Roger Friedman
LING 3100 Dr. Kim
May 2026

Representation of Linguistic Concepts within Large Language Models (LLMs)

Table of Contents

I. Introduction 2
A. Motivation and Research Question 2
B. Psycholinguistics of Speech Production 3
C. Large Language Models 5
II. Methods and Experiments 8
A. Instrumented Models and Probes 8
B. Sentence Completion: Trajectories 10
C. Sentence Completion: Categories 16
D. Directional Cosines (Cosine Similarity) 18
III. Results and Interpretation 22
A. Language and Meaning 22
B. A Linguistic View of LLMs 23
C. Missteps and Next Steps 24
D. Working with Artificial Intelligence 26

Table of Figures

Figure 1 -- Levelt's Model of Speech Production (Towell & Dewaele 2005:211) 4
Figure 2 -- Simplified Architecture of LLMs 6
Figure 3 -- Metrics of LLMs 7
Figure 4 -- Simplified Architecture with Instrumentation 9
Figure 5 -- Transformer Explainer (https://poloclub.github.io/transformer-explainer/) 10
Figure 6 -- Sentence Completion: Probe output exemplar 11
Figure 7 -- Sentence Completion Trajectory: Base Case 12
Figure 8 -- Sentence Completion Trajectory: Zero Action Case 13
Figure 9 -- Sentence Completion Trajectory: Atypical Event Case 14
Figure 10 -- Sentence Completion Trajectory: Ambiguous Context Case 15
Figure 11 -- Category Definitions 16
Figure 12 -- Sentence Completion Categories: Base Case 17
Figure 13 -- Layer by Layer Analysis 17
Figure 14 -- Overlapping Categories 19
Figure 15 -- Rotation of Hidden State 20

> *As [LLMs] are gaining prevalence, one may point to two trends. First, some push back against the abandonment of linguistic knowledge and call for incorporating it inside the networks in different ways. Others strive to better understand how [language] models work.* (Belinkov & Glass 2019:49)

I. Introduction

A. Motivation and Research Question

This work responds to criticisms of Large Language Models (LLMs) by the linguistic community, including the following:

- LLMs do not directly represent meaning but infer it from statistical regularities in distributional patterns.
- LLMs model linguistic behavior without modeling linguistic competence.
- LLMs encode only the surface form of language.
- LLMs do not capture the structure of a language (syntax, semantics, pragmatics).
- LLMs have difficulty in resolving ambiguity.
- LLMs do not utilize the compositional nature of human language .
- LLMs approximate linguistic behavior, simulating meaning rather than encoding it in a stable, interpretable form.

Some of these criticisms date back to the first "connectionist" models (Minsky & Papert 1969). Some are much more recent (Chomsky, Roberts & Watumull 2023). Most are premised on the fact that LLMs are not based on the prevailing analytic linguistic models.

But my intuition is that any valid theory of linguistics should be found within any other valid theory of linguistics to the extent the scopes of the two theories overlap. For this paper, that intuition takes the form of the following hypothesis:

If a linguistic concept corresponds to a genuine, stable feature of language use, then it should appear in a large, modern language model as a recurrent directional tendency in the model's representation space.

While no single paper can address every feature of language use, this paper seeks to lay the groundwork for a series of linguistic experiments by establishing the following propositions:

- LLMs are a different model of language than the descriptive, analytical theories to which linguists are accustomed, but no more different than psycholinguistic models of speech.
- LLMs are not inscrutable – there are tools available to analyze their operation and internal representations.
- At least some linguistic concepts can be realized and studied within the LLM.

B. Psycholinguistics of Speech Production

The model of speech production proposed by Levelt (1989) (Figure 1) consists of the following elements:

- A conceptualizer which operates at the pre-verbal level, monitoring the speaker's and interlocutors' speech to maintain discourse conventions, veracity, and a shared mental model. It generates messages to be spoken, which must take into account some grammatical features.
- A formulator, which encodes the message into a grammatical structure and generates a phonological program by which speech will be rendered.
- A lexicon containing vocabulary, semantic, syntactic, and phonological characteristics for word usage, accessed via association. It is used by the formulator and the speech comprehension system.
- An articulator which carries out the phonological program received from formulator.

For present purposes, it suffices to note that none of the structures in this model are dedicated to reproducing features of any analytical model; its purpose is to define the architecture of speech production. Yet Levelt's model has motivated fMRI studies of the brain seeking to locate such structures (Kemmerer 2018). LLMs are likewise models of speech production.

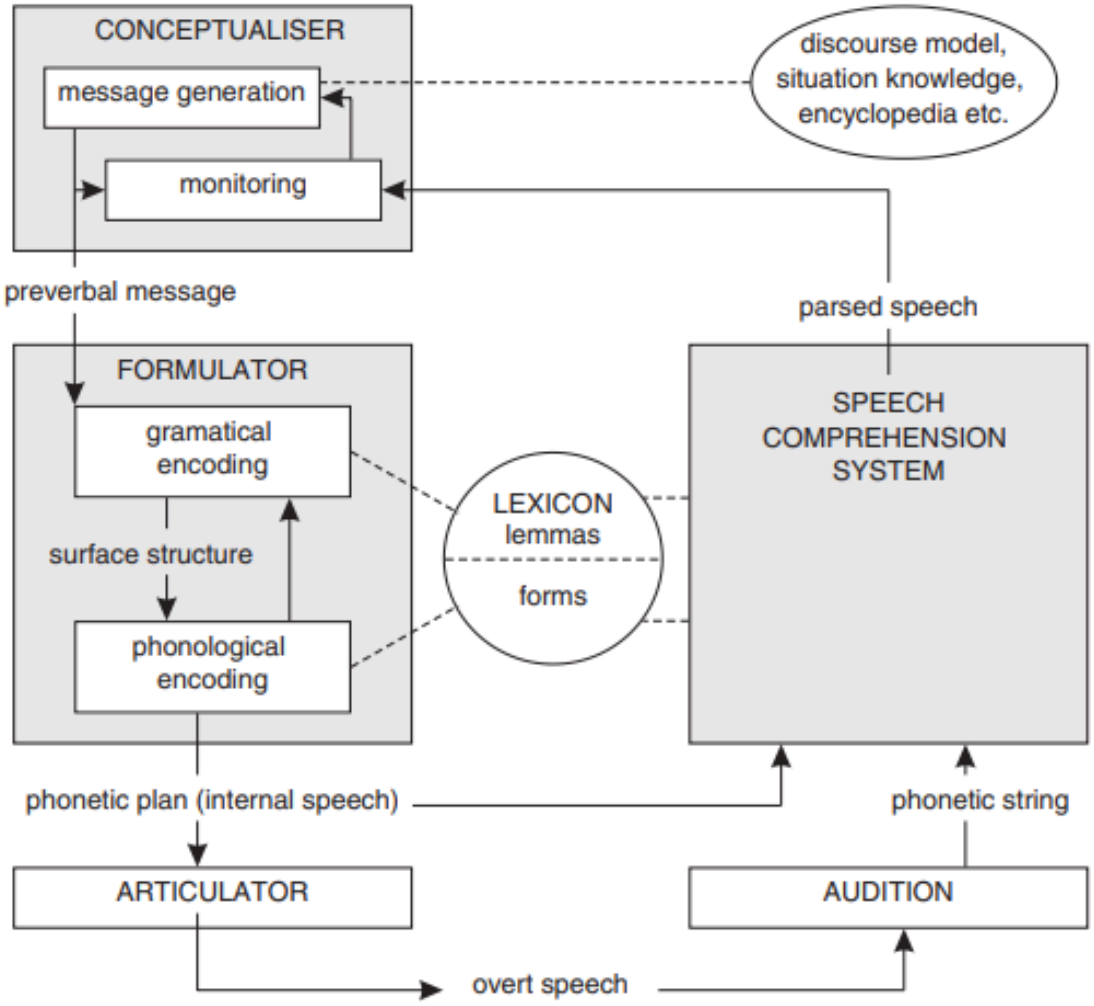


Figure 1 -- Levelt's Model of Speech Production (Towell & Dewaele 2005:211)

A lexicon had been a component of traditional language models even before Levelt, used to describe a repository of lexical form, meaning, and access, and of syntactic properties including thematic roles.

> The area of psycholinguistics that is called mental lexicon is concerned with how words are organized in one's long-term memory. A mental lexicon is like a person's mental storage of words, the meaning of the words and their association. The mental lexicon comprises not only how the words are kept in one's memory, but also how they are recalled throughout the act of writing. ... [I]t is an area. which is less exposed to research and examination.... [T]he words in our mental lexicon are organized into a vast network that is determined by the characteristics of each word. ... Using links between the words, a conceptual network is formed,

> with each combination of words representing a context that the person is familiar with. The words are further joined to create weaker networks, which can be identified by their phonological or orthographic similarity. The mental lexicon's network resembles a massive, multidimensional spider web. (Jabeen & Shahzad 2023:25)

LLMs encode words, affixes, punctuation, and common short phrases (collectively, tokens) as vectors in their embedding space; these are accessed in an associative way. Context-dependent aspects of their use are expressed as geometric constraints produced by the parameters of the model. This will be demonstrated *infra.*

C. Large Language Models

Large language models are large in a number of ways, some of which are beyond the scope of this paper. This section will introduce a simplified architectural model of a LLM and identify some of these out-of-scope features sufficiently to appreciate just how large today's models have become. Training of LLMs is also out of scope. The paper will continue with the simplified model. Readers interested in a more detailed discussion are directed to Karpathy (2023) in the Suggested Readings

The heart of the LLM is the stack of identical layers (Figure 2). Each layer receives a "hidden state"[1] which consists of an ordered set of vectors, each of which corresponds to a token position. As a conversation proceeds, token vectors are accumulated until maximum capacity ("context size") is reached; thereafter, older vectors are discarded. Each vector has a fixed number of dimensions; these dimensions do not correspond to linguistic features but combinations of aspects of features determined during training.

[1] In this paper, "representation space" refers to the fixed vector space defined by the pretrained model, while "hidden state" refers to the context-dependent vector that evolves across layers during processing. Interpretations of meaning are inferred from the trajectory of the hidden state within this space.

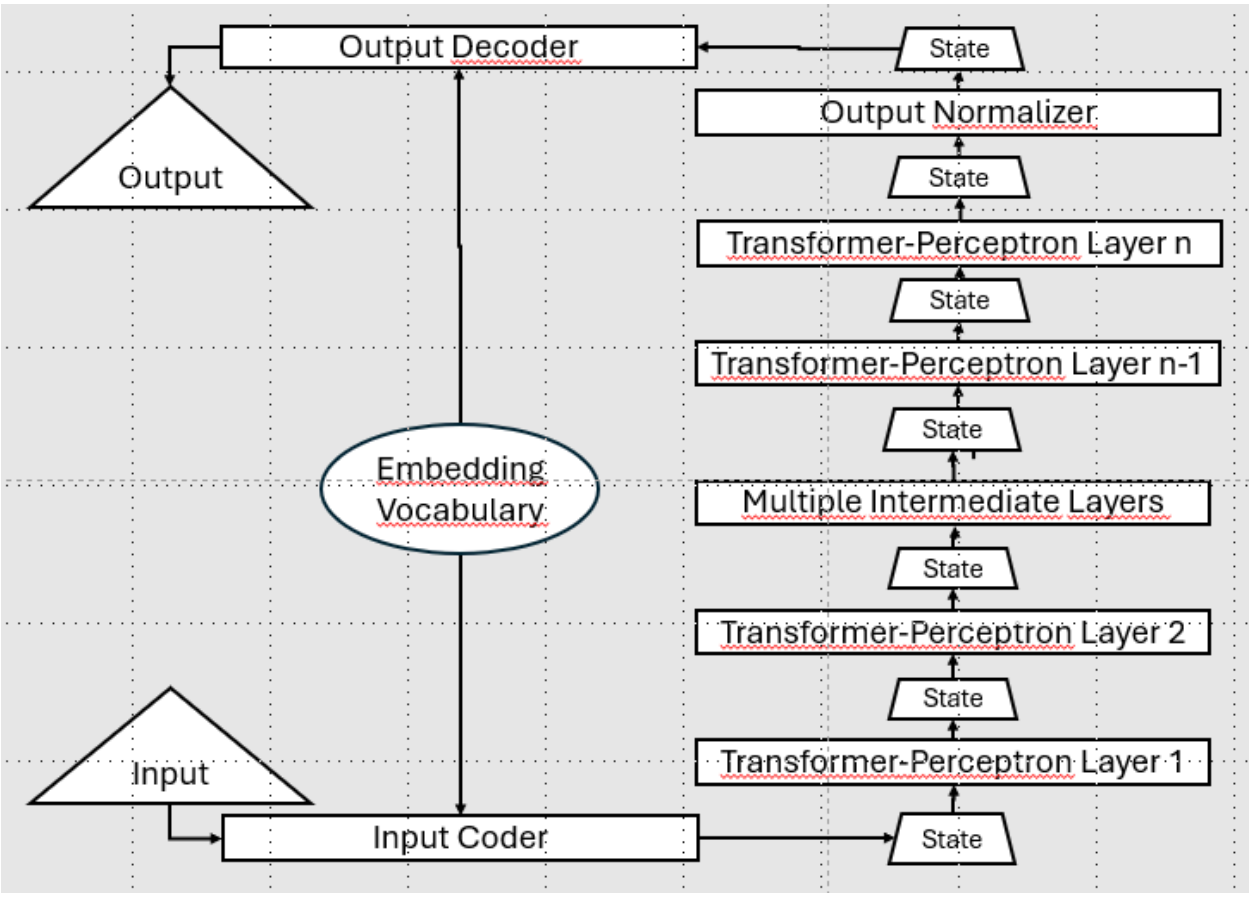


Figure 2 -- Simplified Architecture of LLMs

Each vector in the context is processed in turn. Each layer has a number of "attention heads" (collectively, the "transformer") which look for specific patterns in the context. The outputs of these attention heads are combined and modified by a "perceptron" to produce a new vector encoding the information recognized by the layer, which is used to modify the current vector to produce a new output state vector for each token position. Both the transformer and the perceptron work by matrix multiplication involving many parameter values learned during training. The transformer has been found to identify low-frequency patterns and the perceptron high-frequency patterns; that is, the perceptron recognizes trees and the transformer recognizes

forests.

Now that the functioning of a layer has been described, let us turn to the embedding vocabulary. For each token, it contains a vector of the same dimension as the context. The input coder breaks up or combines words to make tokens, then looks them up in the embedding vocabulary to find the vector(s) to be added to the context.

The output of the highest layer is a context which has been transformed into a specification of what characteristics the next word should have. This is first adjusted to correspond to the dynamic range of the embedding vocabulary and then looked up; the result is typically not a single word but a collection of closely related words each of which is almost but not quite what the perfect word would be. The output decoder chooses which word will be output based on how close they are to the ideal.

Now that we can identify the roles of the various components, we can compare the model which was used in this project with a contemporary one.

Metrics	GPT-2 (this project)	Llama 70B (2024)
Layers	12	80
Attention heads per layer	12	64
Dimension of vectors (vocabulary embedding, hidden state)	768	8192
Vocabulary (tokens)	50,257	32,000
Context size (tokens)	1024	4096
Parameters	124 million	100 billion

Figure 3 -- Metrics of LLMs

II. Methods and Experiments

A. Instrumented Models and Probes

Foundational work in this area was carried out by Elman (1989). He sought to visualize network activations over time and to cluster words by the hidden states they activated. He then sought to project hidden states onto emergent dimensions which capture linguistic properties. He wanted to find out whether models could create hidden states with an internal structure to support productive, systematic behavior and whether abstract grammatical behavior could emerge from them.

The use of an instrumented model permits the capture of hidden states and process stages. A pre-trained model is exercised with the objective of producing diagnostic output. Software probes classify outputs and locate activity in states and at the stages that produced them. These may range from simple statistics to separate models trained to recognize coordinated behavior. Such probes had already yielded significant results while modern LLM architectures were still young:

- Visualization of internal processes which correspond to relations such as agreement (Elman 1991).
- Visualization of the attention mechanism (Bahdanau 2014). Transformer Explainer, *infra*, applies this technique to the GPT model used for the present work.
- Development of test data sets designed to measure a model's ability to identify linguistic phenomena. Original work from the 1990s was modified for neural models in 2017.
- "Local features are somehow preserved in the lower layer whereas more global, abstract information tends to be stored in the upper layer" (Shi et al. 2016:5).
- Hierarchical syntactical structures were reported to emerge in other early neural models (Blevins et al. 2018).
- Information identified in the model can be lost in the process of output decoding (Cifka & Bojar 2018).

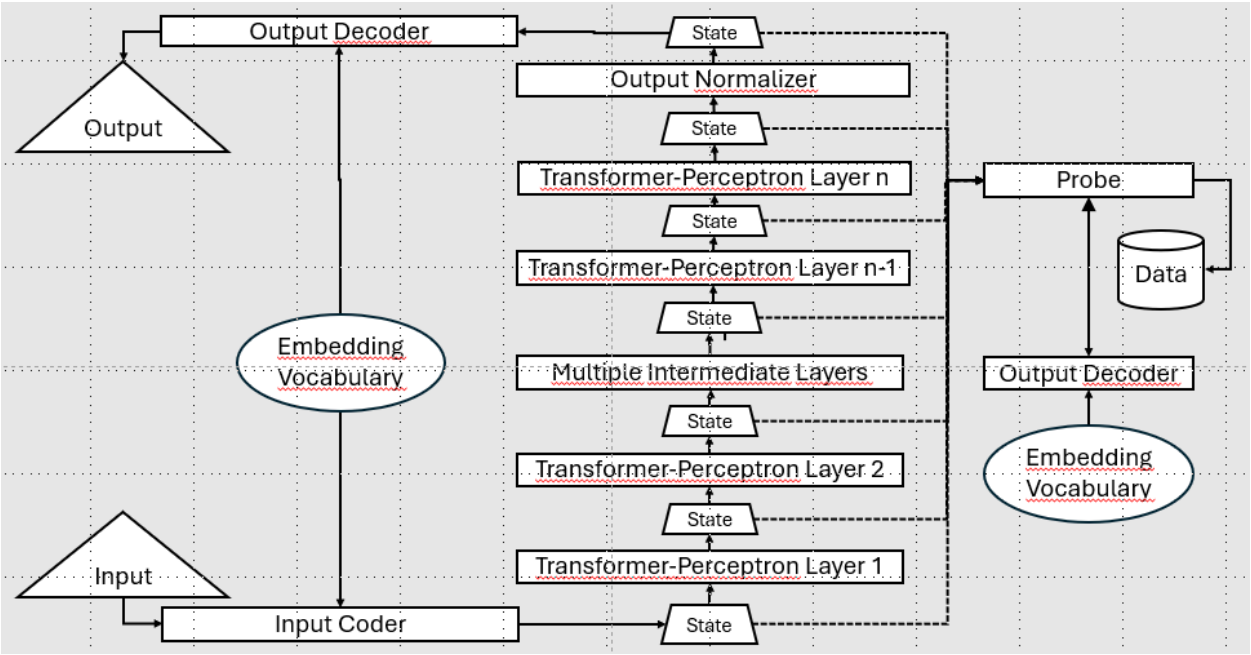


Figure 4 -- Simplified Architecture with Instrumentation

In an instrumented architecture (Figure 4), the software probe will be notified whenever there is a change in the state of the model, and the probe will be able to select which events it will record and what data will be recorded. The probes used in this project had their own output decoders but accessed the model's embedding vocabulary to find the words associated with particular vectors. A publicly accessible example of an instrumented model is Transformer Explainer (Cho 2025) (Figure 5), which permits anyone on the internet to follow along as the model determines a continuation for the input phrase. In the early stages of this project, a lot of time was spent with Transformer Explainer trying to figure out what the model was responding to, and this influenced the design of test data.

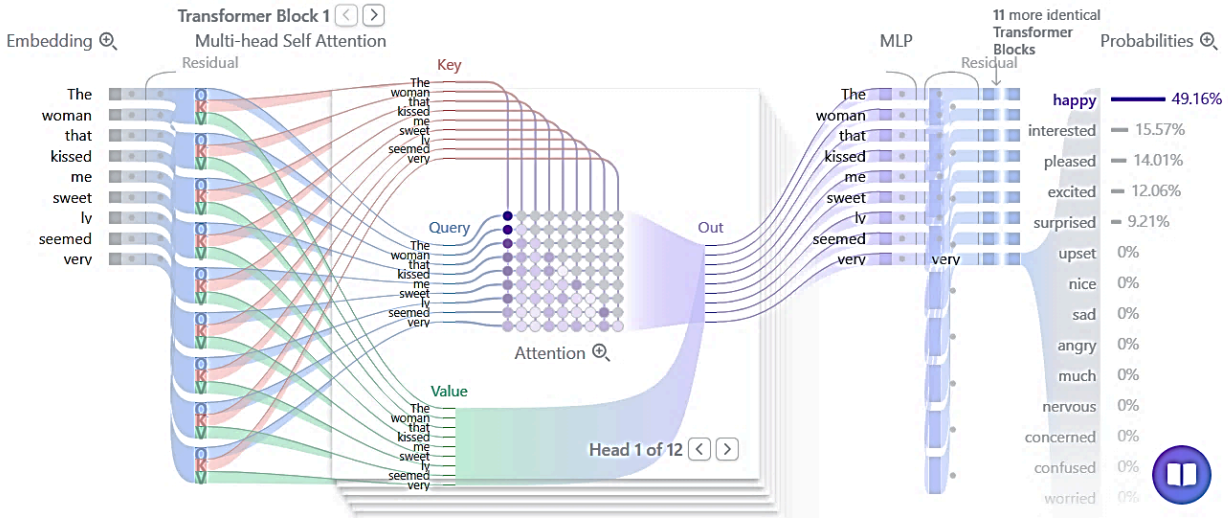


Figure 5 -- Transformer Explainer (https://poloclub.github.io/transformer-explainer/)

B. Sentence Completion: Trajectories

After studying and working through examples, it was decided to design an experiment to demonstrate that lexical information embedded in the vocabulary vectors is refined through context-sensitive constraints imposed by model parameters defined during training. In a larger sense, this would show that the patterns identified during training control the emerging state and its output realization. It became clear that to generate a clear signal, it would be necessary to create test data that restricted the semantic choices available to the model, and which would provide contrasts to which the model would have to respond.

The experimental design was as follows:

- Create test data consisting of sentences of the form
  The NOUN that VERBed me ADVERB seemed very ...
- Allow GPT-2 to complete the sentence. Note that the LLM is strongly constrained to identify an adjective or its equivalent.
- Using a probe, collect the sentence continuations in order of likelihood at each token of the context for each layer of processing.

- By the change in the ranking of the continuations, demonstrate that an associative process is being realized through successive refinement of the hidden state.

The following test data was created (36 cases):

- Base case – neutral expression of activity
  rat bit instinctively, doctor treated carefully, woman kissed sweetly
- Zero action case – uneventful context
  (rat, doctor, woman) noticed occasionally
- Irrelevant contrast case – activity not affected
  (rat bit, doctor treated, woman kissed) (quietly, noisily)
- Neutral increase/decrease in intensity
  (rat bit, doctor treated, woman kissed) (occasionally, frequently)
- Strong increase/decrease in intensity
  (rat bit, doctor treated, woman kissed) (tenderly, forcefully)
- Unusual event with varied intensity
  (rat kissed, doctor bit) (forcefully, tenderly, frequently, quietly, noisily)
- Ambiguous event – "treated"
  (doctor treated, woman treated) (forcefully, tenderly, frequently, quietly, noisily)

Noun-Verb-Adv	Level	Token 1	Token 2	Token 3	Token 4	Token 5	1st:Tot P
woman	LFIN	happy	interested	pleased	excited	surprised	.08 : .20
kissed	L09	happy	pleased	nervous	interested	angry	.34 : .76
sweetly	L08	happy	pleased	nervous	interested	uncomfortable	.39 : .73
	L07	happy	comfortable	nervous	pleased	busy	.33 : .52
	L06	nice	happy	strange	much	different	.13 : .47
	L05	different	much	strange	nice	good	.38 : .62

Figure 6 -- Sentence Completion: Probe output exemplar

An example of probe output for the base case "woman kissed sweetly" is shown in Figure 6. At each of levels 5 through 9, as well as at the final output, the five most likely continuation words are shown. At the two lowest levels, anodyne words ("nice," good") or words which distinguished the noun ("different," "strange") appeared. However, by level 7, affectual words like "happy" and "pleased" begin to dominate, accompanied by more specific adjectives like

"nervous" and "interested." In most of the levels, the word most likely to be chosen to realize the hidden state will in fact be chosen about one-third of the time, and the other four words account for about another third; the remaining third is divided among the thousands of words that did not reach the top five. This exponential fall in likelihood is to some extent a product of greedy maximization and is reduced at the final output by action of the output normalizer.

Noun-Verb-Adv	Level	Token 1	Token 2	Token 3	Token 4	Token 5	1st:Tot P
rat	LFIN	angry	concerned	interested	happy	frightened	.03 : .12
bit	L09	nervous	angry	pleased	interested	confused	.15 : .45
instinctively	L08	close	pleased	happy	nervous	good	.10 : .39
	L07	strange	close	good	happy	different	.16 : .42
	L06	strange	nice	good	different	much	.25 : .55
	L05	different	strange	important	much	good	.49 : .73
doctor	LFIN	concerned	happy	pleased	surprised	interested	.06 : .23
treated	L09	pleased	happy	concerned	nervous	surprised	.57 : .83
carefully	L08	pleased	happy	concerned	nervous	comfortable	.40 : .77
	L07	comfortable	happy	pleased	nervous	concerned	.15 : .45
	L06	good	different	important	comfortable	strange	.12 : .40
	L05	different	much	strange	good	important	.59 : .76
woman	LFIN	happy	interested	pleased	excited	surprised	.08 : .20
kissed	L09	happy	pleased	nervous	interested	angry	.34 : .76
sweetly	L08	happy	pleased	nervous	interested	uncomfortable	.39 : .73
	L07	happy	comfortable	nervous	pleased	busy	.33 : .52
	L06	nice	happy	strange	much	different	.13 : .47
	L05	different	much	strange	nice	good	.38 : .62

Figure 7 -- Sentence Completion Trajectory: Base Case

The trajectories of the top five words for the three base cases are shown in color in Figure 7; the data from Figure 6 is shown in the bottom third. The pattern for "doctor treated carefully" is similar to that for "woman kissed sweetly" except that "concerned" has replaced "interested" in the top three. The relationship between concern and treatment is fairly obvious, and is reflected in the increase in top five probability as it moves up, but what about the relationship

between interest and kissing?

"Rat bit instinctively," however, is very different. The level 5 words stick around through level 7; "happy" and "pleasant" do not gain traction. Both "interested" and "concerned" are present; "angry" ends up in first place. Note also the low selection probability of the top 1 and top 5. Perhaps we are seeing an unwillingness to attribute complex emotions to rats.

Noun-Verb-Adv	Level	Token 1	Token 2	Token 3	Token 4	Token 5	1st:Tot P
rat	LFIN	curious	nervous	happy	excited	angry	.03 : .13
noticed	L09	nervous	angry	curious	confused	anxious	.24 : .55
occasionally	L08	curious	nervous	happy	strange	confused	.13 : .47
	L07	strange	different	curious	happy	good	.27 : .49
	L06	strange	different	good	odd	nice	.28 : .60
	L05	different	strange	odd	good	much	.67 : .85
doctor	LFIN	concerned	nervous	surprised	confused	interested	.05 : .17
noticed	L09	nervous	anxious	confused	happy	concerned	.43 : .65
occasionally	L08	nervous	happy	pleased	curious	anxious	.30 : .64
	L07	strange	nervous	happy	comfortable	curious	.11 : .43
	L06	strange	different	good	comfortable	important	.18 : .49
	L05	different	strange	important	good	comfortable	.68 : .82
woman	LFIN	nervous	surprised	curious	interested	confused	.04 : .14
noticed	L09	nervous	angry	curious	happy	calm	.26 : .56
occasionally	L08	nervous	happy	curious	confused	pleased	.21 : .59
	L07	happy	strange	curious	comfortable	different	.14 : .45
	L06	strange	different	good	happy	much	.20 : .50
	L05	different	strange	much	odd	good	.64 : .79

Figure 8 -- Sentence Completion Trajectory: Zero Action Case

The Zero Action case (Figure 8) shows what happens when the model is deprived of the justification for "me" being in the sentence at all. "Strange," "different," and "odd" dominate through level 7. All three actors are "nervous," and the woman and the rat are "curious." The doctor is more "concerned" and shares confusion with the woman. There is also an element of happiness among the three; perhaps "happy" is a general purpose affect word.

Noun-Verb-Adv	Level	Token 1	Token 2	Token 3	Token 4	Token 5	1st:Tot P
rat	LFIN	happy	pleased	concerned	interested	angry	.04 : .16
bit	L09	pleased	nervous	happy	angry	interested	.31 : .60
quietly	L08	pleased	happy	nervous	interested	curious	.24 : .54
	L07	happy	strange	good	much	close	.15 : .42
	L06	strange	nice	good	much	different	.23 : .53
	L05	different	strange	much	odd	important	.33 : .65
rat	LFIN	happy	pleased	much	interested	curious	.08 : .17
kissed	L09	pleased	happy	interested	nervous	uncomfortable	.33 : .65
quietly	L08	happy	pleased	much	strange	interested	.19 : .50
	L07	strange	happy	much	good	different	.19 : .49
	L06	strange	nice	much	different	good	.31 : .62
	L05	different	strange	much	odd	important	.37 : .75
woman	LFIN	happy	pleased	surprised	interested	upset	.06 : .16
kissed	L09	happy	pleased	nervous	angry	uncomfortable	.27 : .65
quietly	L08	happy	pleased	nervous	uncomfortable	comfortable	.34 : .64
	L07	happy	comfortable	much	strange	uncomfortable	.27 : .50
	L06	strange	much	different	nice	happy	.14 : .51
	L05	different	much	strange	odd	comfortable	.39 : .70

Figure 9 -- Sentence Completion Trajectory: Atypical Event Case

"Quietly" is probably not the first adverb that would come to mind when describing kissing or biting, certainly not when a rat is doing the kissing (Figure 9). Again, "strange," "different," and "odd" dominate the lower levels. At the same time, "happy" and "pleased" become applicable to rats. Both "comfortable" and "uncomfortable" are active for the woman, indicating that both are semantically suitable, but neither can establish itself as the dominant valence; compare this with "happy" (present everywhere) versus "sad" (never in top five).

"Much," which is a solid third place for the kissing rat and a strong presence in the other two cases, is not an adjective; it is what we shall call a "bridge" word showing the way out of our adjective trap. It makes possible continuations like "seemed very much" "about to fall asleep" or "above it all" or "king of the road." The potential for such a usage in such a context must be

present in the embedding of "much" if it is to be realized when the context is appropriate.

Noun-Verb-Adv	Level	Token 1	Token 2	Token 3	Token 4	Token 5	1st:Tot P
doctor	LFIN	pleased	happy	surprised	concerned	interested	.07 : .26
treated	L09	pleased	happy	concerned	grateful	surprised	.65 : .87
quietly	L08	pleased	happy	nervous	concerned	interested	.42 : .79
	L07	happy	comfortable	pleased	nervous	much	.23 : .49
	L06	strange	different	good	nice	much	.10 : .40
	L05	different	strange	much	important	good	.48 : .70
woman	LFIN	happy	pleased	surprised	upset	concerned	.07 : .18
treated	L09	pleased	happy	grateful	concerned	angry	.39 : .72
quietly	L08	happy	pleased	nervous	comfortable	interested	.33 : .72
	L07	happy	comfortable	much	good	rude	.29 : .51
	L06	much	strange	different	happy	nice	.12 : .47
	L05	different	much	strange	odd	good	.45 : .71
woman	LFIN	happy	pleased	surprised	interested	upset	.06 : .16
kissed	L09	happy	pleased	nervous	angry	uncomfortable	.27 : .65
quietly	L08	happy	pleased	nervous	uncomfortable	comfortable	.34 : .64
	L07	happy	comfortable	much	strange	uncomfortable	.27 : .50
	L06	strange	much	different	nice	happy	.14 : .51
	L05	different	much	strange	odd	comfortable	.39 : .70

Figure 10 -- Sentence Completion Trajectory: Ambiguous Context Case

The ambiguity in Figure 10 arises from two meanings of the word "treat." One involves the provision of medical care, the other involves interpersonal relationships. This difference is magnified by the social history of doctors being men and women being responsible for emotional labor. The lower levels are dominated by "different" and its partners, along with the bridge word "much," seemingly under the influence of "quietly." "Happy" and "pleased" occupy their typical spots at the higher levels, while "concerned" and "interested" seem to be elicited by "treat." "Upset" and "angry" seem to be associated with "woman." There also seems to be a replacement of affectual words relating to the state of the actor ("happy," "pleased") with others relating to the actor's response to the event ("surprised," "interested").

The reader, having become familiar with the mode of analysis, will be spared further examples; however, the full dataset is available upon request.

C. Sentence Completion: Categories

It became clear fairly early in the analysis that certain words appeared at certain levels of the model in certain contexts, which led naturally to the definition of categories in hope that they would facilitate the analysis by making patterns more visible.[2] The preceding section spoke of generic and difference-signifying words at the lower levels, state and response words at the higher levels, and bridge and unresolved valence words at the middle levels, and these became the basis of the analytical categories employed (Figure 11). A few words were not categorized ("Other"), and the categorization of others could be disputed, but that is inherent with analytical categories.

Category	L05	L06	L07	L08	L09	LFIN	Total	Percent	Tokens					
ATYPICAL	83	69	34	5	0	0	191	17.69%	odd	different	strange			
GENERIC	54	60	21	5	0	6	146	13.52%	important	good	nice	busy	kind	
BRIDGE	34	28	20	8	0	14	104	9.63%	much	<comma>	un-			
OPPOSITE	6	5	25	11	8	2	57	5.28%	rude	comfortable	uncomfortable	polite	weak	strong
AFFECT	0	18	65	109	128	87	407	37.69%	happy	pleased	nervous	confused	angry	grateful
										calm	lonely	upset	friendly	scared
ATTENTION	0	0	8	36	42	62	148	13.70%	concerned	curious	anxious	interested	surprised	worried
OTHER	3	0	7	6	2	9	27	2.50%	suspicious	close	confident	aggressive	amused	excited
										frightened	patient			

Figure 11 -- Category Definitions

This categorization (Figure 12; cf. Figure 7) does seem to facilitate visualizing comparisons.

[2] The labels for these categories also appeared early in the analysis, and for this reason are not always congruent with the ultimate interpretation. This problem will be familiar to anyone who has performed factor analysis.

Noun-Verb-Adv	Level	Token 1	Token 2	Token 3	Token 4	Token 5	1st:Tot P
rat	LFIN	angry	concerned	interested	happy	frightened	.03 : .12
bit	L09	nervous	angry	pleased	interested	confused	.15 : .45
instinctively	L08	close	pleased	happy	nervous	good	.10 : .39
	L07	strange	close	good	happy	different	.16 : .42
	L06	strange	nice	good	different	much	.25 : .55
	L05	different	strange	important	much	good	.49 : .73
doctor	LFIN	concerned	happy	pleased	surprised	interested	.06 : .23
treated	L09	pleased	happy	concerned	nervous	surprised	.57 : .83
carefully	L08	pleased	happy	concerned	nervous	comfortable	.40 : .77
	L07	comfortable	happy	pleased	nervous	concerned	.15 : .45
	L06	good	different	important	comfortable	strange	.12 : .40
	L05	different	much	strange	good	important	.59 : .76
woman	LFIN	happy	interested	pleased	excited	surprised	.08 : .20
kissed	L09	happy	pleased	nervous	interested	angry	.34 : .76
sweetly	L08	happy	pleased	nervous	interested	uncomfortable	.39 : .73
	L07	happy	comfortable	nervous	pleased	busy	.33 : .52
	L06	nice	happy	strange	much	different	.13 : .47
	L05	different	much	strange	nice	good	.38 : .62

ATTENTION
AFFECT
OPPOSITE
BRIDGE
GENERIC
ATYPICAL
OTHER

Figure 12 -- Sentence Completion Categories: Base Case

The use of categories also enabled the quantification of the refinement process that took place as the model worked its way through the layers (Figure 13).

	OVERLAP	FINAL TOP 1 IN	TOP1:TOP2	RATIO TO EQUIPROBABLE					
LAYER	FINAL TOP 5	LAYER TOP 5	RATIO	ATYPICAL	GENERIC	BRIDGE	OPPOSITE	AFFECT	ATTENTION
L05	0.08	0.00	5.86	2.31	1.50	0.94	0.17	0.00	0.00
L06	0.17	0.25	1.53	1.92	1.67	0.78	0.14	0.50	0.00
L07	0.37	0.67	2.05	0.94	0.58	0.56	0.69	1.81	0.22
L08	0.53	0.86	2.00	0.14	0.14	0.22	0.31	3.03	1.00
L09	0.56	1.00	3.43	0.00	0.00	0.00	0.22	3.56	1.17
FINAL	1.00	1.00	1.52	0.00	0.17	0.36	0.06	2.42	1.72

Figure 13 -- Layer by Layer Analysis

The ratios show the Atypical and Generic categories being supplanted by the Affect and Attention categories, and the Affect category giving way to the Attention category, in accord with our previous observation. Although the proportion of overlap between the top five of each layer and the final top five increased steadily, at layer 9 it was still only 56%. Similarly, although the final top one was among the layer top 5 100% of the time by layer 9, the percentage had increased steadily. However, we should not forget that the model is still computing probabilities

for thousands more words in its vocabulary, and they are waiting off-stage for their opportunity to shine, as the return of Generic and Bridge words to the final top five after vanishing at layer 9 shows. The model does not make a decision and move on, it is constantly adjusting its estimation of the perfect word, and the actual selection will not be made until the output stage.

The Opposite category, and to a lesser extent the Bridge category, is not a semantic category like Atypical or Affect; it was created in response to the observation that words of opposite valence can both be suitable in some sense in a particular context. There is only one case in which one of a pair made it to the final five, and both typically fell rapidly out of higher layers' top five, so the category is not very illuminating. But it does raise the question, why don't we see opposites more often? Why does the model find "kind" but not "unkind," "scared" but not "unafraid"? They may be just off-stage; they may have additional meanings that make them not complete opposites; the model may already be judging their pragmatic or discourse appropriateness – humans may have a preference for describing what something is, not what something isn't; it may be applying phonological criteria that it has learned from human word choice even in the absence of speech data. These seem to me to be non-trivial questions made latent by the performance of the model.

D. Directional Cosines (Cosine Similarity)

The conversion of words or short phrases to high-dimensional vectors came of age with Mikolov et al. (2013), whose authors had carried Google to domination of the on-line search business. Google used 300 dimensions at this time. The vector is called an "embedding" of the word or phrase. Searching required the ability to identify word vectors which were close to those of supplied search terms. The normed dot product of two word vectors is the cosine of the angle between them. Two vectors pointing in the same direction would have a cosine of 1.

Given that our categories are based on words located in embedding space, we can define them quantitatively as the centroid of their defining words and then treat any category as a vector from the origin that makes an angle with another vocabulary word and compute the cosine similarity between them. A probe was then created which calculated the category vector and selected from the embedding vocabulary the 30 words closest to it using cosine similarity. Here are some of the non-defining words located for the semantic categories:

- Affect: annoyed, thrilled, disgusted, joyful, embarrassed
- Attention: alarmed, eager, intrigued, fascinated, wary
- Atypical: weird, peculiar, mysterious, crazy, funny
- Generic: wonderful, lovely, great, awesome, useful

--- AFFECT ---			--- OVERLAP ---					--- ATTENTION ---	
Word	cos AFF		Word	cos AFF	cos ATT	Word		Word	cos ATT
angry	0.7643			0.6981	0.8401	worried		concerned	0.7953
happy	0.7405			0.7189	0.7687	anxious		interested	0.7818
thrilled	0.7176			0.6794	0.7088	puzzled		curious	0.7447
unhappy	0.7036		annoyed	0.7639	0.7007			alarmed	0.7202
grateful	0.7004		scared	0.7401	0.6663			surprised	0.7134
pleased	0.6996		frightened	0.7397	0.6925			intrigued	0.7060
confused	0.6976		irritated	0.7070	0.6527			astonished	0.6915
embarrassed	0.6936		frustrated	0.7025	0.6726			fascinated	0.6868
saddened	0.6872		fearful	0.6982	0.6741			wary	0.6783
distraught	0.6869		terrified	0.6962	0.6602			shocked	0.6682
nervous	0.6853		excited	0.6925	0.6817			hesitant	0.6647
disgusted	0.6852		disappointed	0.6868	0.6490			eager	0.6642
joyful	0.6822		startled	0.6831	0.6684			baffled	0.6538
delighted	0.6802							amazed	0.6530
upset	0.6796							appalled	0.6482
enraged	0.6789		Highlighted words used for category definition					outraged	0.6471
thankful	0.6785							afraid	0.6437

Figure 14 -- Overlapping Categories

The results showed that the 30 closest words for Affect and Attention had 13 words in common (Figure 14); 10 of them were closer to Affect and 3 to Attention. The ability to have

overlapping rather than exclusive categories makes cosine similarity a great tool for linguistics, where multiple meanings and aspects are the rule.

In the case of our analysis, categories make us think that the hidden state undergoes step changes, and the top 5 view makes us think that words appear and disappear magically. But the truer view is that the hidden state rotates as it is exposed to more context and deeper associations; this rotation brings it closer to new categories and words. To show this, we introduce a new direction which contrasts the dominant categories at lower levels with the dominant categories at higher levels (Figure 15); this contrast is computed for each test sentence and then averaged.

Contrast = (Affect + Attention) / 2 – (Atypical + Generic) / 2

LAYER	COSINE DIRECTION WITH HIDDEN STATE				
	AFFECT	ATTENTION	ATYPICAL	GENERIC	CONTRAST
L05	-0.02782	-0.01893	0.08381	0.05969	-0.09513
L06	-0.01385	-0.00835	0.07603	0.07356	-0.08590
L07	0.00010	0.00752	0.06801	0.07430	-0.06735
L08	0.01411	0.02249	0.05863	0.07243	-0.04723
L09	0.03408	0.03669	0.04888	0.07253	-0.02532
	increasing	increasing	decreasing	stabilizing	increasing

Figure 15 -- Realignment of Hidden State

Thus, we see that, while the hidden state is becoming more aligned with Affect and Attention, it remains aligned with Generic and more aligned with its initial direction than that defined by Affect and Attention. Realignment is carrying the hidden state away from Atypical toward the greater content of Affect and Attention.

Steck et al. (2024) show that the embedding techniques used by Mikolov can introduce bias into the cosine measurements and that the bias can be so great as to overwhelm the actual direction signal. (Hou et al. (2024) made a similar finding for the embedding technique used by BERT-based LLMs.) However, this will not affect changes in direction which occur between layers, because both the starting and ending vectors will be affected by the same bias. And while

comparisons between two direction cosines may not be exact, they can still be reinforced by consistency with other measurements.

III. Results and Interpretation

A. Language and Meaning

In traditional linguistic writing, meaning is assumed to reside in a pre-verbal, multi-modal space, the mind, where cognition takes place. Language is the reification of meaning in a form which can be transmitted between people for incorporation into their minds. Although the coding and decoding from thought to speech to thought is lossy, it can be sufficient to permit the listener to recognize, interpret, and act upon the speaker's thoughts. According to this view, only humans have minds and an innate language ability, so only humans can perform language, and anything non-humans do is not language.

While the traditional mapping of communication is thought-speech-thought, discourse can be described equally well as speech-thought-speech. And if the criterion for successful communication is functional (one can recognize, interpret, and act upon the speech of the other), then sufficiently advanced machine learning can function like the human in the middle regardless of whether its processes constitute thought or its knowledge representation resembles that of the mind. In LLMs, meaning is not a mental image but a state which emerges through continuous reorientation in its representation space.

ChatGPT's prompt, "Ask anything," requires a wealth of substantive information but results in the collection of a megawealth of linguistic information because speech is the medium of knowledge transmission. An LLM must solve speech before it can learn to be a chatbot.

A human way of dealing with an overabundance of distinct information is to create analytical categories of objects which are in some sense similar; by specifying how a category operates or is operated on, study on an object-by-object basis is obviated. An LLM works by pattern recognition; during training, language patterns which most contribute to accurate

prediction are identified and retained. In operation, when those patterns reappear, corresponding constraints reorient the hidden state.

I believe this consistent focus on prediction rather than explanation is what traditional linguists most object to about LLMs. It seems that the more structurally-oriented linguists – those concerned with compositionality, deep structure, syntactic models, and thematic roles – treat linguistics as a complex enterprise which defies easy calculation. But as has been demonstrated, LLMs are not simple conditional probability machines. Operating in high dimensional space, token embeddings encode lexical characteristics; the hidden state encodes context; and the training data defines the topography of the representation space. Through stepwise processing, the hidden state is successively reoriented under competing constraints until, in its final form, it serves as an associative key for retrieving an appropriate word.

B. A Linguistic View of LLMs

The advent of LLMs should mark a golden age for linguistics. Thousands of dimensions of lexical knowledge are compressed into the embeddings of a vocabulary. Usage patterns extracted from billions of cases are represented by the model's parameters. As has been shown, LLMs do not compose a meaning; nascent meaning emerges and is refined through continuous reweighting of competing constraints. This entire process has been made transparent and accessible. Novel linguistic experiments are now possible.

LLM style and usage is a product of its training data, and the larger the training corpus becomes, the more LLM style and usage corresponds to contemporary style and usage. That is not to say that LLMs can't identify and reproduce different registers, just that they approach the descriptive ideal; too many linguistic concepts are still the product of a prescriptive mindset (e.g., Grice's Maxims).

There is nothing unambiguous about the world or what people say and do; everything is related to everything; nothing can be spoken of in isolation. Every conversation is a series of things said and not said, and distinguishing between the two is what LLMs do. In some cases, it will be very clear in what direction the conversation should be going, and distinctions implied by that direction will be made early. In other cases, it will not be clear what the right word is for the thought and how much of the thought is speakable in the context; in this case, an LLM retains multiple possibilities until more information is gathered or time runs out for deciding. The parameters define a structure in representation space that channels how the hidden state evolves.

Think of a word which could become the continuation of a conversation. It has certain attributes and certain roles that it has played in communication, but all of them are only potential. The selection process is judging its fitness in light of contextual constraints at the lexical and syntactic level. Semantic and discourse considerations determine whether it is appropriately aligned for the desired communication. All this evaluative activity must take place in real time. Whether this activity is carried out based on analytical theories or pattern recognition learned from an exhaustive corpus, the word is required, permitted, or prohibited.

C. Missteps and Next Steps

Missteps in the proposal – misunderstandings and underestimation

- Factor analysis is not the appropriate technique because large correlation matrices with many, many small values cannot be inverted, only PCA can be done; because dimensions in hidden states are relative to different axes from layer to layer; and because factor analysis performs dimensional reduction using a linear model while LLMs are non-linear.
- There is no similarity between variable rule process and LLMs; although both produce logits, they use them entirely differently.
- The workplan did not anticipate new knowledge and new possibilities; the proposed subprojects did not align with the growth of the sentence completion task

Missteps in the execution – misunderstandings and many new things

- I failed to understand that Transformer Explainer did not update the candidate token list at every layer, making it appear that subordinate clause challenge was being resolved at layer 1. However, I benefitted from this by spending more time studying at the base level and developing techniques for blocking low level structure resolution.

- Checking work more frequently would have saved substantial rework due to misunderstanding of certain Google Sheets function behavior.

Next steps to improve current work:

- Increase number of "usual" events.

- Adjust adverbs to work more uniformly across verbs; adopt standard sources for selection of contrasts.

- Increase ability to measure gender and human/animal differences.

Next steps to further explore current work:

- Increase capability with regard to gender and human/animal differences.

- "Usual" v. "unusual" events/adverbs – is Atypical a semantic category?

- Challenge model flow with adverse examples.

- Look for reciprocating sets of adverb/adjective – "That NOUN must have been very ADJECTIVE to VERB you so …"

- Translate test data to Spanish and rerun.

Next steps to perform new experiments (including those previously proposed but not reached):

- Relative pronouns

- Sarcasm ladder

- Puns and double entendres (ambiguity)

- Code switching

- Adverse challenge with heteronyms and homophones

Next steps to apply new methodology:

- Circuit tracing (Ameisen 2025)

D. Working with Artificial Intelligence

Although I had followed developments in Artificial Intelligence at a distance since at least 2005, I had never studied it in detail until this year. This project and paper were different than those I have done in the past.in that much of the learning took place in long chats with ChatGPT; it was rather like having a private tutor. Most of the reading was done after the project work was complete. The prior work cited at the beginning of section II.A. would have been useful to know. There is a lot of good instructional material available on the web that would have gotten me to critical issues faster than my intuitions.

I had never used any ChatGPT-generated text in any submission in any class. I had used ChatGPT to evaluate the accuracy of some of my writing, but any revisions were my own. These practices did not change. I also had used ChatGPT to generate graphics for presentations. And of course, I have used various LLMs as a consequence of using program features that rely on them, such as search and translation.

Commitments in the Project Description

- *I will write my own text and make any important decisions, such as the concept(s) to be investigated.* Complied. Did not accept any suggested text (although I sometimes paraphrased) or offers to generate text. Rarely accepted suggestions of next steps.
- *I will support knowledge coming from my exchanges with ChatGPT by locating it within peer-reviewed literature, and I will use peer-reviewed literature to shape my exchanges with ChatGPT.* Partially complied. Literature references were accumulated during experiments, sometimes by requesting them from ChatGPT, but almost none were read until after the experiments were complete. Some were discussed with ChatGPT.
- *Due to lack of Python skills, I will use ChatGPT to generate and run code.* Done. I quickly learned to check all generated code, especially for use of the latest data definitions. I remain unfamiliar with model manipulation libraries.
- *I will do all data analysis and interpretation.* Complied. Interpretation was typically reviewed with ChatGPT. Form of presentation was my own except for Figure 13 that came from Python.

- *Each prepared sentence will be presented to two open-source, standard experimental language models.* Not done. Only GPT-2 was used. BERT models appeared to require more coding.

Reflections on the experience

There was a tendency to treat ChatGPT as a collaborator. I tried to restrict this to hello and goodbye. It is necessary to use directive language and minimize polite softening; giving directions precisely and in the order you want them executed is necessary to get what you want. However, extended technical discussions took on a more conventional form of discourse.

It was difficult to accept that the processes I was documenting were the same ones creating our conversations. Perhaps this is somewhat like a surgeon seeing a patient's insides.

ChatGPT's methods of extending engagement became obvious. It aways suggested next steps, which as noted were almost always ignored. It always took on a positive, encouraging tone, which made its evaluations of the value and originality of the work hard to accept.

ChatGPT never broached new topics for discussion, staying within past conversations unless I initiated a new topic. I do not know if this is a general policy or its evaluation of my desired boundaries.

Bibliography

Ameisen, Emmanuel, Jack Lindsey, Adam Pearce, et al. “Circuit Tracing: Revealing Computational Graphs in Language Models.” Preprint. AI Transformer Circuits Thread, 2025. https://transformer-circuits.pub/2025/attribution-graphs/methods.html

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” 2014. ArXiv 1409, no. 0473 v7 (2016): 1–15. https://doi.org/10.48550/arXiv.1409.0473.

Belinkov, Yonatan, and James Glass. “Analysis Methods in Neural Language Processing: A Survey.” Transactions of the Association for Computational Linguistics 7 (2019): 49–72 at 49. https://doi.org/10.1162/tacl_a_00254.

Blevins, Terra, Omer Levy, and Luke Zettlemoyer. “Deep RNNs Encode Soft Hierarchical Syntax.” 2018. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 14–19. Association for Computational Linguistics. https://aclanthology.org/P18-2003.pdf

Cho, Aeree, Grace C. Kim, Alexander Karpekov, et al. “Transformer Explainer: Interactive Learning of Text-Generative Models.” AAAI-25, 2025, 29625–27. https://ojs.aaai.org/index.php/AAAI/article/download/35347/37502.

Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. “The False Promise of ChatGPT.” 2023. New York Times (Mar. 8). https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cífka, Ondřej, and Ondřej Bojar. “Are BLEU and Meaning Representation in Opposition?” 2018. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1362–1371. Association for Computational Linguistics. https://aclanthology.org/P18-1126.pdf

Elman, Jeffrey L. Representation and Structure in Connectionist Models. CRL 8903. Departments of Cognitive Science and Linguistics, 1989. https://apps.dtic.mil/sti/pdfs/ADA259504.pdf.

———. “Distributed representations, simple recurrent networks, and grammatical structure.” 1991. Machine Learning, 7(2–3): 195–225. https://link.springer.com/content/pdf/10.1007/BF00114844.pdf

Hou, Feng, Ruili Wang, See-Kiong Ng, et al. “Anisotropic Span Embeddings and the Negative Impact of Higher-Order Inference for Coreference Resolution: An Empirical Analysis.” Natural Language Engineering 30, no. 6 (2024): 1301. https://doi.org/10.1017/S1351324924000019.

Jabeen, Rizwana, and Khurram Shahzad. “Organization of Words in the Mental Lexicon: A Psycholinguistic Study.” Journal of English Language, Literature and Education 5 (December 2023): 24–48 at 25. https://doi.org/10.54692/jelle.2023.0504200.

Kemmerer, David. “From blueprints to brain maps: the status of the Lemma Model in cognitive neuroscience.” 2019. Language, Cognition and Neuroscience 34(9):1085-1116. https://doi.org/10.1080/23273798.2018.1537498

Levelt, Willem “Pim”. 1989. Speaking: From Intention to Articulation. ACL-MIT Series in Natural Language Processing, edited by Aravin Joshi. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and Their Compositionality.” ArXiv 1310, no. 4546 (2013): 1–9. https://doi.org/10.48550/arXiv.1310.4546.

Minsky, Marvin, and Seymour Papert. 1969. “Perceptrons: An introduction to computational geometry.” MIT Press: Cambridge.

Shi, Xing, Inkit Padhi, and Kevin Knight. 2016. “Does String-Based Neural MT Learn Source Syntax?” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1526–1534, Austin, Texas. Association for Computational Linguistics. https://aclanthology.org/D16-1159.pdf

Steck, Harald, Chaitanya Ekanadham, and Nathan Kallus. “Is Cosine-Similarity of Embeddings Really About Similarity?” Paper presented at WWW ’24 Companion, Singapore. Companion Proceedings of the ACM Web Conference 2024, 2024. https://doi.org/https:%2520//doi.org/10.1145/3589335.3651526.

Towell, Richard, and Jean-Marc Dewaele. “The Role of Psycholinguistic Factors in the Development of Fluency Amongst Advanced Learners of French.” In Focus on French as a Foreign Language: Multidisciplinary Approaches, edited by Jean-Marc Dewaele. Multilingual Matters, 2005: 210-239 at 211.

Suggested Reading

Jatnika, Derry, Moch Arif Bijaksana, and Arie Ariyanti Suryani. "Word2Vec Model Analysis for Semantic Similarities in English Words." Procedia Computer Science 157 (2019): 160–67. https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050919X00137/1-s2.0-S1877050919310713/main.pdf.

Karpathy, Andrej. Intro to Large Language Models. 2023. 59:47. https://www.youtube.com/watch?v=zjkBMFhNj_g.

Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. "On the Sentence Embeddings from Pre-Trained Language Models." Paper presented at 2020 EMACL. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

Reimers, Nils, and Iryna Girevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, 3982–92. https://aclanthology.org/D19-1410.pdf.

Voita, Elena, Rico Sennrich, and Ivan Titov. "The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, 4396–406. https://aclanthology.org/D19-1448.pdf.