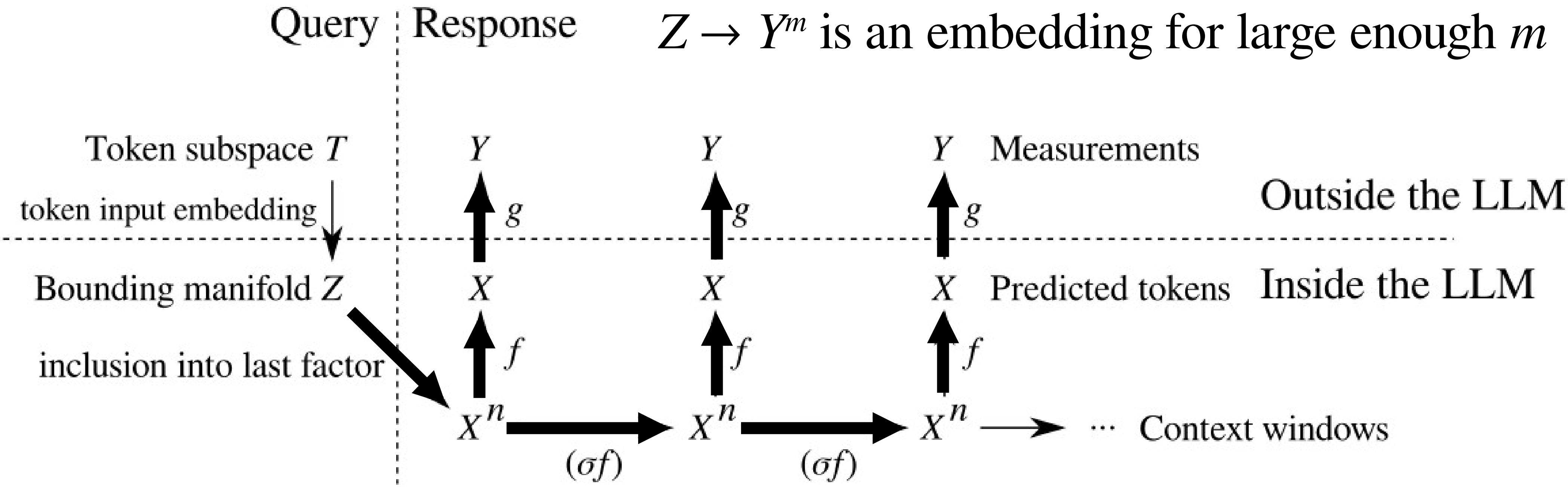
Generically $(f^{-1}\circ g^{-1})(y_1) \subseteq Z$ is a submanifold of positive codimension and $((\sigma f)^{-1}\circ f^{-1}\circ g^{-1})(y_2) \subseteq Z$ is a submanifold of positive codimension and $((\sigma f)^{-1}\circ(\sigma f)^{-1}\circ f^{-1}\circ g^{-1})(y_3) \subseteq Z$ is a submanifold of positive codimension

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Intersection submanifold

- Multi-jet transversality says, "intersect enough of these and you'll end up with an empty set!"

$Z \to Y^m$ is an embedding for large enough m

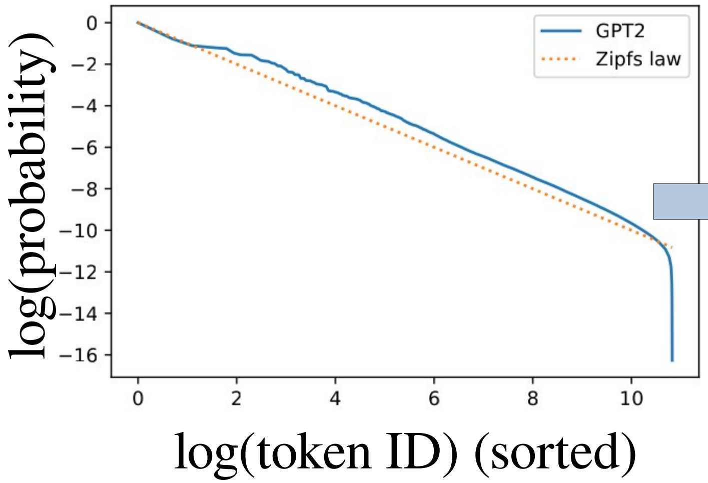


Generically $(f^{-1} \circ g^{-1})(y_1) \subseteq Z$ is a submanifold of positive codimension and $((\sigma f)^{-1} \circ f^{-1} \circ g^{-1})(y_2) \subseteq Z$ is a submanifold of positive codimension and $((\sigma f)^{-1} \circ (\sigma f)^{-1} \circ f^{-1} \circ g^{-1})(y_3) \subseteq Z$ is a submanifold of positive codimension

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Practical considerations

- “Measurements” may be slow to converge…



Typically power-law;
Many infrequent tokens
May not actually reflect the
actual next token distribution

- … but we have more options for collection

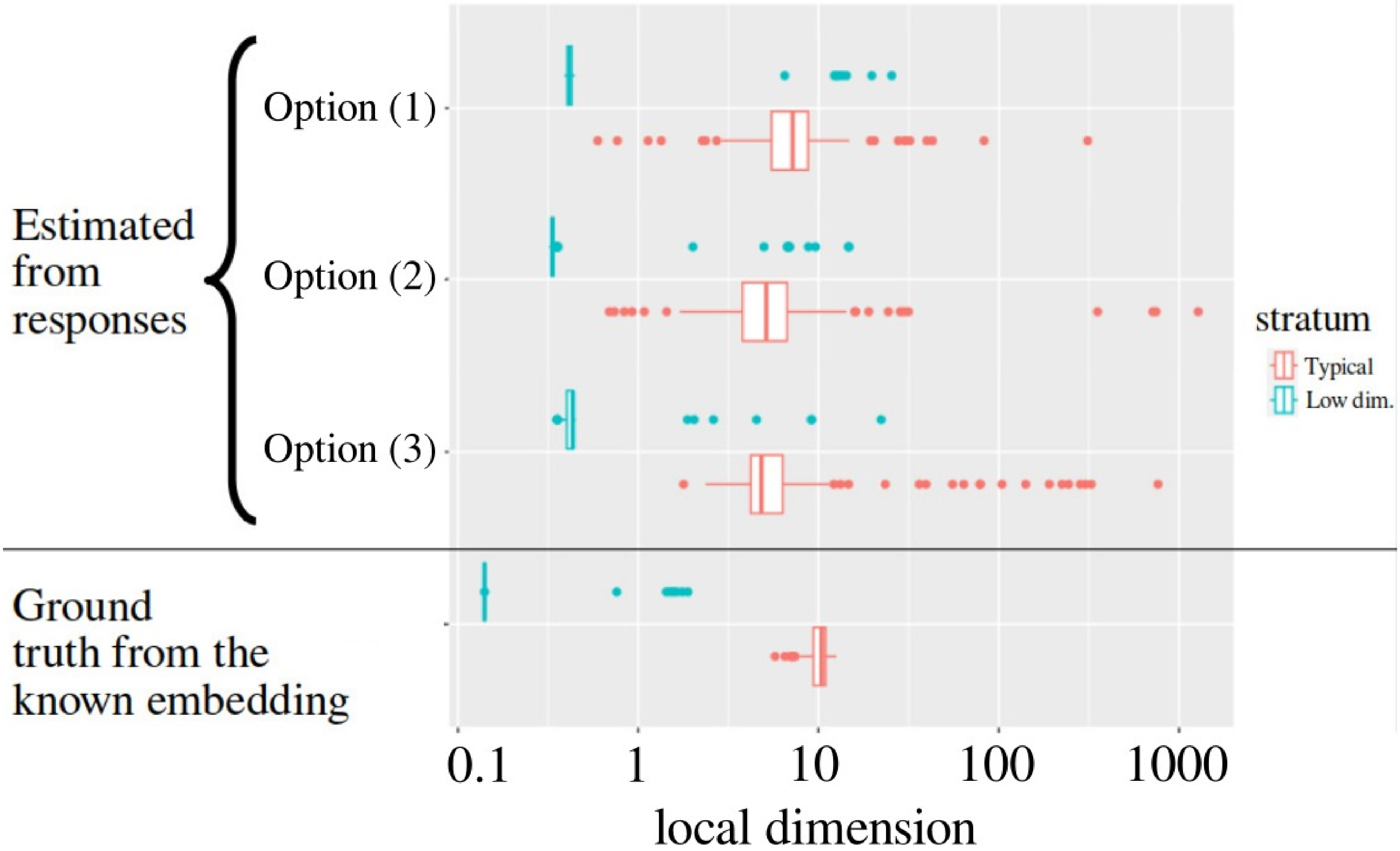
Option (1): Collect $m = 30$ response tokens and $\ell = 3$ probabilities for the top three tokens at each response token position (ignoring what the tokens actually were),

Option (2): Collect $m = 30$ response tokens and $\ell = 32016$ probabilities, one for each token, but aggregated over the entire response, and

Option (3): Collect $m = 1$ response token and $\ell = 32016$ probabilities, one for each token being the first token in the response.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

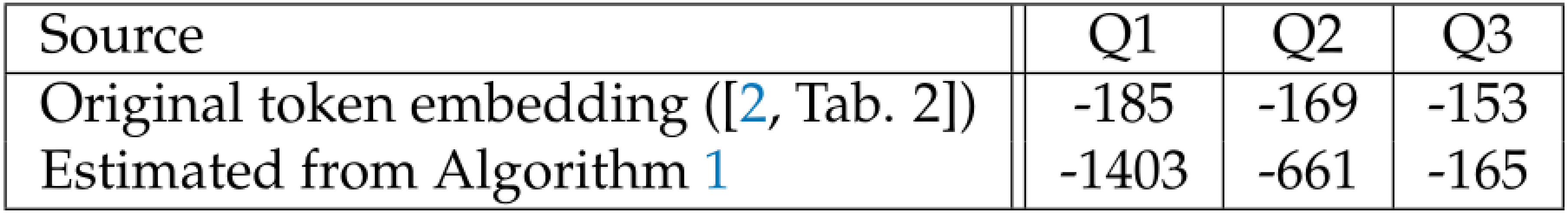
Results: dimension is preserved



This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Results: geometry is destroyed

- The distribution of Ricci scalar curvature changes



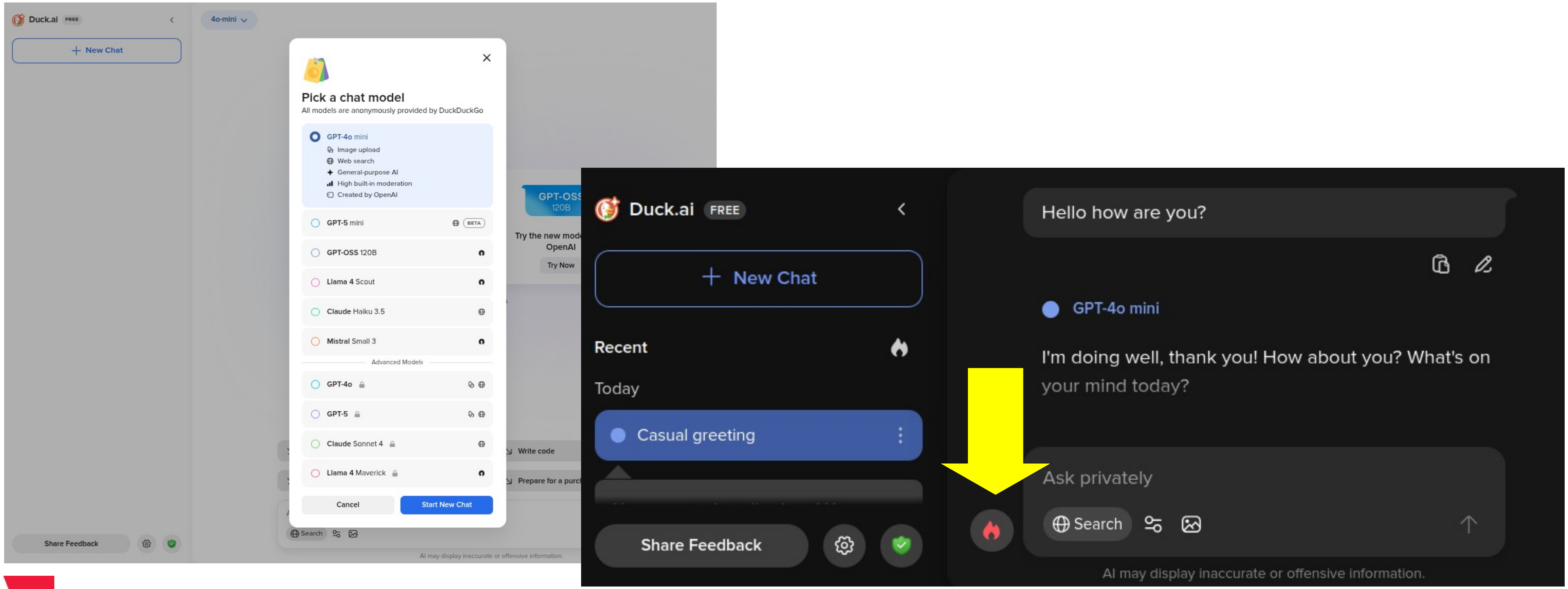
Source	Q1	Q2	Q3
Original token embedding ([2, Tab. 2])	-185	-169	-153
Estimated from Algorithm 1	-1403	-661	-165

- This is expected… the embedding coordinates have nothing to do with the transformer *f* at all!
- Caution: if you thought distances in the token embedding space were meaningful, they are not preserved...

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

If you want to try this…

- You can’t use the web interface for ChatGPT, because you don’t control the context window
- Instead, try https://duck.ai since the context is controlled by you. Delete it after every prompt



Pick a chat model

All models are anonymously provided by DuckDuckGo

● GPT-4o mini
- Image upload
- Web search
- General-purpose AI
- High built-in moderation
- Created by OpenAI

○ GPT-5 mini BETA

○ GPT-OSS 120B

○ Llama 4 Scout

○ Claude Haiku 3.5

○ Mistral Small 3

Advanced Models

○ GPT-4o

○ GPT-5

○ Claude Sonnet 4

○ Llama 4 Maverick

Hello how are you?

GPT-4o mini

I'm doing well, thank you! How about you? What's on your mind today?

AI may display inaccurate or offensive information.

Implications and next steps

- That topology can be extracted (expensively) even if the model is proprietary
- Topology of the internal representation of tokens in an LLM directly impacts its behavior
  - If the token subspace is not a manifold, gradient descent is not well defined!
  - Prompt engineering is, as a result, an artisan craft!
- What about geometry? That's next up...
- Details: https://doi.org/10.3390/math13203320
- Questions? Ask! michaelr@american.edu

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)