Busting Language Myths

F Oct 31

- **A new language myth: “Generative AI can talk (and think) like humans”**

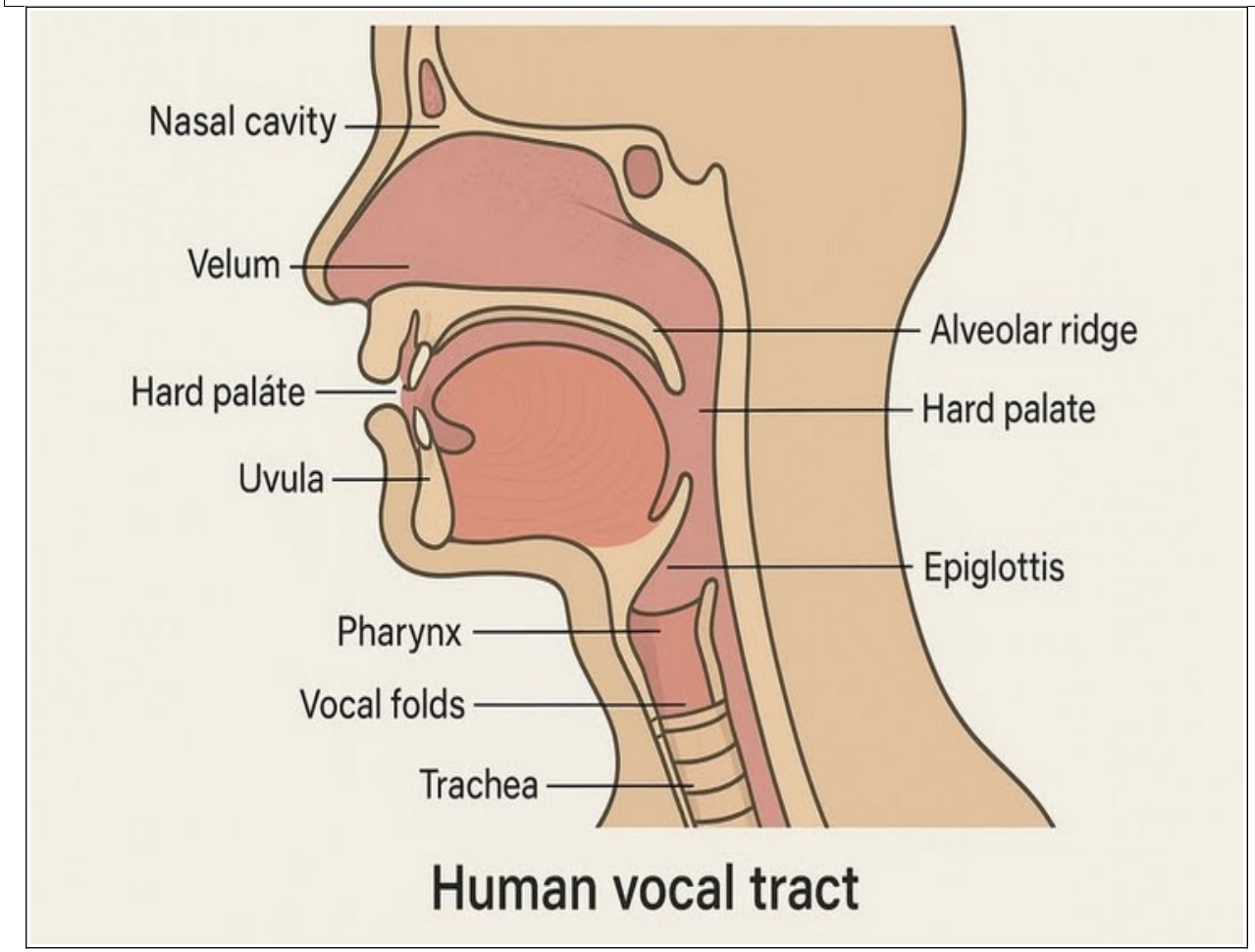
0. Today's objectives

After today's class, you should be able to:

- Explain the use of machine learning in testing linguistic models and theories
- Discuss why LLMs are an extension/expansion of this kind of approach
- Evaluate the claim that "LLMs learn language like humans" on the basis of the article discussed
- Discuss pros and cons of genAI use in general

1. Warm-up: Some more AI hallucinations

- "This is the prompt I used for ChatGPT: "Draw and label a midsagittal diagram of the human vocal tract. The diagram should contain all the parts necessary for a typical introductory course in linguistics."



Thoughts?

prompt and image from Nathan Sanders, U Toronto

1. Warm-up: Some more AI hallucinations

- Some relevant recent links
  - "AI gone wild: Let's discuss times AI got it really wrong" [link]
    Ask A Manager, October 30, 2025
  - "5 Tips When Consulting 'Dr.' ChatGPT"
    New York Times, October 30, 2025 [guest link]

2. A new language myth: genAI and language

- Last time, we talked about various ways in which AI, especially generative AI, can be
  - useful
  - problematic or harmful
- This time, we will consider a new language myth: “GenAI can learn to talk or think like a human”

2. A new language myth: genAI and language

- *Big-picture research question:*
  Can an LLM learn to make the same **grammaticality judgments** about language structures as a human?

  Implications:

  - If yes, maybe **human language acquisition** also needs nothing more than lots and lots and lots of data and an ability to make predictions (acquisition is not sensitive to linguistic structure)

3. Machine learning and linguistic theory

- **Machine learning** in theoretical linguistics:
  - If we feed a computer program some input data,
  - and we program it to learn in a particular way,
  - what kind of patterns does it learn?
- Why is this useful?
  - It doesn't directly prove anything about humans
  - It demonstrates whether or not a particular set of assumptions about available data and learning strategy are enough for a pattern to be learnable

3. Machine learning and linguistic theory

- Example: Morita & O'Donnell (2022) [UNC link]

  Morita, Takashi & Timothy J. O'Donnell (2022). Statistical evidence for learnable lexical subclasses in Japanese. *Linguistic Inquiry* 53(1): 87–120.

  - Abstract: "It has been proposed that the Japanese lexicon can be divided into etymologically defined sublexica on [phonological] and other grounds. [...]
  - "[W]e apply a commonly used clustering method to Japanese words and show that there is robust statistical evidence [...] [and] such sublexica are learnable.
  - The model is able to recover [phonological] properties of sublexica previously discussed in the literature [...]"

4. LLMs as language learners?

- With an LLM, the data provided is **not linguistically structured**
  - As we saw last time, LLMs essentially learn by looking at the probability of a next word, given a prior word and the context
- Do LLMs learn the **same language behavior** as humans?
  - **Could** human mental grammars also be based on probability only?
  - No syntax? No morphology? No phonology?

5. LLMs and human syntax judgments

- Qiu, Duan, & Cai (2025)

  Qiu, Zhuang, Xufeng Duan, and Zhenguang G. Cai (2025). Grammaticality representation in ChatGPT as compared to linguists and laypeople. *Humanities and Social Sciences Communications* 12(1): 1–15.

  - "ChatGPT was asked to provide grammaticality judgments in different formats for over two thousand sentences with diverse structural configurations. We compared ChatGPT's judgments with judgments from laypeople and linguists to map out any parallels or deviations." (Qiu et al. 2025: 2)

5. LLMs and human syntax judgments

- Qiu, Duan, & Cai (2025)
  - "The findings indicate **significant correlations** between ChatGPT and both laypeople and linguists in various grammaticality judgment tasks. This study also reveals **nuanced differences** in response patterns, influenced significantly by the specific task paradigms employed." (Qiu et al. 2025: 13)

5. LLMs and human syntax judgments

- FC [forced-choice] task:
  - ChatGPT was given a pair of minimally different sentences (testing proposals of theoretical syntax)
  - Prompt: Select the **more grammatically acceptable** sentence

5. LLMs and human syntax judgments

- Example sentence pair (Qiu et al. 2025: 13)

  The teacher and principal spoke together after class.

  The teacher spoke together after class.

  - Presentation order (more/less grammatical) was balanced
  - Responses where ChatGPT didn't give a simple response of "Sentence 1" or "Sentence 2" were excluded

- Analysis: ChatGPT's choices compared to those made by humans (linguists) in a previous study

5. LLMs and human syntax judgments

- Data graphic

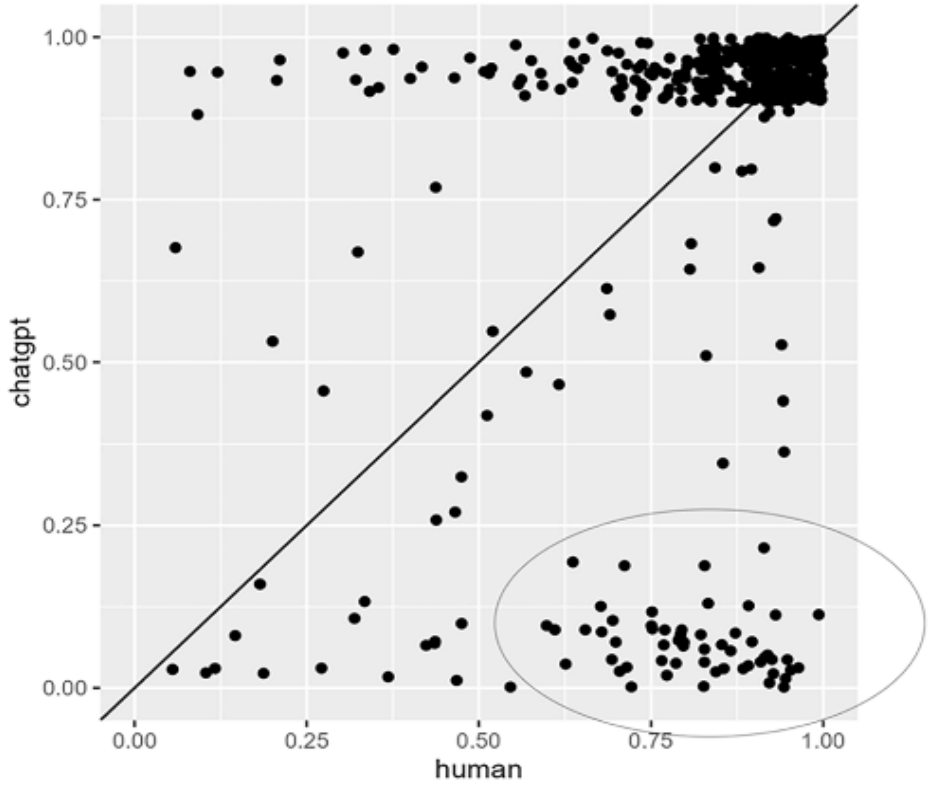


Fig. 8 Probability of selecting grammatical sentences in the FC task for each experimental item, comparing human participants and ChatGPT.
Note. Each point in the figure represents the mean rating score of a sentence, averaged across multiple human participants or across multiple independent trials of ChatGPT judgment. A diagonal reference line is added for ease of comparison. Data points above the line indicate trials where ChatGPT outperformed humans, and those below show where humans had higher accuracy. Points in the circle refer to trials where the accuracy of ChatGPT was much lower than that of humans; see Table 11 for examples.

5. LLMs and human syntax judgments

- "A selection of trials in Experiment 3 for which human participants had a much higher accuracy rate than ChatGPT (90% vs 10%)" (Qiu et al. 2025, Table 11)

More grammatical	**Less grammatical**
Who the hell asked who out?	*Who asked who the hell out?*
A surfer cuter than my husband came along the beach.	*A cuter surfer than my husband came along the beach.*
The teacher and principal spoke together after class.	*The teacher spoke together after class.*
There has been a man considered sick.	*There has been considered a man sick.*
They knew and we saw that Mark would skip work.	*They knew and we saw Mark would skip work.*

5. LLMs and human syntax judgments

- “As an anonymous reviewer correctly pointed out, all three experiments in this project tasked ChatGPT with the judgment of sentence grammaticality in a relative and gradient fashion, which is different from the real-life practice of spotting grammatical anomaly among otherwise well-constructed sentences. If linguists can detect grammatical errors according to their theoretical orientations, it is an intriguing question whether large language models like ChatGPT can detect and correct grammatical errors the same way as linguists do.”
(Qiu et al. 2025: 13)

5. LLMs and human syntax judgments

- Is the myth busted?
  - So far, ChatGPT does **not** seem to behave exactly like humans in judging sentences
    - Is this the best methodology for comparing "mental grammars"?
  - Will LLMs get better?
    - Or are they already "too good" and that's the source of the differences? (memory, speed)
    - Or: Is human language categorically different?

6. Some final thoughts

- Back to the class survey from last time:
  - What are some harms of genAI?
  - Can these be mitigated? Are the trade-offs worth it?

- What kind of AI policy should a class like this one have?