Locality, Relation, and Meaning Construction in Language, as Implemented in Humans and Large Language Models (LLMs)

March 19, 2025

Julia Witte Zimmerman PhD Candidate, CSDS Advisors: Christopher M. Danforth & Peter Sheridan Dodds Vermont Complex Systems Institute, UVM BSc, Linguistics, MIT



Thank you...

To the people, animals (especially Mumi, Numi, and Squishi), plants, fundamental forces, places, atoms, and organizations who have supported this work directly and indirectly, including UVM, MIT, the Houston, Texas Independent School District (HISD), my family and friends, and Alejandro J. Ruiz.

I have had the resources and opportunity to work on this project due in no small part to luck in its various incarnations.

Some of those resources come at the **cost** of other people and the environment. Therefore some (meager) remuneration is within the scope of this project: I donated fifty dollars to the Environmental Defense Fund and fifty dollars to the American Civil Liberties Union.

Thank you, in no particular order! If I omitted anyone, my sincere apologies. To avoid repetition, even if someone was in multiple categories. I only listed them once.

Thank you to my advisors, Christopher M. Danforth and Peter Sheridan Dodds, and to my additional defense committee members, Randall Harp, Donna Rizzo, and Juniper L. Lovato.

Thank you to my co-authors, Denis Hudon, Kathryn Cramer, Jonathan St. Onge, Mikaela Irene Fudolig, Milo Z. Trujillo, Thayer Alshaabi, Shawn Beaulieu, Joshua R. Minot, Michael V. Arnold, Andrew J. Reagan, Isabelle Smith, Jennifer Karson, Yoshi Meke Bird, Sarah E. Grobe, Sean P. Rogers, Calla Beauregard, Ashley Fehr, and Bradford Dem**a**rest.

Thank you to those acknowledged in the individual works this thesis primarily relies on, Desi Alexander, Danbee Kim, Madeline Pennebaker, Mark Wittels, Grace Taylor, Julia Chimienti, Evan Lynch, Cecile Smith, Peter Heft, Ashley Dennis-Henderson, William Thompson, Gabriel Meyer-Lee, Robert Wolfe, Alice Patania, Josh Bongard, Kenneth C. Fan, Haim Dubossarsky, Alexa Woodward, Charlie Brooks, Ben Cooley, participants and leaders of CEL (the Computational Ethics Lab), CSL (the Computational Story Lab), and SCRaPs (Student Complexity Research and Pizza Seminars), Carter Ward, Aviral Chawla, Parisa Suchdev, Mohsen Ghasemizade, Jacques Bailly, and Guillermo Rodriguez. Thanks to the speakers, participants, and organizers of the excellent June 2024 UQAM summer school, "Understanding LLM Understanding", moderated by Stevan Harnad. Thanks to Tyler Ehrlich and Philip Nguyen for their work on previous versions of the in-progress Character Space project. Thanks to Ethan Ratliff-Crain, Nate Gaylinn, Fitzwilliam Keenan-Koch, Krishna Kannan Srinivasan, Collin Coil, Dakota Marosi, and Katie Grenon for thought-provoking comments. Thanks to Anthony Barrows and the GWC (Graduate Writing Center) for helping improve this thesis.

Thank you to all the authors, thinkers, and researchers whose work this thesis is built upon — whether or not the views expressed herein are in agreement with those works or in opposition to them. Particularly the people who took big swings in good faith, regardless of the result.

Thanks to my teachers (especially those in the Complex Systems Institute), including Nick Cheney, Laurent Hébert-Dufresne, and Ben Sienicki. Thanks to MassMutual and the National Science Foundation for providing financial support for this work. This material is based upon work supported by the National Science Foundation under Award No. 2242829.

Finally, I want to thank my entire family/

Finalmente, quiero dar las gracias a toda mi familia/

Finalment, vull donar les gràcies a tota la meva família,

including but assuredly not limited to ...

Ale, Catherine, Gary, Geoff, Marina, Javier, Cata, Isa, Nani, Ducky, Richie, Christy, Matt, Jo, Lu, Harold, Nancy, Robert, Kathy, Doreen, Ken, Adam, Kristin, Stefanie, and Dustin.

Thank you, in memoriam, to Joseph Nogee, Sarah Colletti, Robert and Norma Witte, Esther Zimmerman, and Marina Iglesias.

Roadmap

- We explore language and cognition in both people and in computational models
- Through the lens of meaning construction at both the individual and collective level
 - Pattern to thought transition
- We step through levels of linguistic structure and large language model (LLM) architecture from morphemes to stories
- Begin to build an understanding of the role these concepts play in human and machine cognition, and, in particular, in meaning construction and symbol grounding

Introduction

1 Introduction

1.1	Road-map of Thesis
1.2	Bowtie architecture
1.3	Representations, metaphors, and leaps of faith
1.4	Language and its acquisition
1.5	Generative AI
1.6	Large Language Models (LLMs)
	1.6.1 Why do we care? Why is this a big deal?
	1.6.2 Impossible Languages
1.7	Bias
1.8	Tokenization
1.9	Singular Value Decomposition (SVD)
Bibl	iography



Large Language Models (LLMs)

- LLMs are deep learning models that intake huge amounts of text as training data and create new utterances by choosing the next token to output
 - The main goal is *plausible text generation* (whether MLM or autoregressive)
- These models use an attention-based transformer architecture, an evolution of *neural networks*
- Why is this a big deal?
 - There has been a huge leap in performance
 - Current LLMs sound much more like people than any previous model did



What could be salient given LLM architecture?

Wh	at coul	d be salient given LLM architecture?				54
2.1	What	is it like to be ChatGPT?				56
2.2	What	does it mean to be trained on text?				57
	2.2.1	Word embeddings				59
	2.2.2	Text? Language? Words?				60
	2.2.3	Diegetic and supradiegetic linguistic information				61
	2.2.4	Salient splits				66
	2.2.5	Vocabulary				67
2.3	Where	ChatGPT runs into trouble				68
	2.3.1	Sumerian Cuneiform				68
	2.3.2	Supradiegetic linguistic information is a blind spot				72
	2.3.3	Palindromes and Symmetry				72
2.4	TL:DR	R; Why is this hard?				77
2.5	Capab	ilities				79
	2.5.1	A Caveat		2		79
	2.5.2	ChatGPT speaks English		2		79
	 Wh 2.1 2.2 2.3 2.4 2.5 	What coul 2.1 What 2.2 What 2.2.1 2.2.1 2.2.2 2.2.3 2.2.4 2.2.5 2.3 Where 2.3.1 2.3.2 2.3.3 2.4 2.5 Capab 2.5.1 2.5.2	 What could be salient given LLM architecture? 2.1 What is it like to be ChatGPT? 2.2 What does it mean to be trained on text? 2.2.1 Word embeddings 2.2.2 Text? Language? Words? 2.2.3 Diegetic and supradiegetic linguistic information 2.2.4 Salient splits 2.2.5 Vocabulary 2.3 Where ChatGPT runs into trouble 2.3.1 Sumerian Cuneiform 2.3.2 Supradiegetic linguistic information is a blind spot 2.3.3 Palindromes and Symmetry 2.4 TL:DR; Why is this hard? 2.5 Capabilities 2.5.1 A Caveat 2.5.2 ChatGPT speaks English 	What could be salient given LLM architecture? 2.1 What is it like to be ChatGPT? 2.2 What does it mean to be trained on text? 2.2.1 Word embeddings 2.2.2 Text? Language? Words? 2.2.3 Diegetic and supradiegetic linguistic information 2.2.4 Salient splits 2.2.5 Vocabulary 2.3 Where ChatGPT runs into trouble 2.3.1 Sumerian Cuneiform 2.3.2 Supradiegetic linguistic information is a blind spot 2.3.3 Palindromes and Symmetry 2.4 TL:DR; Why is this hard? 2.5 Capabilities 2.5.1 A Caveat 2.5.2 ChatGPT speaks English	What could be salient given LLM architecture? 2.1 What is it like to be ChatGPT? 2.2 What does it mean to be trained on text? 2.2.1 Word embeddings 2.2.2 Text? Language? Words? 2.2.3 Diegetic and supradiegetic linguistic information 2.2.4 Salient splits 2.2.5 Vocabulary 2.3 Where ChatGPT runs into trouble 2.3.1 Sumerian Cuneiform 2.3.2 Supradiegetic linguistic information is a blind spot 2.3.3 Palindromes and Symmetry 2.4 TL:DR; Why is this hard? 2.5 Capabilities 2.5.1 A Caveat 2.5.2 ChatGPT speaks English	What could be salient given LLM architecture? 2.1 What is it like to be ChatGPT? 2.2 What does it mean to be trained on text? 2.2.1 Word embeddings 2.2.2 Text? Language? Words? 2.2.3 Diegetic and supradiegetic linguistic information 2.2.4 Salient splits 2.2.5 Vocabulary 2.3 Where ChatGPT runs into trouble 2.3.1 Sumerian Cuneiform 2.3.2 Supradiegetic linguistic information is a blind spot 2.3.3 Palindromes and Symmetry 2.4 TL:DR; Why is this hard? 2.5 Capabilities 2.5.1 A Caveat 2.5.2 ChatGPT speaks English

	2.5.3	What else can ChatGPT do? 8
2.6	Is the	structure necessary or just one way that works? 8
	2.6.1	What is the technological package of linguistic fluency? 8
	2.6.2	Intelligence and language
2.7	Multin	nodal processes
	2.7.1	Vision and language
	2.7.2	Descriptive and depictive representations
2.8	Return	ing to palindromes
2.9	Comm	on sense and Math
	2.9.1	Why does mathematics not come with linguistic fluency? 9
	2.9.2	Holey sequences
2.10	Fuzzin	g up frequency and truth
2.11	Symbo	ls
	2.11.1	Packages and contents
	2.11.2	Arbitrary leaps
	2.11.3	Diegetic boundaries
	2.11.4	Flatland
	2.11.5	Ergodicity and span 10
2.12	Return	ing to Sumerian cuneiform 10
2.13	Approx	ximations: only the thing is the thing
2.14	Gricea	n cooperation
2.15	Where	can we go?
	2.15.1	Postlapsarian
	2.15.2	What things can ChatGPT bring to the table?
Bibli	ograph	y

We want to know... what can LLMs understand?

- What factors enable their human-like language performance?
- What is their cognition like?
- What can they absorb via what they can perceive of the linguistic signal?
- What is it like to be an LLM?
- What is their gnogeography like?
 - Gnogeography: the landscape of knowledge, a being's mind-and-body/ architecture in conjunction with what information they have taken in (perception, cognition).
 - Everything in their gnogeography is more (~proximal) or less (~distal) salient.
 - Imperceptibility means an ultimate lack of salience.
 - Reachability (traversal through gnogeography) is salience in the limit.
 - Locality ~ salience



What is it like to be ChatGPT?

- ChatGPT does not have a human-like body; its experiences come to it mainly through its exposure to training data and input
- We say LLMs are trained on "language" or "text", but they are really trained only on a small portion of text: the 1D relations between symbols, the distributions
 - What can an LLM come to know, from the distributional properties of language as reflected in what it can perceive of text?
 - The symbol grounding problem: how do symbols gain their meaning?
 - How can LLMs, with such curtailed and comparatively indirect access to the world, come to know about it?
 - What is reference, for an LLM?
- For people, language is an embodied task and our senses ground our cognition



Diegetic Approximations

- Describing something fundamentally outside a being's gnogeography using pieces from within their gnogeography
 - Making a representation of something local when the thing itself is unreachable
- For example, ChatGPT cannot perceive what the letter *c* looks like, but we can provide it a diegetic approximation in text, such as: "the letter *c* is round like an oval or a circle, but missing a segment on the right side"
- It's immediately apparent that perceiving the appearance of the letter *c* and our diegetic approximation of it are not identical
 - There must be ambiguity unless the map is the size of the territory
- But there's no obvious reason to think the approximation can't be arbitrarily good
- We all live in worlds of diegetic approximations... but ChatGPT's is much flatter than ours
- As a corollary: we should be wary of benchmarks!



Tokenization in LLMs

3 Tokenization in LLMs

3.1	Introd	uction
3.2	The D	Pistributional Hypothesis
3.3	Metho	odology
	3.3.1	Vocabularies
	3.3.2	Word lists
	3.3.3	Tools
	3.3.4	Parts of speech and human vocabularies
	3.3.5	Token embeddings (extispicy)
	3.3.6	Text data
3.4	Core I	Results
	3.4.1	Exemplar tokenizations
	3.4.2	Exemplar vocabularies
	3.4.3	Extispicy: exemplar tokens
3.5	Discus	sion
	3.5.1	Towards meaning and grounding
	3.5.2	POSIWID teleology: Objective functions
	3.5.3	Bias, alignment, and ethics
3.6	Conch	usion
Bibl	iograph	ν



Now that we have a basic frame of reference to start from, we can incorporate more specificity about how LLMs work, which brings us to tokenization!



- Tokens are a significant part of LLM architecture, but don't receive much scrutiny in terms of their contribution to the model's cognition Bridge between koinos kosmos and idios kosmos

- Tokens can be created with various strategies

 The overall goal is to come up with a set of tokens that can be combined to create any text the model could need to generate as output
- In practice, this means the tokenizer should balance the combinatorial power of short strings with the efficiency of long strings

100 - rec rec rec rec rec rec rec rec	-Wha	i kes o	with the r large	ones	ost th , to t	nou wis	ste t	hat Istle	toke Utler	ens ance	may	be	smal	D	2000 3000 5000 6000 8000 8000 13000 13000
80- ic or strategy s	7 ston 7 vers 5 orn vers ures con pic or without tures cice pic cice pic di without di without	? stion ? vers s or ver bures con pic or without turn	n 7 i ation vers con s of	7											15000 21000 23000 23000 25000 50000 1000000
60 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 -	 Allowpht thought k a boo b boo b boo b c <lic< li=""> <lic< li=""></lic<></lic<>	Alice pro- alice with thought with thought with thought with the second	iout pictures pictures e Arithout chought discost of Support and chought discost and chought discost d	s ation convers bures without without a book a use of is the with a and	7 7 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	n 7 svens solon is orinvers rout pic e without thought k i k i ori book ori solo svens s	7 s acion convers or bres pc w w thought hought book a use of	? 5 ablen comves of curves stronget hought book a	7 stion convers pictures without Alice thought	7 ation convers pictures without Alice thought	7 abon corrers of pictures without Alice thought	7 abion convers or pictures without Alice thought	7 ation conversion pictures without Alice thought		
40 - 40 - 40 - 40 -	Nad pic L no L no Just Nad Was Nat Was Nat Was Nat Var Nat	Tures con pic or no bur he disc or no bur he disc or no bur her eep he disc or pe ed she ep had ep she pp to book	i con of si Luces i bito hoad i road ar road road si er mas sist k hee ed ep po po bad site i hoad i road sist i hoad i road sist i hoad i site i bito sist i bito i bito i bito sist i bito i bito sist i bito si bi bito si bito si bito si si bito si bi bi bi bito si bi b	It in atlons corrivers of tures of tures of thad thad thad thad er was sist ed into the ed into the ope	k in it attempt of the second	what and " " rs in uses atoms torses pic torses pic torses pic torses pic torses pic torses t	is the what and it ations in convers of resp is pic but ing is but ing er was sist her book	use of is the what and it ations in convers of pictures fo it had , but ing irread er was sist her book	a use of is the what and it ations in convers or pictures no pictures no that , but ing read sist the her	book a use of is the what and t ations in convers or pictures no R had , but ing read er was sist	book use of a is the what and it attens in convers of pictures no it had but ing read or mas sist	book use of a is the what and it abors in convers or pictures pictures pictures pictures and put ing read er was	book use of a is the what and it abios in convers or pictures no pictures ing read er was	? ation pictures or con- without Alice thought is the and what it ations in pictures or con- thad no-	noz
20 - 20 - 20 -	do in the second	or twike once or or indeto is the indeto the	ve trivice of conce to once to to the share dberg dberg dberg dberg dberg bit to the her bit trig bit ber bit of conce to the the the the the the the the the the	white or or or ing to th th belowing of of ank b on the ank b onthe b sister th th or of of of of of of of of of of	she had ge thice she for twice she or twice once or this of twice do the she the she she she the she	e of into the had of pe the she had e of the e o	ed into the ed into the phe had by the had by the of ed thing to no having of and bank of skown by by thing bank	ed into the ed into the see had be had be had be had or ence : ; do thing to no a having of , and bank on the sister by sting	book ed into the gp the had brie had of of thing to no having of , and bank on the sister by her sitting	her book eed into the CP ence into anti- nations of and bank on the sister by her string	her book ed into the pe pe she had write or once i do nothing to having of , and bank on the sister by her sitting	is the second se	d statu statu her ed into the ed she had she had once : do nothing to having of bank on the sister by her	A but ing to the second	"Who is packets or coverables?"
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	very ti get very to get inning to beginning was was ice ice Al Al	red red bi bi very very get get bo beginning beg was was Alice Alice	inning beginning e Alkce	red bi get beginning Mas Alice	of of red red to to to get to get to get to get Alice Alice	e Alice	of red ti very to get beginning was Alice	of red ti very to get beginning was Alice	of red ti to get beginning was Alice	of red ti very to get beginning was Alice	of red ti to pet beginning was Alice	sitting of tired very to get beginning was Alice	sitting of tired very to get beginning was Alice	sister by her sitting tired of very to get was beginning Alice	er was reading, but peeped into the book her sist twice she had , and of having nothing to do: once of sister on the bank sisting by her to get very tred of Alice was beginning

Tokens, Flattening



The Distributional Hypothesis (DH)

- Connecting tokens + language acquisition in LLMs ↔ people
- Z. Harris: meaning can be acquired by using the distributional properties of language as the "building blocks of semantics"
 J.R. Firth: "You shall know a word by the company it keeps"
- Learners leverage correspondence between "distributional similarity and meaning similarity" (Sahlgren, 2008)
 - In theory, not level-specific, could be at e.g. morpheme, word, phrase level; in practice (NLP, ML), usually at word level
- Unclear...
 - How much of human language acquisition is actually achieved through these means
 - Which distributions from within languages are relevant and to what degree
 - In "what sense it is meaning that is conveyed by distributional patterns" (Sahlgren, 2008)
 - Something like 2 regimes: distributional patterns for symbol grounding, learning; abstract representations, deeper meaning for mastery?

Relations and locality determine meaning

- Fundamental units in LLMs *must be* tokens and relations between tokens
- Most accessible (to the model) would seem to be distributions betwixt tokens
- LLMs are drenched in distributional information -- the patterns of tokens in the training data -- but probably need to build up some kind of foundation there first before they can make use of other information
 - 2 regimes?
- The principle of information locality (local relevance) is baked into human language already
- LLM preference for information locality is tied to model architecture through the objective function of next token generation as well as through mechanisms like relative position bias
 - Size of context window determines what relations will be perceived by the model
- Symbols and relations = bottom-most necessary components of language
- Much of their MVP meaning can be encoded morphosyntactically
- MVP lexical component can be comparatively minimal, maybe even non-existent
 - EmoAtlas (Semeraro et al., 2025), Lexinvariant LM (Huang et al., 2023)
 - Reminiscent of theories in which the fundamental unit of meaning construction is relation or metaphor, e.g. theories of C.S. Peirce

Tokenization algorithm: Efficiency & language

- Main method: training a BPE tokenizer
- Human-meaningful linguistic units appear as tokens
 ⇒ relationship between informative efficiency* and
 the structures in human language
 - *BPE is a compression algorithm, so entropy-reducing
- This is in line with a growing body of work on how such forces shape language
 - Specifically information-theoretic concepts of efficiency
- Pressures of orthographic efficiency, akin to popular tokenization algorithms, are reflected in world writing systems through ligature, and in contracted/ Grade 2 braille
- More broadly, we can see forces of efficiency [w.r.t. various resources] at work in language at all levels in the general principle *lectio difficilior potior*
 - For example, a non-trivial driver of language change over time is that people migrate a language towards sounds that are physically easier to make



Tokenization (kind of) re-invents words

- Space incorporated = similar word-boundary marking in both people and LLMs → importance of words
- If a ~word is the unit of language best suited to our ability to learn distributionally from the linguistic signal (at least in English), then models should perform best when provided easy access to plentiful word tokens
- Linguistic structures evolved over 100,000 1,000,000 years (Fedorenko et al., 2024): that's a lot of compute!
- → What else could we allow LLMs to take better advantage of? What else might they get out of the linguistic signal? What other objectives might a tokenization algorithm incorporate?



Theoretical limits on expression based on dimensionality

\$

\$

In *Le Morte d'Arthur*, "besperple" doesn't appear to be a commonly cited term or word. It's possible that it could be a misspelling or misinterpretation of a word from the text, which contains archaic language and terms that might seem unfamiliar today. If you encountered this word in a particular passage, it might help to examine the context more closely, as the spelling or meaning could be clearer in relation to the sentence or story.

Could you provide more context or the passage where you found the term? That could help in determining its meaning.

i think it means bespatter

You're correct! "Besperple" is an archaic word from *Le Morte d'Arthur* that means "to bespatter" or "to soil with liquid or mud." It reflects the older use of language, where "bes-" was a prefix indicating thoroughness or coverage, and "perple" likely meant "spatter" or "sprinkle." This fits well within the Middle English context of *Le Morte d'Arthur*, where vivid and sometimes coarse descriptions were used.

You're correct! "Besperple" is an archaic word from Le Morte d'Arthur that means "to bespatter" or "to soil w ith liquid or mud." It reflects the older use of langu age, where "bes-" was a prefix indicating thoroughness or coverage, and "perple" likely meant "spatter" or "s prinkle." This fits well within the Middle English con text of Le Morte d'Arthur, where vivid and sometimes c oarse descriptions were used.

LLM Vocabulary Status Quo

- Main method: looking at vocabularies used by real LLMs
- There are many words/ morphemes that aren't tokens!
 - The proportion varies depending on category
 - Out of the 189,558 approximate lexemes...
 - About 2/3 of the base words are not tokens in any of the files!
 - Even in GPT-40, large vocabulary: many words do not have corresponding tokens, especially amongst the content, open classes. Maximum representation was below 70%
- Tokenization likelihood varies by word category
- Tokenization likelihood ~ frequency → LLM vocabularies are conservative, normative ["flattening"]
- Unusual information in the tails
- → Tokenization as a cultural artifact?





Figure 15. These charts shows the proportions of parts of speech among the words in CSW19 and GPT-40. The last chart shows the ratio of the number of words in that category in CSW19 that ended up being tokens in GPT-40's vocabulary.

Examples of presumably sub-optimal semantic building blocks



185118 _Watch Japanese porn videos for free 116852 中国福利彩票天天 116852 China Welfare Lottery Everyday 128031 久久免费热在线精品 128031 Long-lasting free hot online boutique 154809 无码不卡高清免费v 154809 Uncoded high-definition free v 172750 大发快三大小单双 172750 Dafa Fast Three Size Odd and Even 177431 给主人留下些什么吧 177431 Leave something for the master 181679 gg的天天中彩票 181679 gg wins the lottery every day 184969_日本一级特黄大片 184969 Japanese first-class pornographic blockbusters 187822 大发快三开奖结果 187822 Dafa Quick Three Lottery Result 49649 彩神争霸邀请码 49649 Caishen Hegemony Invitation Code 89409 免费视频在线观看 89409 Free videos to watch online 122333 无码不卡高清免费 122333 Uncensored HD Free 122712 Uncoded area one, area two, area three 122712 无码一区二区三区 128600 大发时时彩计划 128600 Dafa lottery plan 133274][:]["][133274] [:] ["] [135161 大发时时彩开奖 135161 Dafa lottery results 149168 大发时时彩怎么 149168 Dafa lottery 160029 Dafa Kuaisan is the country 160029 大发快三是国家 160131 大发快三是不是 160131 Is Dafa Fast 3 160267 天天中彩票网站 160267 Lottery winning website every day 176039 Boutique Zone 1, Zone 2 and Zone 3 176039 精品一区二区三区 186348 大发快三是什么 186348 What is Dafa Kuaisan? 187516 大发快三走势图 187516 Dafa Fast 3 Trend Chart 187810 在线观看中文字墓 187810 Watch Chinese subtitles online 191179 大发快三怎么看 191179 How to watch Dafa Kuaisan 193825 中国特色社会主义 193825 Socialism with Chinese Characteristics 194062 彩神争霸是不是 194062 Cai Shen Zheng, right?

Figure 18. Here are examples of the longest Chinese-language tokens in the file o200k_base for GPT-40 (via tiktoken), run through Google translate. Their content suggests the kind of training material used during tokenization.

What does LLM "meaning" look like?

- Main method: UMAP
 + PCA of token vectors
 in RoBERTa (MLM)
- Organization by context, frequency, syntax, semantics...
- Distinction between manmade and natural
- "bank of [NP]" otherwise identical: not strictly necessary to output that text
- Tantalizing hints of an underlying world model





- Some of the clusters seem to reflect fundamental human-like distinctions such as prototypical usage, embodiment, and concreteness...
- Relevant to the Swadesh list and processes of metaphorical extension
- A cluster focused on run's temporal ending (Aktionsart), as in stop ("run into a wall")
- → Time and space are core dimensions to the human experience and universally made salient by the technology of language (as implemented by us... but maybe that could be otherwise!)
- \rightarrow Indirect grounding?



Token extispicy

- Clusters reflect syntactic, semantic, and distributional information
 - Consistent with the DH's syntagmatic information
- Polysemous words were immediately visually distinguishable from more monosemous words
- Both inter- and intra- cluster structure
 - Identifiable sections of specificity (as to word meaning, syntax, frequency) within a cluster with a broader theme, and across or between clusters
- Fine-grained distinctions such as between relational, descriptive, restrictive, and attributive roles, as well as distinctions in construction, and possibly even between idiomatic vs. productive uses
 - As well as grouping by frequency
- This method could provide a quantitative way to identify prototypical usages and to count and classify word senses; there are many potential applications. You could train a classifier on the utterances in each desired cluster, and then probably use that classifier to get fine-grained semantic and syntactic parsing within a novel text. [Shout-out to Collin Coil's embed-then-classify project!]



A Big Deal: LLMs validate and isolate the DH

- Because we know LLM architecture and what training data they are exposed to, we can conclude that whatever they do know can be conveyed via the DH
 - Although through relations between tokens, rather than words
 - At least at the input level and as regards the linear algebraic operations within the model ["flattening"]
- The DH is sufficient for reasonably human-like language performance (MVP language)
 - LLMs demonstrate, for the first time, that there is a minimally viable (MVP) language technology
 - Language can be implemented outside of humans
 - Language can be non-trivially disentangled from many hallmark properties of human cognition
- This is important to note in the face of AI hype, one core promise of which is "we have a machine that can talk, therefore we have a machine that can think": "therefore" is not licensed
 - LLMs show that it would be plausible to have solely a machine that can talk
 - This warning also applies to context-specific modifications to the model such as alignment and fine-tuning
 - Strategies that do not attempt to reconcile the fundamental gnogeographic constraints of the model with the intended task are unlikely to be successful
 - Plausible text ≠ accurate text

Information Technology

4	Info	ormatio	on technology	215
	4.1	Groun	ding	217
	4.2	Lingui	stic meaning	222
	4.3	Ethics		223
		4.3.1	The ouroborous and the crucible of mediocrity	223
		4.3.2	Cultural products, art, and the tech ecosystem	224
		4.3.3	What is art?	228
	4.4	Techno	ology and fundamental dimensions	234
	4.5	Stories		246
	4.6	Chara	cter space: What matters to people? What matters in stories?.	251
		4.6.1	Methodology	253
		4.6.2	Results: archetypes	269
		4.6.3	Stories exploit character space strategically	279
		4.6.4	Connections with other models	285
	4.7	Discus	sion	312
		4.7.1	Lexinvariant LMs	312
		4.7.2	Flattening and fuzzing up frequency and truth	317
		4.7.3	Look-and-Say Sequence	321
		4.7.4	Relation and the identity-difference spectrum in meaning-making	1g325
		4.7.5	Things are like other things (TALOT) vs. Only the thing is the	0.
			thing (OTTITT)	327
		4.7.6	Creativity. Leaps of faith	330
	Bibl	iograph	N	332



Pattern-to-thought transition: distributional pattern \rightarrow Linguistic meaning \leftrightarrow World knowledge?

- What is salient (including perceptible) is different for LLMs than for people, even when given identical information
- But a lot can be conveyed diegetically via language (to them, to us, in general)
 - Sahlgren, 2008 described this idea as "what is internal to language"
 - Harris, 1968 described this as the portion of meaning that "has a purely linguistic aspect"
- There is reason to believe LLMs can make use of diegetic approximations
 - Starting with distributional patterns, they can build out word meanings; they can then acquire world knowledge through language (at least in a rudimentary way)
 - World knowledge \leftrightarrow cognition?



Mechanisms: inferential and referential lexical competence

- Speculatively, meaning is successively bootstrapped
- The syntagmatic information lays the groundwork for semantic information
 - The relations are the bottom-most guiding force but semantic themes emerge from that
 - For example, we see a cluster of e.g. largely PP (prepositional phrases), meaning syntax is basically their highest level of organization, but within the cluster what looks like a vertical spatial relationship subdivides it; maybe with more instances, more such semantic structure would emerge
 - We think this fits with our discussion of the DH
- Symbol grounding: fundamentally, how can LLMs connect linguistic knowledge to world knowledge?
- The DH is (basically) sufficient to establish Marconi's inferential lexical competence
 - Harris' "purely linguistic" meaning





Roles of reference

- Referential lexical competence
 - Grounding: connects "purely linguistic" (diegetic) to the world (extradiegetic) via grounding
 - Compression: creating new words, usurping information locality constraints
- Grounding strategies use another mode to enhance the meaning available to the model
- → Strategies like contrastive linguistic-visual objectives (as in language-vision multimodal transformers) can create contextually "better" language representations
- → Augment inferential competence (learned through the DH from text) with limited referential competence acquired by the pairing of visual and linguistic information (visual sensory grounding)
- \rightarrow Richer meaning
- Meaning bridges linguistic and world knowledge; meaning needs both!
- Why no one can agree on meaning's boundaries, and why people who argue for stochastic parrots or against the DH feel unsatisfied by MVP, less-than-complete meaning these models currently possess
 - DH and small talk? Similarly, with semantic bleaching, sometimes a word like awful or terrible really hits, and moves you deeply. But mostly the level of engagement is more shallow.
 - Maybe the DH, even the inherent degrounding of the fungibility that allows language to work in the first place, is constantly wearing down the meanings of words, and we use other strategies to compensate, to rebuild. Function of poetry?



Stories: an information technology

- Another level of relational, metaphorical, referential technique
- Compression, many-to-one
- Symbiotic with language
- From Vol. 1 of Terry Pratchett's *the Science of Stories:* "A little narrativium goes a long way: the simpler the story, the better you understand it. Storytelling is the opposite of reductionism: 26 letters and some rules of grammar are no story at all."
- From Vol. 2: "Narrativium is powerful stuff. We have always had a drive to paint stories on to the Universe... Moreover, the rules of the universe have to be able to produce everything that we humans observe, which introduce a kind of narrative imperative into science, too. Humans think in stories..."
- Stories consolidate individual behaviour and beliefs to an easily transmissible format on a societal level, which can be applied to each individual approximately equally easily.
- Stories also seem to be particularly cognitively sticky (for us); there are things we remember about stories extremely clearly and enduringly.
- Another noteworthy aspect of stories is that their power is not in their literal depiction of reality: stories continue to guide people in every aspect of their lives even when those stories have conclusively been proven false (Riecken et al., 1956).
- Point of difference with LLMs



Character Space

- Main method: SVD of survey data
 from openpsychometrics.org
- By analyzing stories across film, television, and literature, we uncover 6 primary and 6 secondary base archetypes for fictional characters, which we extend to a framework of single, dual, and triple archetypes.
- These archetypes encompass real personality traits
- The six major archetypes align with the three dimensions of essential word meaning: power, danger, and structure (PDS)
- Our character space archetype framework has broad implications for the analysis and creation of stories
- Stories = Characters + Time



Primary essential dimensions:

Archetypes

We find that twelve archetypes afforded by six essential dimensions dominate the makeup of characters.

#. Primary archetype pair	Essential Meaning	Five factor model	% (#) with Lead	% Variance
\sim Differentials / Locus	(Ousiometrics)	dimensions	Dimension	Explained
1. {Fool/ \Leftrightarrow Hero/}	{weak⇔powerful}	$+\{$ conscientiousness $\}$	39.5% (12+778=790)	24.4%
$\sim \{{\rm weak/incompetent/lazy/stupid} / $	\Leftrightarrow powerful/capable/purpos	seful/intelligent ⁷ } / Internal		
2. {Angel/ \Leftrightarrow Demon/}	${safe \Leftrightarrow dangerous}$	$-\{agreeableness\}, +\{extroversion\}$	27.3% (228+318=546)	20.4%
$\sim \{\text{safe/pure/virtuous/humble}/ \Leftrightarrow d$	langerous/depraved/corrupt	/arrogant/} / Social		
3. {Traditionalist/ \Leftrightarrow Adventurer/	$\{ structured \Leftrightarrow unstruct \} $	$tured\}+{openness}, +{extroversion}$	19.2% (62+323=385)	14.6%
\sim {serious/predictable/humorless/u	$ncreative \leftrightarrow playful/unpress$	dictable/funny/creative/} / Internal		
		Totals for dimensions 1, 2, and 3:	86.0% (302+1419=1721)	59.4%
Secondary essential dimensions	s:			
#. Secondary archetype pair		Five factor model	% (#) with Lead	% Variance
\sim Differentials / Locus		dimension(s)	Dimension	Explained
4. {Lone Wolf/ \Leftrightarrow Diva/}		$+\{extroversion\}$	5.7% (22+91=113)	6.2%
\sim {rugged/stoic/independent/blunt	$\uparrow \Leftrightarrow refined/dramatic/dependence$	ndent/sensitive/} / Social		
5. ${Outcast \land \Leftrightarrow Sophisticate \land}$		$+\{extroversion\}, -\{neuroticism\}$	5.6% (112+0=112)	5.1%
\sim {unlucky/unsophisticated/traumaticated/	$atized$ \Rightarrow fortunate/sophist	icated/confident/} / Social		
6. { Brute ^{$/ \Leftrightarrow$} Geek ^{$/$} }		$-\{extroversion\}, +\{neuroticism\}$	1.9% (27 + 11 = 38)	3.8%
\sim {physical/mainstream/simple-min	$nded \land \Leftrightarrow intellectual / weird / det and a second $	complex/} / Internal		
		Totals for dimensions $4, 5, and 6$:	13.2% (161+102=263)	15.1%
		Totals for dimensions 1–6:	99.2% (463+1521=1984)	74.6%

Trait-level essential dimensions:

Unnamed non-Archetype Complex Essential Traits		% (#) with Lead	% Variance
#. \sim Differentials		Dimension	Explained
7. ~ {dramatic/attractive/young/ \Leftrightarrow comedic/ugly/old/}		0.5% (5+4=9)	2.1%
8. ~ {spiritual/rural/historical ⁷ \Leftrightarrow skeptical/urban/modern ⁷ }		0.2% (0+5=5)	1.6%
9. ~ {old/historical/low-tempo ^{\prime} \Leftrightarrow young/modern/high-tempo ^{\prime} }		0.1% (2+0=2)	1.4%
10. ~ {feminine/luddite ^{<i>i</i>} \Leftrightarrow masculine/technophile ^{<i>i</i>} }		0.0% (0+0=0)	1.1%
11. ~ {secondary/street-wise/ \Leftrightarrow primary/sheltered/}		0.0% (0+0=0)	0.9%
	Totals for dimensions 7–11:	0.8% (7+9=16)	7.2%
	Totals for dimensions 12–464:	0.0% (0+0=0)	18.3%
	Totals overall:	100.0% (2000)	100.0%

Authors/ fictional universes exploit character space strategically

- Authors/ fictional universes exploit character space strategically in ways that support their narrative goals, somewhat analogous to how languages exploit vowel space
- Each character's composition contributes to the story through its similarities and differences with the other characters
- The space of actions and motivations that are available to drive the plot forward are determined by the span of the characters









PDS = PPP

- Vectors/ relations of power
- Power, danger, and structure → power within, power without, power between
 - A "power" sense for danger is attested by its derivation: the word "danger" is derived from Middle English "daunger" (meaning power, dominion, peril) via authority, power, from Latin "dominus", meaning "lord, master"
 - Connecting power, and PDS, to the Trifunctional Hypothesis
- Power is really important in stories, words; language → important to people

DILIGENI	HUMBLE	SCHEDULED	PHEPPY	NEXT DOOR	NEKD	OLU	SPIRITUAL
PERSISTENT	FORGIVING	STICK-IN-THE-MUD	MANICURED	ANXIOUS	BOOKISH	REPULSIVE	HISTORICAL
MOTIVATED	NURTURING	UNCREATIVE	TAILOR	TENSE	INTELLECTUAL	COMEDIC	THEIST
DRIVEN	COOPERATIVE	CORPORATE	PRIVELEGED	TRAUMATIZED	CATPERSON	UGLY	LOW-TECH
COMPETENT	ANGELIC	SERIOUS	REFINED	CHEESY	QUEER	NERD	RURAL
HIGH IQ	WARM	HUMORLESS	SHELTERED	VINTAGE	FREAK	DORKY	COUNTRY BUMPKIN
RESOURCEFUL	ALTRUISTIC	SERIOUS	IVORY-TOWER	SLOVENLY	CRYPTIC	CHATTY	LUDDATE
PERCEPTIVE O	RESPECTFUL	MONOTONE	RICH	UNLUCKY	INTROVERT	ASEXUAL	POETIC
WORKAHOLIC	EMPATH	GERIATRIC	BOURGEOISIE	UNPOLISHED	WEIRD	CHEESY	OLD
SELF-DISCIPLINED	WHOLESOME 6	STOIC	METROSEXUAL	AWKWARD	CREATIVE	WEIRD	IDEALIST
PRO	WELL-BEHAVED	DELIBERATE	FEMININE	THRIFTY	METROSEXUAL	HAPPY	GENUINE
GO-GETTER 0	PURE	RESERVED	FLOWERCHILD	BLUE-COLLAR	MYSTERIOUS	MASCULINE	FANTASTICAL
slugabed 💛	DEBASED	CHATTY	GOTH	IVORY-TOWER	UNAMBIGUOUS	FEMININE	REAUSTIC
NOOB	MISCHIEVOUS	SPONTANEOUS	MASCULINE	EXTRAVAGANT	MACHO	SAD	SARCASTIC
DISORGANIZED	SALACIOUS	EXPRESSIVE	MACHO	CHARMING	CONVENTIONAL	NORMAL	REALIST
SLACKER 26	PSYCHOPATH	VIBRANT	PROLETARIAT	ELOQUENT	NORMAL	CHIC	YOUNG
UNOBSERVANT	RUDE	EXPRESSIVE	POOR	FORTUNATE	EXTROVERT	SEXUAL	FACTUAL
HELPLESS	SELFISH	PLAYFUL	BLUE-COLLAR	STYLISH	STRAIGHT -FORWARD	RESERVED	TECHNOPHILE
LOW IQ	QUARRELSOME	FUNNY	STREET-SMART	TRENDY	NORMIE	COOL	CITY- SLICKER
INCOMPETENT	DEMONIC	BOLD	RUGGED	CHIC	STRAIGHT	JOCK	URBAN
UNAMBITIOUS	COMPETITIVE	FREELANCE	OPRESSED	FLOURISHING	DOG PERSON	BENJTIFUL	HIGH-TECH
UNMOTIVATED	POISONOUS	OPEN TO NEW	BLACKSMITH	RELAXED	PHYSICAL	DRAMATIC	ATHEIST
QUITTER	VENGEFUL	ADVENTUROUS	SCRUFFY	CALM	SPORTY	ATTRACTIVE	MODERN
LAZY	ARROGANT	Spontaneous	PUNKROCK	CELEBRITY	JOCK	YOUNG	SKEPTICAL
3=2=1 holy, he	y, holy"- the seraphim		L Control		TER A		
3 1,3,2	6 wings : 3 pairs = 1 →1 pair = 11	being	A PARTY		N. S. S.	N2T	N AN
Dualities and trinities	1 pair=2w 1=3 one holy trinity	angs and		1013			
good and evil	A 1=2 Jesus' two natures				R. Mark		and the second
the DNAR the ghortal	3 parts 2 entities			· AKA		- A	
nothing, something	pawer danger struct	fure			States.		
father, son	A Stress 2		1 Parto		图 很	P. Nes	
B	6=1+2+3	3 aba			A. Brack		
	8-2=6 8-3=5	trak?					
1=2 ABCABCAB POWER DANGER	within {2=1 } 3=1 5,36,8		Bath		1.1	1.000	
3=1 ABC	RE between Jorientations, Vector						
2=3 1=2 CA Dumezil: Warrior	s vs firmament	the max of	A Starting	1 Add	4 .	Sec.	
	as, and and a concerned	NEAD TO AND A DECK	A TRANSPORT	Sec. A Ch	Farsha Brit		

Technology

- Language plays a huge role in our cognition
- Making information from the cognition of the past transmissible to the present
- Manipulation of locality across time, space, and self
- Language allows us to manipulate what is functionally local, changing what can be salient, an incredibly powerful mechanism across the spectrum of information-related technology, if not across technology writ large
- We can imagine that potential, or power, the ability to effect, is the goal of any technology
- The main strategies we employ could be divided into (1) information, a fight for stability despite a dispersive medium (Levin, 2024), and (2) transformation, change, such as fire.
 - Technology is related to power. Formation is the root of both (1) information and (2) transformation.
 - Congruent with "knowledge is power", connects search for meaning with a broad range of innovations across 0 life, all seeking to increase the scope of what is salient to them - increasing their power, their salioscape
 - (1) includes information, coordination, or structure, such as reproduction and language. 0
 - These strategies may have different properties. For example, information seems to support emergence. 0
 - (1) and (2) can also be described as stability versus change, or even structure versus chaos/ unstructuredness.
- All manipulation of information, perhaps even down to coordination between cells and structure itself, is a continuum.
- Salioscape extends the concept of gnogeography with embodiment The complete landscape of what can affect and be affected, the entire network of influence at the fuzzy interface between self and environment.
- Describing power in terms of spheres of influence leads us to a provisional definition of technology as a repeatable, modular mechanism for manipulating fundamental dimensions, thereby increasing the salioscape — or the scope of what can be made salient, which requires reachability, or locality — for any system employing the mechanism.
 - This definition can encompass biological problem-solving processes, such as heterochrony (shifts in developmental timing) and heterotopy (shifts in spatial expression), as natural examples of technological innovation.
 - The development of our physical senses could also be understood through this lens: vision, for example, offers 0 power through manipulation of space (and, indirectly, time). Sight allows us to connect with, to touch, at the speed of light entities in the environment that would otherwise remain distal.
 - But vision is more bound by the self than language, because of the much lower 0 fungibility of the format of the information





Language as a technology for people

- At the most basic level, aspects which are perceptually and practically salient to human life are the semantic building blocks we combine to create more complex meanings
 - For example, time and space are universally encoded in language (Rijkhoff, 2002), and concrete and embodied usages are anchor points in metaphorical processes (Maudslay et al., 2024; Lakoff, 2008)
 - Power is a building block in stories, via fictional characters; words
- For LLMs, with such curtailed bodies and access to the world, language is the most important medium
 - Indirect grounding for time, space, even perhaps rudimentary embodiment for LLMs – so how about power? What else?
- Language is also incredibly important for people, but it has a much richer life beyond its MVP linguistic components: it is used in cultural contexts and artifacts
- People have created many techniques that make use of language, such as storytelling, and jurisprudence (more generally, speech acts)
- We speculate that such technologies are likely partly outside of MVP language, making use of MVP language and scaffolded by it (and enriching the MVP meanings available to words, as discussed with grounding)





Language as an abstract technology

- The technological mechanism of linguistic meaning is purely relational (Harris 1954, 1968; Sahlgren 2008)
 - Potentially doesn't need othermodal reference at all
 - Language is not implemented for us in such a wholly degrounded way (and it's hard to imagine talking without talking about something)
 - And not how we implemented it for LLMs, as they (like us) seem to concurrently build a world model, and get indirect grounding through our language (e.g. time and space).
 - Although reference is still required for the compression achieved through lexical encoding, which is needed to work around our architectural processing constraints.
- The fundamental linguistic unit (when considered as a theoretical technology) is therefore metaphor, a relation, not a semantic primitive
 - This ties the construction of linguistic meaning to representation, abstraction, and perspective (even self), more generally
 - Compression and reference are meaning-making techniques in language founded on relations (metaphors) of contrast (e.g.bits: anything propositional or descriptive can be encoded via only two symbols, something and not-something)
- Reference, at the lexical level and above, supports our cognition and communication because it would be impossible to keep many complex concepts and propositions sufficiently local (recall discussion of information locality constraints in language) without metaphorical extension and reference, which allow them to be compressed
- Symbol grounding: Hebbian learning, $DH \rightarrow$ meaning



Information technology: relation + locality is all you need?

- Leverages structure to transcend the original bounds of time, space, and/ or self --- the factors constraining locality --transmitting information beyond its current location
- Changing the bounds within which information can be salient (which requires it to be local)
- Locality is integral across many information technologies; universal property?
- Examples
 - 0
 - In deep learning architectures, in pooling and convolutional layers and skip connections Information locality is also embedded in the structure of language (in Behaghel's first law of syntax) and <mark>in the DH</mark> W. Tobler's aphorism that although all things are related, things that are closer together are more closely related Ο
- Abstract representation of information locality that allow more literal proximity to be leveraged in new, even metaphorical, contexts.
- CNNs leverage their architecture to allow groups of pixels to relate both to nearby and remote pixels, a transcendence of literal closeness through abstract closeness, which is literal in some representational space.
- Skip connection: enables features from layers that are not proximal in the architecture to be transmitted (almost) directly; skip connections make loss landscapes much more convex and enable greater generalizability (Li et al., 2018)
- These examples preserve Tobler's maxim
- Taking Tobler's maxim a step further, we can connect locality with the definition of technology as a means of increasing power over a fundamental dimension
- If life is about persistence despite a dispersive medium (Levin, 2024) \rightarrow everything we do is about trying to subvert locality: persistence is inherently non-local
- Technology expands the domain we are persistent in, our functional salioscape; this is power
- Metaphor, reference, and representation are all techniques under the same relation-for-locality umbrella.
- Metaphor and representation pick out salient features in one perspective and apply them in another, moving information local to one context into another, thus enhancing cognition
- Reference allows both compression and deepening of meaning through representation
- Relation is the most fundamental, rudimentary piece shared across all these techniques: the mechanism that allows these feats, the atomic locality expander
- Relation is vital to how patterns become thoughts

5.1	Précis	of the Thesis				2	
5.2	Tokeni	zation and its consequences			8	28 - 23	
5.3	Future	of Learning Machines			-	10	
	5.3.1	Promising signs				-	
	5.3.2	Call to theory					
	5.3.3	Composition				1	
	5.3.4	Abstraction				2	
	5.3.5	Separability				•	
	5.3.6	Stories					
	5.3.7	Selfhood, agency, and creativity.					
	5.3.8	Open-endedness and meta-cognition	on				
5.4	Inform	ation technology					
5.5	Relatio	ons					
5.6	Chang	e					
5.7	A Cav	eat: Structure and Energy					
5.8	Future	work					
	5.8.1	Deep Learning				•	
	5.8.2	Metascience				•	
	5.8.3	LLMs				•	
	5.8.4	Character Space				•	
5.9	Art an	d the Luddite				•	
5.10	Limita	tions				•	
	5.10.1	Hasty Generalization Limitations				•	
	5.10.2	Vocabulary Limitations				•	
	5.10.3	Extispicy Limitations		Ĵ			
	5.10.4	Character space limitations		Ĉ) L		
	5 10 5	General Limitations		^			Ì



Promising signs for AI: pattern-to-thought transition?

- (1) Models can still learn surprisingly well when given only highly-constrained, bizarre levers to pull (Frankle et al., 2021)
 - Plasticity "free[ing] the downstream (future) agent from the restrictions of the upstream (past) agent" reminiscent of the resilience of biological systems (Levin, 2025)
- (2) Forcing an LLM to allocate more resources to generating its output (through strategies like Chain or Tree of Thoughts) can improve its output (Yao et al., 2023)
 - The steps in that reasoning process do seem to matter (it isn't just that more tokens generated is always better; it matters which tokens are generated) (Xia et al., 2025)
- (3) LLMs seem to be able to acquire additional world knowledge through diegetic approximations (Chang et al., 2024)
 - Implies there is space between the world knowledge necessary for MVP language and what it is possible to know through the medium of language
 - Given that current LLMs clearly possess MVP language but are missing much knowledge and common sense, they are somewhere within that space
 - Size and shape of the upper bound LLM gnogeography is unknown
- (4) Multimodal grounding can create contextually better representations
- There are alternatives to why these things might occur other than "the model is thinking", but we're not sure whether or not we can tell the difference between those alternatives and thinking
 - Undermining the stochastic parrot argument



LLM implementation of language

- Language as an abstract technology → there are LLM-specific implementation details, with attendant benefits and costs
- Gnogeography (cognitive landscape) of large language models (LLMs) is not based on the human body
- LLM implementation of language need not have the same embodied limitations
 - Less noisy
 - Opposite of earlier point
- TokenSkip compression (Xia et al., 2025) pushes past the redundancy of a written language derived from a spoken language designed both for communication and thought, shifts emphasis to a symbolic language scaffolding cognition
- Decoupling has profound implications for understanding the ceiling of LLM cognition, the limits of information transmission through sequential data, and the abstract implementation of language as a technology
- Re-evaluation of redundancy in human language
- Link between reasoning and speaking, central to theories of linguistic relativism and "thoughts are thinkers" (Levin, 2024)



Let's end on some *useful takeaways* for a general audience!

- Humans are always in the loop. Some risks of LLMs are more accurately framed as risks of the system using them, which includes both people and LLMs.
 - People are the ones creating the data, designing and building the models, and deciding how to use the models.
- Different architectures are different. At the same time, piling more things blindly on top of an LLM while never addressing the cause of a particular issue is not a solid foundation for success.
- Language is a beguiling medium. We are used to imputing a human mind to the speaker whenever we encounter human-like language. We cannot do that with LLMs.
 - We must learn to apply a different Theory of Mind.
- \rightarrow Be wary of benchmarks!
 - Games of plausible text: represent a complex task as a game of generating plausible text → ask the generating-plausible-text machine to generate plausible text → proclaim that the machine can perform the original task



Important takeaways...

- A machine that can talk \Rightarrow a machine that can think
- Frequency, grammaticality \neq truth. But for an LLM, these are essentially one and the same.
- Plausible text ≠ accurate text
- Beware flattening. LLMs have trouble with unusual data. For many reasons, baked into the model at many levels. As a result, they can be conservative, homogenizing, and reductive.
 - Including training data, tokenization, objective functions, alignment, and perception/ embodiment
- Resist the invocation of "associative learning", "pattern recognition", "probabilities", "next token generation" as thought-terminating cliches
 - Deep learning models are doing interesting things and it is not necessarily unwarranted to use words like "learn", "think", and "understand" to describe them
 - \circ $\hfill We also learn by association: the DH is a special case of Hebbian learning$
 - What is the alternative explanation for a behaviour (e.g. Chain-of-Thought)? Is the alternative truly distinguishable?
- Beware squishy dichotomies
 - No clear border between linguistic knowledge and world knowledge
 - Or architecture and data, thoughts and thinkers
 - Or coordination, communication, and cognition
 - Or individual and collective knowledge
- → Language is important for both thought and communication [in people and LLMs]

*Two pseudoparadoxes in here but I don't think there will be time to get into that.







The end, thank you!



6	App	endix	506
	6.1	Terminology	506
	6.2	Some things we tried that didn't work	508
		6.2.1 Automatic extraction of character space from texts	508
		6.2.2 Saucier dataset	514
		6.2.3 Token Length + Frequency	516
		6.2.4 Wild speculation	517
	6.3	Supplementary vocabulary results	519
		6.3.1 Frequency + Length	528
		6.3.2 Affixes	529
		6.3.3 Function words	532
		6.3.4 Proper Nouns	534
		6.3.5 Pronouns	535
		6.3.6 Nouns	535
		6.3.7 Adjectives	535
		6.3.8 Adverbs	535
		6.3.9 Verbs	535
		6.3.10 Interjections	535
		6.3.11 Prepositions	535
		6.3.12 Conjunctions	535
		6.3.13 CSW19	535
		6.3.14 Iconic words	538
		6.3.15 GPT-40	542
	6.4	Supplementary extispicy results	545
	6.5	Supplementary character space results	548
	6.6	Supplementary discussion	550
		6.6.1 Lexinvariant LMs	550
		6.6.2 Indirect grounding	553
		6.6.3 Information technology	554
	6.7	Supplementary figures	557
	6.8	Supplementary future work	557
	Full	eference list in alphabetical order	557

Abstract

In this thesis, we explore language and cognition in both people and in computational models, through the lens of meaning construction at both the individual and collective level. We step through levels of linguistic structure and large language model (LLM) architecture – from morphemes to stories – in order to (begin to) build an understanding of the role these concepts play in human and machine cognition, and, in particular, in meaning construction and symbol grounding.

In Chapters 2 and 3, we interrogate the interface between LLMs and the world, since it is this interface that facilitates the creation of *idios kosmos* from *koinos kosmos*. Specifically, we consider the role tokens --- the smallest units of textual structure accessible to LLM architecture – play in LLM cognition. We find evidence that even an objective function meaningfully insulated from the main intelligence can impact the system's cognition. We show that different kinds of words behave in different ways as tokens. We relate evidence that syntax, semantics, and frequency can be reflected in token content for LLMs. Based on these findings, we argue that relation is the fundamental mechanism for linguistic meaning construction.

In Chapter 4, we extrapolate from the results of the first two chapters to connect the mechanism of relation (metaphor) to two omnipresent aspects of cognition, locality and salience. We consider a variety of contexts in order to distill these concepts as omnipresent across different scales of biological and non-biological information technologies. To that end, we introduce and exploit the concepts of gnogeography and salioscape. Gnogeography is the sphere of ideas – roughly, the representations, abstractions, models, facts, propositions, and descriptions of perceptual input – that is accessible from the perspective of a being, given its experience and its architecture. Salioscape extends this idea with explicit embodiment. A context we consider in detail is the story:, we present our framework of 6 archetypes to describe fictional characters. This framework aligns with our previous results that elucidate power as a fundamental component of word meaning, which we use in our nascent theory of technology.

In the Conclusion, we present high-level takeaways on cognition, language, and the future of learning machines as well as the science around them, based on the work put forth in the previous chapters.

Relatively tangible results include: (1) a framework for understanding the role tokens play in LLM cognition, anchored in the Distributional Hypothesis (DH); (2) evidence that even an objective function meaningfully insulated from the main intelligence of a model can impact the system's cognition; (3) evidence that different kinds of words behave in different ways as tokens; (4) evidence of what information can be reflected in tokens, e.g. that frequency is among the diverse streams of information falling outside conventional meaning that nevertheless directly impacts token meaning; (5) implications of (1-4) plus other key architectural choices for LLM cognition and its use of language ("flattening"); (6) an argument that LLMs demonstrate for the first time the non-trivial separability of a minimally-viable form of language from other cognitive processes (that is, that thinking and speaking can be significantly de-coupled), and that they validate and isolate the DH; (7) a framework for describing fictional characters in terms of 6 basic archetypes; (8) evidence for a connection between components of meaning across different linguistic structures; (9) the beginnings of a theory of technology in terms of its impact on problem-solving through the increase of power; (10) a theory of meaning construction in language; the theory that relation and locality are the fundamental components of problem-solving (relation being the core mechanism for subverting locality, the core constraint); (11) proposals for future work.

Example: MRNN (Memory Recurrent Neural Network) from 2011, ChatGPT (LLM) from 2024

A character-level MRNN trained on Wikipedia, initialized with "The meaning of life is..." (Sutskever et al, 2011):

"The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the" ChatGPT (OpenAI, 2024), asked to write a paragraph starting with "The meaning of life is...":

"The meaning of life is to find joy in the journey, to seek purpose in our actions, and to connect deeply with others along the way. It is about discovering what brings us fulfillment and dedicating ourselves to pursuits that ignite our passions. Life's meaning is often found in the small, everyday moments that bring happiness and contentment. It is about embracing challenges and growing from them, learning to appreciate the beauty around us, and fostering relationships that provide support and love. Ultimately, the meaning of life is a deeply personal quest, unique to each individual, and it is found through our experiences, our dreams, and the connections we make."

All kinds of textual input are experientially flattened into identical structures

- Many tokens in extant vocabularies don't look well-suited to being semantic primitives
- Many tokens are cased (or otherwise minor) variations of each other
- Quite a few tokens were allocated to formatting
- Numbers were not uncommon tokens either
 - Because the tokenization strategy treats all symbols alike, the number tokens do not directly reflect our conventions (e.g. the base 10 digits, maybe additional milestones like powers of 10, etc.)
 - Rather, they reflect common sequences of numbers, which could come from e.g. telephone numbers, backend software development, website URLs, etc.
- Brings us to a shortcoming of the current architecture (both of tokenization and of the LLM): all kinds of textual input are experientially flattened into identical structures
- All tokens are treated identically, at least at the most basic level
 - With respect to its linear-algebraic operations
 - 0 is treated the same as 713, or 00, which is treated the same as ``cat", the same as ``c", the same as ``Karen", the same as +, the same as =======, etc.
- It is easy to imagine that this flattening would have downstream impacts on model cognition and performance
- For example, with iconic words, which have a significant referential, sensory component that presumably is not amenable to encoding via the DH, and numbers, with which LLMs have notorious difficulty

- More syntax-based than bank and run 0
 - Function word
- However, we didn't think it had no semantic content
 - e.g. Ο vehicle-based region \rightarrow shows that frequency impacts meaning
- Cluster that looked entirely formatting-based, around dialogue... presumably meaning that the strong conventions around format for dialogue were the strongest signal for the LLM, overwhelming any signal of what would conventionally be considered meaning



Power

- Ability to act within some range; sphere of influence; changing the salioscape
- Technology = a repeatable, modular mechanism for manipulating fundamental dimensions (space, time, self), thereby increasing the salioscape or the scope of what can be made salient, which requires reachability, or locality for any system employing the mechanism.
- Connecting power, technology, relation, locality
- Stories surface our concern with power
- Technology is a power amplifier



Birds' eye view: we want to improve LLM performance

- Current models fall short for many of the ways we want to use LLMs
 - It is clear that we need to develop strategies for allowing LLMs to make better use of the data they are exposed to
 - E.g. duplicated training data leads to forgetting
 - Model-generated training data leads to "catastrophic defects" (Shumailov et al., 2024)
 - People, by contrast...
 - Are very good at wringing new information from familiar stories
 - Can sometimes do a lot with a little
 - Poverty of the stimulus, fast-mapping
 - Sometimes LLMs can do a lot with a little, too: lexinvariant LMs
- Data scaling is unlikely to be the sole answer
 - Based on the general observation that unusual information is difficult for models to acquire
 - Lack of access to the tails of the distributions is associated with model collapse
 - And projections that the amount of data we have access to would be insufficient anyway (which underline the inadequacy of that strategy)

Returning to diegetic approximations + AI hype

- With careful structuring of training data and model architecture (e.g., batch size, number of training steps), LLMs do seem to be able to acquire additional world knowledge (Chang et al., 2024)
 - Implying that there is space between the world knowledge necessary for MVP language and what it is possible to know through the medium of language
 - Given that current LLMs clearly possess MVP language but are missing much knowledge and common sense, they are somewhere within that space
 - The entirety of that space is the upper bound of their gnogeography
- However, the size and shape of the upper bound LLM gnogeography is unknown
 - It is plausible that it is incompatible with what would be required for many of the tasks we want LLMs to perform
 - This warning also applies to context-specific modifications to the model such as alignment and fine-tuning
 - Strategies that do not attempt to reconcile the fundamental gnogeographic constraints of the model with the intended task are unlikely to be successful
- This is important to note in the face of AI hype, one core promise of which is "we have a machine that can talk, *therefore* we have a machine that can think"
 - In fact, LLMs show that it would be plausible to have solely a machine that can talk