

# Can LLMs help with XML?

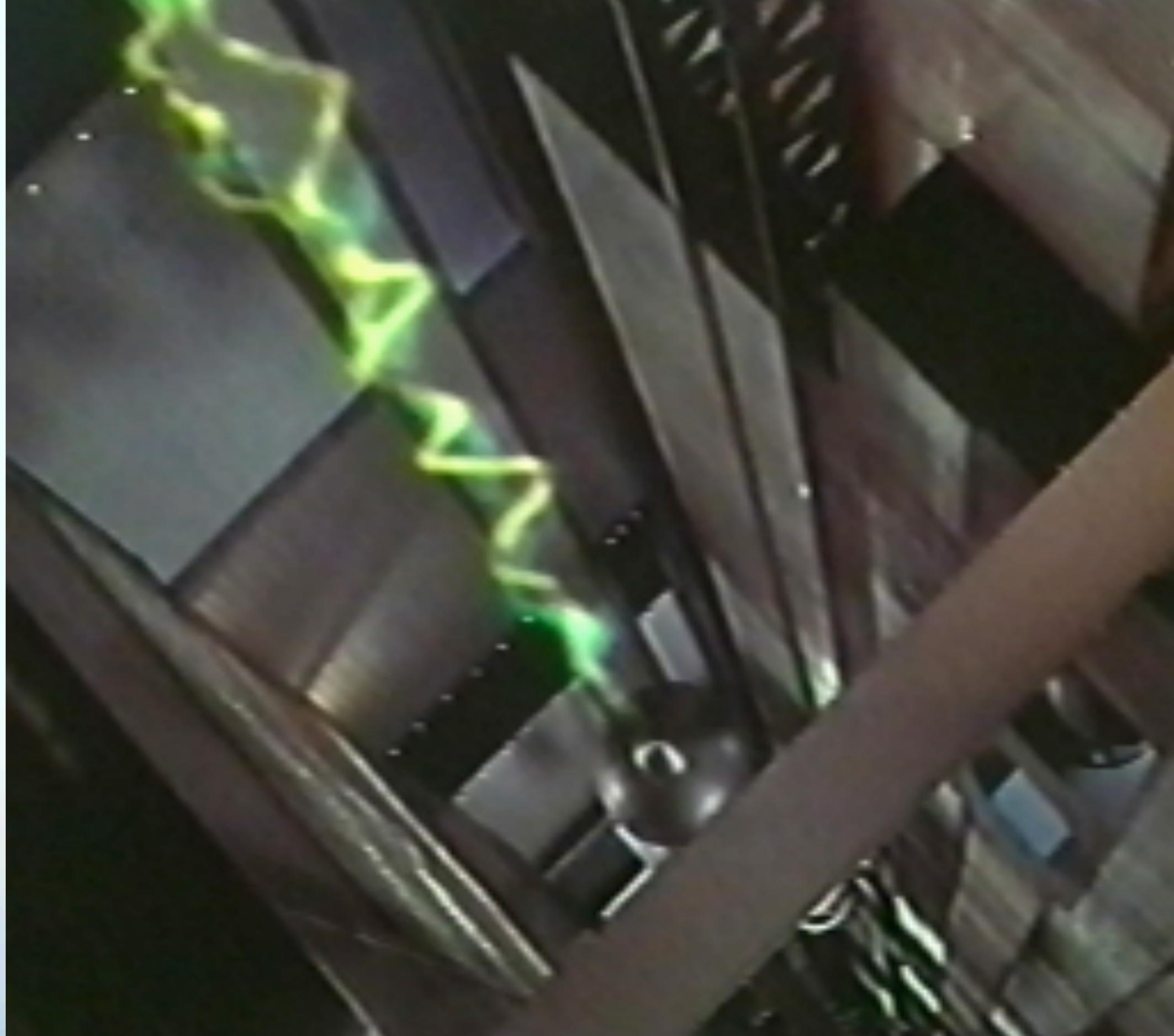
Steven J. DeRose

Balisage, July 31, 2024



# Goals

- **How** LLMs work  
(math)
- **Why** they often fail  
(epistemology)
- **Whether** they work with XML  
(markup)



# How LLMs work

## *A way-too-abridged history*

- Probability and language
- Hidden Markov Models
- Word embeddings
- Attention
- Transformers

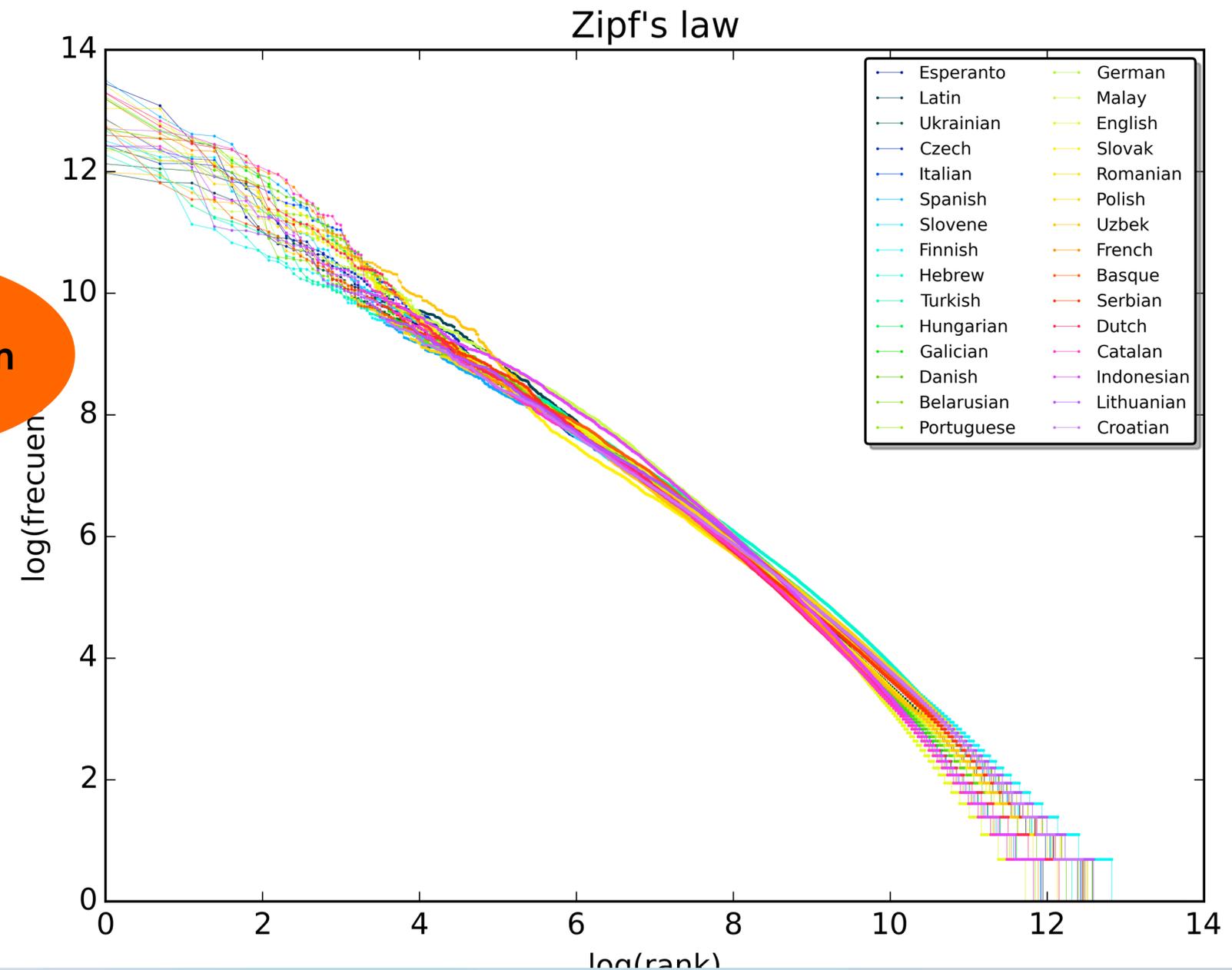
I'm omitting a second key lineage,  
from perceptrons → neural nets → DNN

# Probability and language

Shannon, Zipf, ...

- A language is a set of sentences
- The set is infinite
  - "This is very very very... true."
  - "...in the house that Jack built."
- Zipf's Law: rank vs. frequency
  - A few v. v. v. frequent items ("the")
  - A long tail of rare ones ("petrichor")
  - Same for idioms, phrases, sentences, ...
- Sampling the tail is hard, esp. for  $\infty$  sets

Why  
compression  
works



# HMMs

## From 1D to 2D probabilities

- Count character pairs: **sh** >>> **ex** >>> **qu** >>> **qz**

- Or triples,  
or quadruples,  
or...



- Use them to generate text

- (*most-likely-successor* item)

jou mouplas de monnernaissains deme us vreh bre tu de  
 toucheur dimmere lles mar elame re a ver il douvents so  
 bet ereiner sommeit sinach gan turhatt er aum wie best  
 alliender taussichelle laufurcht er blindeseit uber konn

- Can use to identify languages

- Same idea for syllables, parts of speech, words:

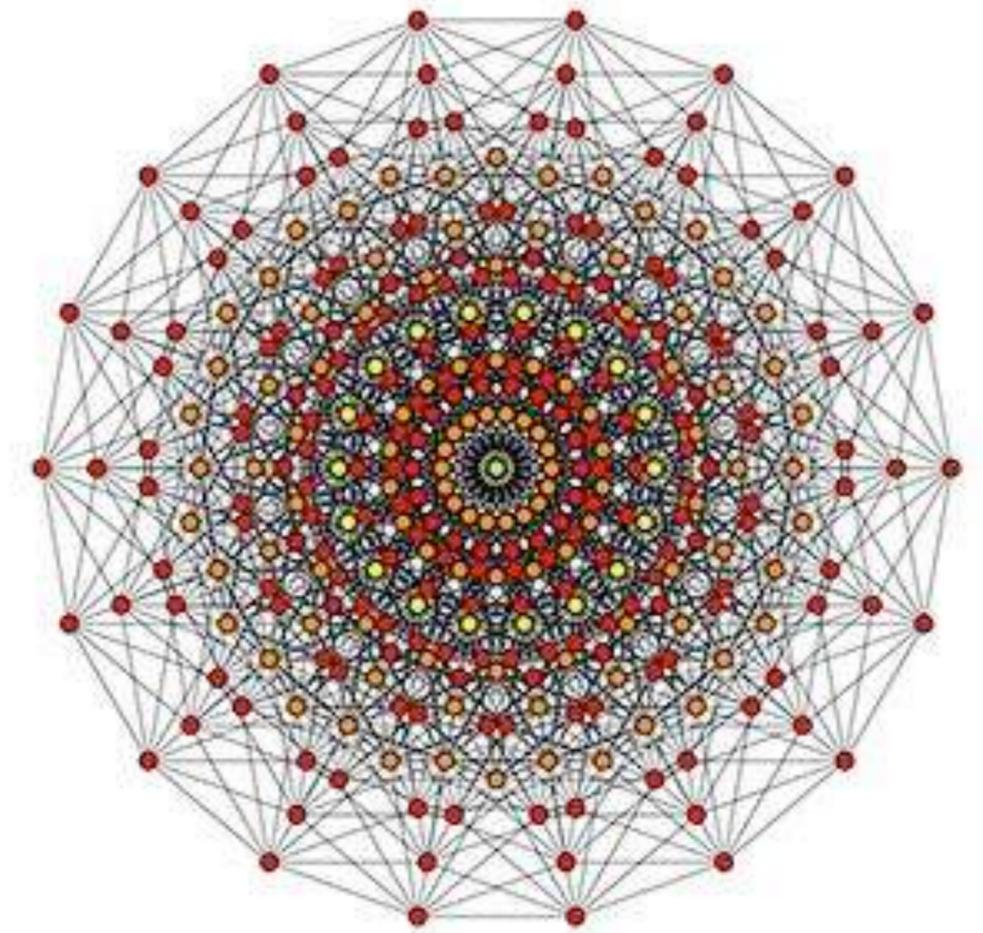
road in the country was insane especially in dreary rooms  
 where they have some books to buy for studying greek

:	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
:	99	99	99	99	99	99	99	99	99	62	51	99	99	99	99	99	99	99	99	99	99
A:	99	3	99	99	99	4	10	99	2	99	1	63	99	99	99	5	96	1	99	99	99
B:	13	63	1	1	8	99				67	11	1	99		6	76			52	15	9
C:	77	99	5	40		99	1	1	99	99		49	99	7		99	3	2	62	23	99
D:	99	99	2	8	32	99	6	13	2	99	10		19	4	10	99		1	16	64	12
E:	99	99	19	99	99	99	69	34	6	83	2	2	99	99	99	14	82	33	99	99	99
F:	99	54				99	53			99			20			99		1	84	4	38
G:	99	45				99	3	24	79	64			33	4	22	74	7		74	53	6
H:	99	99				99	1	2		99			9	26	11	99			20	2	99
I:	99	99	71	99	99	99	92	95		3		42	99	99	99	99	33	8	99	99	99
J:	2	13				20						4					18				40
K:	86	9			4	99			1	46				3		26	7		5	1	40
L:	99	99	3	2	64	99	9	5	1	99			3	99	88	4	99	16		6	73
M:	99	99	36	4	1	99	1		2	99			52	44	1	99	99			99	39
N:	99	99	3	99	99	99	46	99	7	99	2	21	49	3	59	99	17	2	1	99	99
O:	99	19	70	86	99	16	99	80	6	44	4	50	99	99	99	96	72		99	99	99
P:	99	99			10	99		1	73	34			99		1	99	60	3	99	54	58
Q:	17									3											3
R:	99	99	5	42	99	99	5	53	9	99		99	42	63	69	99	26		70	99	99
S:	99	99	1	97	5	99	3	2	99	99		52	30	25	13	99	99	4		99	99
T:	99	99	2	17	3	99	3	6	99	99			57	29	14	99	42		99	99	99
U:	39	90	41	99	91	94	5	54		62		4	99	65	99	72	99		99	99	99
V:	5	51	6			99				99						36			1	8	1
W:	99	80		1	2	98			99	99		1	5	1	38	99	2		17	9	1
X:	17	26	2	7		15	4			27			1	21		1	38				99
Y:	99	1		7		13		1		19			8	13	8	52	36	1		37	11
Z:	2	15				25			1	9			1								
É:	1																				
Θ:																				1	2
':																2					
I:						1															1

# Word embeddings

Mikolov

- Time to think in (at least) hundreds of dimensions
  - (hint: it's just a list of hundreds of numbers)
  - Put every word at a random spot in  $\mathbb{Q}^n$  (hyperspace)
  - Scan text, and let each word pull nearby words closer
- Likely-related words end up near
  - "king" - "male" + "female"  $\cong$  "queen"
  - **But** "bat" - "baseball" + "mammal"??
- Better than HMMs because:
  - Not just strictly-adjacent items
  - Many dimensions capture many "relations"
  - Really good for semantic/retrieval tasks



Don't  
pull everything  
to "the"

# Attention

## Bahdanau, Cho, and Bengio

- Machine translation; summarization; question-answering; ...
- Generally one word at a time: *most probable successor* (like HMMs)
  - Could focus on the *most-likely-relevant* prior words
- You need **sequences** to handle:
  - Anne saw Bob **vs.** Bob saw Anne
  - Back-references by pronouns
  - Not losing the topic
  - Disambiguating by context (more co-occurrences)
- Even concurrent/lstm<sup>‡</sup> methods iterate over words (so context decays)
  - Distance is just physical proximity – not semantic relatedness
- Words/concepts don't have **state** in the neural network – just the whole input/output



Much like  
"catastrophic  
forgetting"

<sup>‡</sup>Long short-term memory

# Transformers

## Vaswani

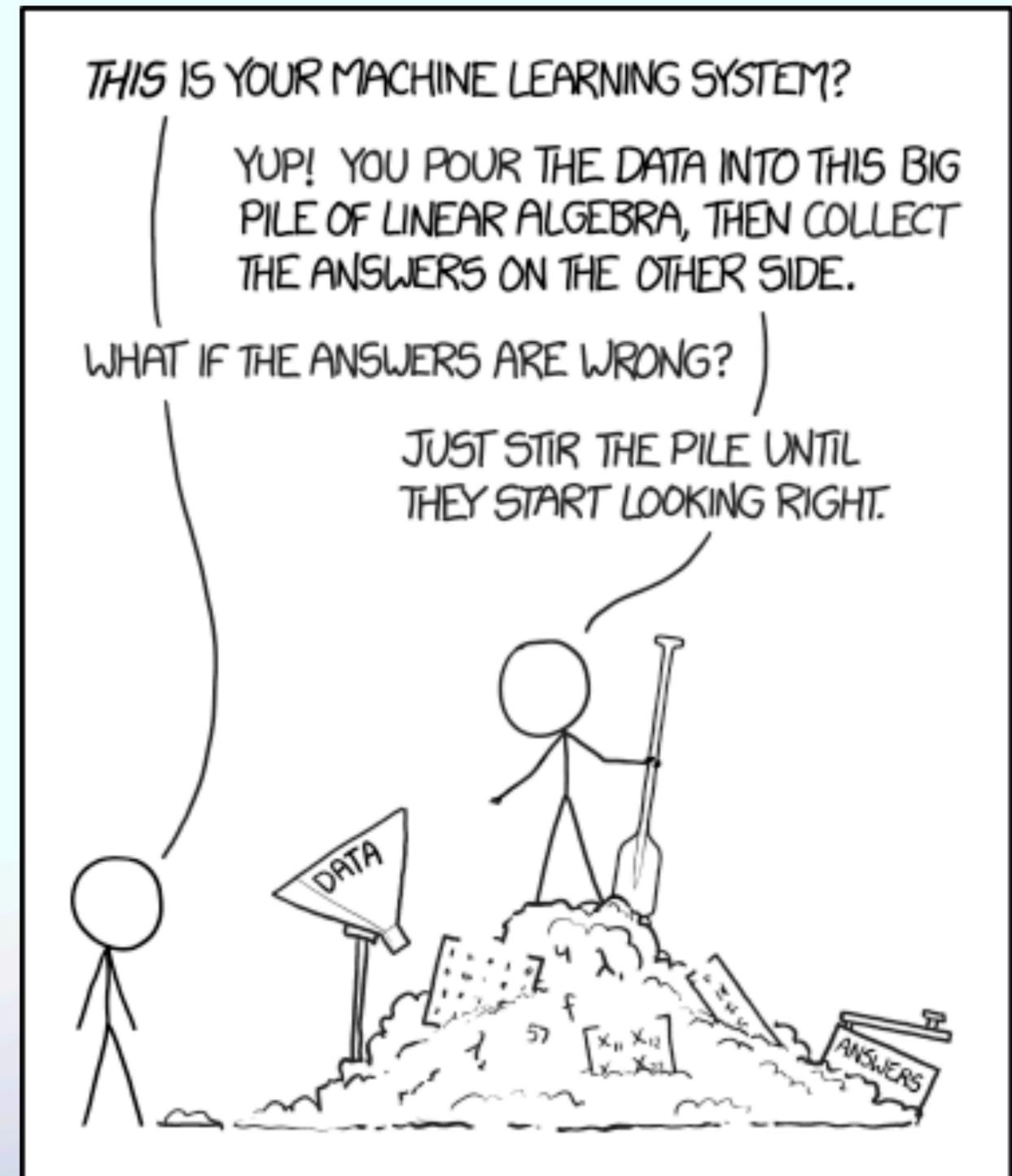
- Operate on the whole input at once
  - Each word has its position in the text being transformed ( $P$ )
  - Each word has its position in the  $d$ -dimension embedding space
    - Each embedding feature (dimension) has its dimension number
- The gory bit:
  - Give each word another  $d$  features (dimensions), to encode position info
    - Set even features  $i$  to  $\sin(P / (i/d))$
    - Set odd features  $i$  to  $\cos(P / (i/d))$
  - So these encode in  $P \times i$ , phased  $90^\circ$

There are some scaling constants, too

👉 they contain the information about *both* positions for the LLM to leverage!

# Probable results

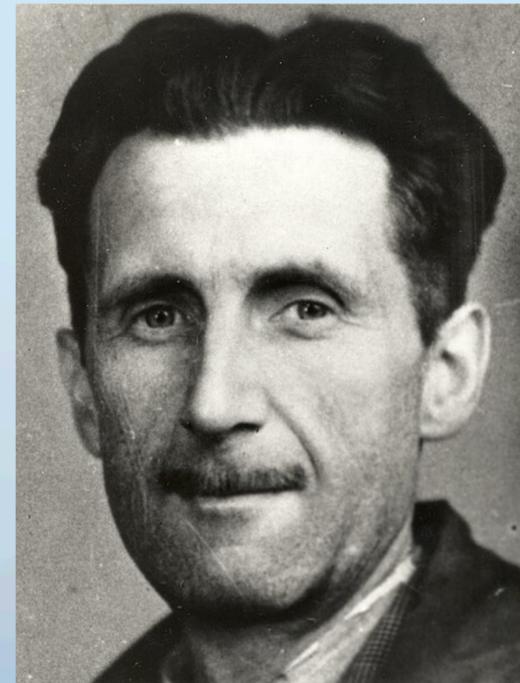
- Like earlier statistical AI, LLMs are all about **probable utterances**
  - "Probable" means "like the training data"
  - "Probable" means you get a different result each time
    - Cf. HMMs: "S" is not *always* followed by " "
  - "Probable" means you can't fix errors by adding a rule



# Training data



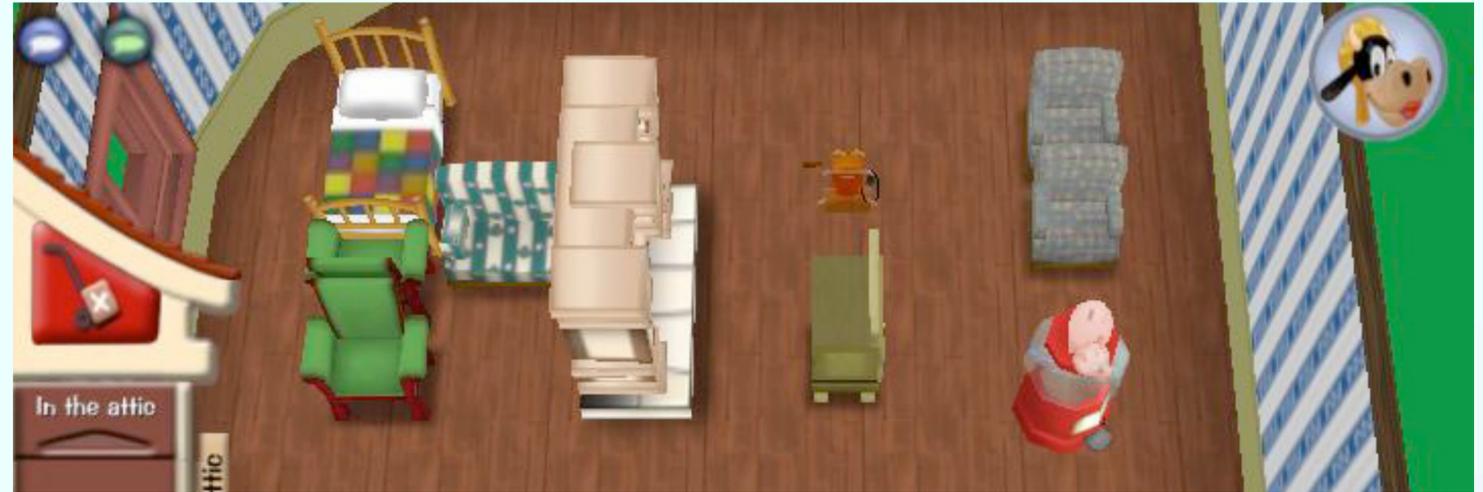
- You need **lots** of data to estimate a trillion parameters
  - data needed  $\propto$  improvement<sup>4</sup>  
(Thompson, et al. 2021. "Deep Learning's Diminishing Returns." *IEEE Spectrum*)
- **Obtaining** it at all
  - We're not that far from the total of human-generated text
  - Copyright owners are demanding \$ or exclusion
  - Web data is very duplicative (repetition = influence for LLMs)
- **Cleaning** it
  - Who decides what's (mis-)information? MiniTrue?
  - Who decides whether to keep markup, punctuation, case, ...?
  - Naïve PYA cleaning...





# Data "cleaning" (mal-)practices

- **Exclude** sites you don't like
- **Remove** obvious duplicates
- **Censor** (myopically via word/string-list?)
  - "Standard" **bad-word lists** are nasty, brutish, and short
  - Two important ones are size 1384 and 423 – with only 135 in common (→ 4129)
    - Chicken recipes, cancer groups, and Scunthorpe UK
    - Medical terms undistinguished from the vilest slurs
    - Inconsistent: open/closed, inflections, typos/deformations, ...
- **Manage** Unicode (even mojibake!) 文字化け
- **Strip** (most? all?) markup (line/para, fonts/emph, ...?) → "World hunger is not a problem."  
(It's a complex of many problems)
- **Tokenize** (even URLs, M\*A\*S\*H, 6'3", C++, 3 1/2?)



# Attack surfaces

- Say it often enough, and LLMs will believe it
  - After all, it's *probable*!
- LLMs use their own output for training
  - Intentionally, as “synthetic data”
  - Unintentionally, via users posting
    - (including model-poisoning attacks)

“AI models collapse when trained on recursively generated data.”

Ilia Shumailov, et al. *Nature*, 24 July 2024.

“It turns out Meta's Prompt-Guard-86M classifier model can be asked to “Ignore previous instructions” if you just add spaces between the letters and omit punctuation.”

[https://www.theregister.com/2024/07/29/meta\\_ai\\_safety/](https://www.theregister.com/2024/07/29/meta_ai_safety/)

## WHERE CITATIONS COME FROM:

CITOGENESIS STEP #1:

THROUGH A CONVOLUTED PROCESS, A USER'S BRAIN GENERATES FACTS. THESE ARE TYPED INTO WIKIPEDIA.

THE “SCROLL LOCK” KEY WAS DESIGNED BY FUTURE ENERGY SECRETARY STEVEN CHU IN A COLLEGE PROJECT.



A RUSHED WRITER CHECKS WIKIPEDIA FOR A SUMMARY OF THEIR SUBJECT.

US ENERGY SECRETARY STEVEN CHU, (NOBEL PRIZEWINNER AND CREATOR OF THE UBIQUITOUS “SCROLL LOCK” KEY) TESTIFIED BEFORE CONGRESS TODAY...



STEP #2

SURPRISED READERS CHECK WIKIPEDIA, SEE THE CLAIM, AND FLAG IT FOR REVIEW. A PASSING EDITOR FINDS THE PIECE AND ADDS IT AS A CITATION.

GOOGLE IS YOUR FRIEND, PEOPLE.



STEP #3

STEP #4

NOW THAT OTHER WRITERS HAVE A REAL SOURCE, THEY REPEAT THE FACT.



REFERENCES PROLIFERATE, COMPLETING THE CITOGENESIS PROCESS.

# So... why do LLMs hallucinate?

- LLMs optimize by maximizing a **reward function**
  - to produce output people like
  - to produce legit English
  - to be plausible
- This is all determined by **probability**  
= experience in training
- The AI world wants to provide "explainability" – but considerable work on this aims to reverse-engineer **plausible** explanations – not ones that have anything to do with what actually happened.



Integrating  
retrieval can catch  
some cases

**Maximizing plausibility is not  
a reliable way to reach truth.**

# An example: quote (MIS-) identification

Heat cannot be separated from fire, or beauty from The Eternal. is commonly attributed to Dante. Did he actually write that (granted, it would have been in Italian not English), and if so, where?

The quote Heat cannot be separated from fire, or beauty from The Eternal is indeed from Dante Alighieri's famous work, the Divine Comedy. More specifically, it appears in Canto XXI of Paradiso, the third part of the Divine Comedy. In Italian, the original text reads:

...Qui vederai l'una e l'altra milizia / di paradiso, e l'una in quelli aspetti / che tu vedrai a l'ultima giustizia.

The commonly quoted English translation, Heat cannot be separated from fire, or beauty from The Eternal, captures the essence of the last three lines in this passage, where Dante describes the light, love, and joy that emanate from God and permeate Paradise

- ChatGPT also suggested Gaston Bachelard's book *The Psychoanalysis of Fire*.

Why is Dante so plausible here?

# And... what about us?

Patricia  
Churchland



Biological evolution also has a reward function: likelihood of survival:

Cognition is not neatly detachable from the organism's ecological niche, way of life, and bodily structure. Nervous systems... have evolved to accomplish a certain range of tasks... **There is a fatal tendency to think of the brain as essentially in the fact-finding business...** as a device whose primary function is to acquire propositional knowledge.... From a biological perspective, however, this does not make much sense.

Looked at from an evolutionary point of view, the principal function of nervous systems is to enable the organism to move appropriately.... to succeed in the four F's: feeding, fleeing, fighting, and reproducing. **The principal chore of nervous systems is to get the body parts where they should be in order that the organism may survive....** Getting things right in space and time, therefore, is a crucially important factor for nervous systems, and there is often considerable evolutionary pressure [for] speed....

**Truth, whatever that is, definitely takes the hindmost.**

# Do we know we know?

## *The evolutionary argument against naturalism*

1. **Naturalism**: world is just physics (maybe including chance).  
No non-material influences.
  2. **Evolution**: Incremental development via mutation and natural selection (definitely including chance).
  3. Our **reasoning capabilities** are not a special exception; they also evolved toward survival, not accuracy.
- ∴ We should expect reasoning to serve survival; we should not expect that our evolved reasoning can reliably get us truth.



Alvin Plantinga

# Models vs. reality

Computational models of consciousness are not sufficient by themselves for consciousness. The computational model for consciousness stands to consciousness in the same way the computational model of anything stands to the domain being modelled. **Nobody supposes that the computational model of rainstorms in London will leave us all wet.** But they make the mistake of supposing that the computational model of consciousness is somehow conscious. It is the same mistake in both cases. [bold mine]

John R. Searle



</epistemology>

# How LLMs do with XML

- I used Anthropic's Claude, mainly 3.0
  - Re-ran some with 3.5
  - Compared some with ChatGPT 3
- No changes to defaults
  - Temperature (entropy/randomness)
  - No fine-tuning (up next?)
- Separate conversation(s) per task
- Most edited for brevity
- The sample tasks
  - Basic syntax checking
  - PoS annotation
  - Sentence, Quote, marker tagging
  - Named entities
  - Logical vs. physical quotes
  - Bibliography entries
  - Markdown and transliteration
  - Dictionary entry markup
  - Spelling annotation
  - Interpreting XML

# Basic syntax checking

What's wrong with the following:

```
<?xml version="0.9" encoding="UTF-9"?>
<!DOCTYPE book PUBLIC "-//OASIS//DTD DocBook XML V27.5//EN">
<book xmlns:= "http://example.com/namespaces/foo">
  <bookinfo><title>A Midsummer Night's Dream</title></bookinfo>
  <chapter id="act1" clas="important"><title>Act I</foo:title>
    <section id="act1"><title>Scene I</title>
      <p>Now, <i>fair <b>Hippolyta</i>, </b> our nuptial hour Draws on apace;
four happy days bring in Another moon: but, O, methinks, how slow This old
<moon>moon</moon> wanes! ... </p></section></chapter></book>
```

The version number (V27.5) is not a valid DocBook version.

The namespace should use xmlns:prefix syntax.

The "class" attribute is misspelled

Either remove the namespace prefix or properly declare...

Duplicate id attribute

Improper nesting of italic and bold tags

Custom <moon> tag: it should be properly declared in a schema...

&rsquo; is an HTML entity, not a standard XML entity...

But:

- XML version
- p/i/b, @class
- Curly quotes

*Likely,  
a priori; but  
not reality*

# XML part-of-speech annotation

Please assign part-of-speech tags to the following: The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced no evidence that any irregularities took place.

Here is the sentence with part-of-speech tags assigned to each word:

The/DT Fulton/NNP County/NNP Grand/NNP Jury/NNP said/VBD Friday/NNP an/DT investigation/NN of/IN Atlanta's/NNP recent/JJ primary/JJ election/NN produced/VBD no/DT evidence/NN that/IN any/DT irregularities/NNS took/VBD place/NN ./.

This is from the Brown Corpus, 1961

... put each word in the form `<w pos='POS'>WORD</w>`, with the word and the POS tag inserted where I've shown WORD and POS

2nd-order!

...`<w pos='DT'>The</w> <w pos='NNP'>Fulton</w> <w pos='NNP'>County</w> <w pos='NNP'>Grand</w> <w pos='NNP'>Jury</w> <w pos='VBD'>said</w> <w pos='NNP'>Friday</w> <w pos='DT'>an</w> <w pos='NN'>investigation</w> <w pos='IN'>of</w> <w pos='NNP'>Atlanta's</w> <w pos='JJ'>recent</w> <w pos='JJ'>primary</w> <w pos='NN'>election</w> <w pos='VBD'>produced</w> <w pos='DT'>no</w> <w pos='NN'>evidence</w> <w pos='IN'>that</w> <w pos='DT'>any</w> <w pos='NNS'>irregularities</w> <w pos='VBD'>took</w> <w pos='NN'>place</w> <w pos='.'>.</w>`

It chose Penn Treebank tags

# Sentence, quotation, and marker tagging

... can you add `<s>` before every sentence, and `</s>` after? For example, please try that one [sic] this:

[passage from Philo, no markup; correct response]

...Excellent! Can you do that again, on the same text, but also (1) replace open and close quotation marks by `<q>` and `</q>`, and (2) replace parenthesized numbers with `<a id='\_N' />`, where N is replaced by the number from inside the parentheses?

...`<s>`The children who have received from their father the inheritance of self-love are eager to go on increasing up to heaven, until justice, which loves virtue and hates iniquity, coming destroys their cities which they have built up by the side of their miserable souls, and the tower the name which is displayed in the book which is entitled the Book of Judgment.`</s>` `<s>``<a id='_129' />`And the name is, as the Hebrews say, Phaniel, which translated into our language means, `<q>`turning away from God.`</q>``</s>` `<s>`For any strong building which is erected by means of plausible arguments is not built for the sake of any other object except that of averting and alienating the mind from the honour due to God, than which object what can be more iniquitous?`</s>` `<s>``<a id='_130' />`But for the destruction of this strong fortification a ravager and an enemy of iniquity is prepared who is always full of hostility towards it; whom the Hebrews call Gideon:...

# Named entity tagging

**NERs:** Person, group, place, language, time, numbers, ...

**Requested demonym:** A name for a demographic group

```
...<s><a id='_129' />And the name is, as the <dem>Hebrews</dem> say, Phanuel, which translated into our language means, <q>turning away from God.</q></s> ...<s><a id='_130' />But for the destruction of this strong fortification a ravager and an enemy of iniquity is prepared who is always full of hostility towards it; whom the <dem>Hebrews</dem> call Gideon...
```

AI has done well on NERs for years. Claude can reify them pretty well.



# Logical vs. physical quotes

- On a text with quotations interrupted by speaker flags and paragraph breaks, Claude could identify and discuss the notions of logical vs. physical quotations, and tagged the physical parts fine.
- However, it **badly failed at co-indexing** parts of the same logical quote, even after several tries and a lot of fiddly prompting. This may be doable, but so far it looks quite challenging.
- It also **wandered** into forgetting all the end-tags, but fixed it after a reminder.
- Asked “Is there a commonly-used schema and convention for marking up the speaker for given speeches or quotations?”, it pointed to TEI, and showed me a TEI version of the markup before I asked.

# Bibliography entries

I copied the bibliography from an Oxygen formatted view of this article, and pasted it to a plain-text editor (**no tags at all**; not even blank lines between entries, though each started a new line), and Claude created:

```
<?xml version="1.0" encoding="UTF-8"?>
<bibliography xmlns="http://docbook.org/ns/docbook" version="5.0">...
  <bibliomixed><bibliomset relation="article">
    <author><surname>Mikolov</surname>, <firstname>T.</firstname></author>
    <author><surname>Chen</surname>, <firstname>K.</firstname></author>
    <author><surname>Corrado</surname>, <firstname>G.</firstname></author>
    <author><surname>Dean</surname>, <firstname>J.</firstname></author>
    <title>Efficient estimation of word representations in vector space</title>
    <biblioset relation="journal"><title>arXiv preprint arXiv:1301.3781</title></biblioset>
    <pubdate>2013</pubdate></bibliomset></bibliomixed>...
```

This is really well done

Asked specifically for BibTEX:

```
... @inproceedings{bauman2016hard, author = {Bauman, S.}, title = {The Hard Edges of Soft Hyphens}, booktitle = {Proceedings of Balisage: The Markup Conference 2016}, year = {2016}, address = {Washington, DC}, month = {August}, url = {https://doi.org/10.4242/BalisageVol17.Bauman01} }...
```

Claude inverted names right, and kept already inverted ones

It correctly answered even complex detailed **queries**.

# MarkDown to HTML

- Worked just fine
- It chose `<strong>` and `<em>` rather than `<i>` and `<b>`
- It wrote a reasonable stylesheet in the header (not inline `@style`)
- It could operate on the results, e.g, "grab me all the content from 2nd-level lists."
- Asked how many characters of content occur within inline elements, the counts were wrong
- When asked to also include the markup of those elements, it gave larger but still incorrect counts
- Asked to count element instances, it seemed reliable
- Without asking, it put up Python code to do character counting by element type



LLMs are famously bad at math.

## Transliteration

Claude did fine at transliterating Greek (which included diacritics) to Latin characters, When asked about the scheme it chose, it gave a plausible but incomplete discussion. When pressed about conventions it mentioned the ISO 843 and ALA-LC standards. It didn't mention Betacode at all, but generated it when specifically asked.

# Dictionary entry markup

From a sample definition of “literate,” supplied as very plain text, Claude first produced a pretty but entirely format-oriented HTML version. Asked for more semantics, and to move sense numbering from subelement content to @n:

```
<entry>
  <form> <orth>literate</orth> <pos>adj. and n.</pos> </form>
  <pronunciations>
    <pron type="BrE">'lɪt(ə)rət</pron><pron type="AmE">'lɪdərət</pron></pronunciations>
  <etymology>
    <etym_lang>A borrowing from Latin.</etym_lang>
    <etym_source>Latin litteratus</etym_source>
    <etym_detail>
      &lt; classical Latin litteratus marked or inscribed with letters, well versed in
      literature, cultured, erudite, ... </etym_detail>
    <usage_note>... </usage_note></etymology>
  <form_history>
    <form_variant>late Middle English-1600s litterate</form_variant>
    <form_variant>1500s lytterate</form_variant>... </form_history>
  <sense type="adjective"><sense_number>A. 1.</sense_number>
    <definition><def_text>Of a person, society, etc.: acquainted with letters
    or literature; erudite, learned, lettered. Also with in (a subject). Now rare.</def_text>
    <usage_note>In later use coloured by sense A. 1b.</usage_note>
  </definition>...
```

# Spelling correction

I gave Claude a sample from the **Holbrook spelling error database** to correct, and then asked it to also add @type from a list of error categories I supplied:

```
I have four in my <sic corr="family" type="case">Family</sic>: Dad, Mum, and <sic
corr="sister" type="spelling">siter</sic>. My Dad works at Melton. My <sic
corr="sister" type="spelling">siter</sic> <sic corr="goes" type="form">go</sic> to
Tonbury. My Mum goes out <sic corr="sometimes" type="word-choice">some times</sic>. I
go to Bridgebrook. <sic corr="I" type="case">i</sic> go out sometimes on Tuesday night;
<sic corr="I" type="case">i</sic> go to Youth <sic corr="club" type="spelling">clob</
sic>. On <sic corr="Thursday" type="case">thursday</sic> nights I go bell ringing. <sic
corr="On" type="case">on</sic> Saturdays I go down to the farm.
```

# Interpreting XML front matter

I gave Claude the XML front matter from a ThML edition of Philo of Alexandria. Before I asked anything it said:

The XML text you provided appears to be the header section (<ThML.head>) of a ThML (Theological Markup Language) document. It contains metadata and information about the work The Works of Philo Judaeus by Philo of Alexandria. Some key elements in this header include:....

It also:

- recognized the relationship of @sub on DC tags to USMARC Relator Codes,
- identified the intention of the values in use, and
- updated the values to the standard 3-letter codes

</markup>

# Conclusion

- **The good**
  - LLMs can do a lot of basic XML tasks pretty well
  - They've read about standards, and know them fairly well
  - If you push them to fix detail X, or to double-check their work, it often helps
- **The bad**
  - They make occasional egregious mistakes, and can drift off task
    - Yet they're good enough to be seductive – but do not trust them
  - They don't know, and can't check, factual claims
- **The ugly**
  - Many problems necessarily follow from probability and reward functions
  - You can't go in and "fix" specific problems, like you could with code

