How many stochastic parrots does it take to change a lightbulb?

1st Workshop on Computational Humor (CHum 2025)
31st International Conference on Computational Linguistics (COLING 2025).

Salvatore Attardo

EAST TEXAS A&M

Perplexity

Not bad, right?

How many stochastic parrots does it take to change a lightbulb?

Answer

This question appears to be a joke or a play on words, as there's no such thing as a "stochastic parrot" in reality. The term "stochastic parrot" is likely a reference to large language models like myself, which have been described as "stochastic parrots" by some AI researchers.

The phrase comes from a 2021 paper titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" by Emily M. Bender, Timnit Gebru, and others. They used this term to describe how large language models can produce fluent-sounding text without true understanding.

As an AI, I don't actually change lightbulbs or perform physical tasks. The question is likely meant to be humorous by combining a classic joke format ("How many X does it take to change a lightbulb?") with a technical AI term.

If you're looking for a witty response, you could say something like: "It depends on how many training examples of lightbulb-changing the parrots have seen!"

Share Rewrite

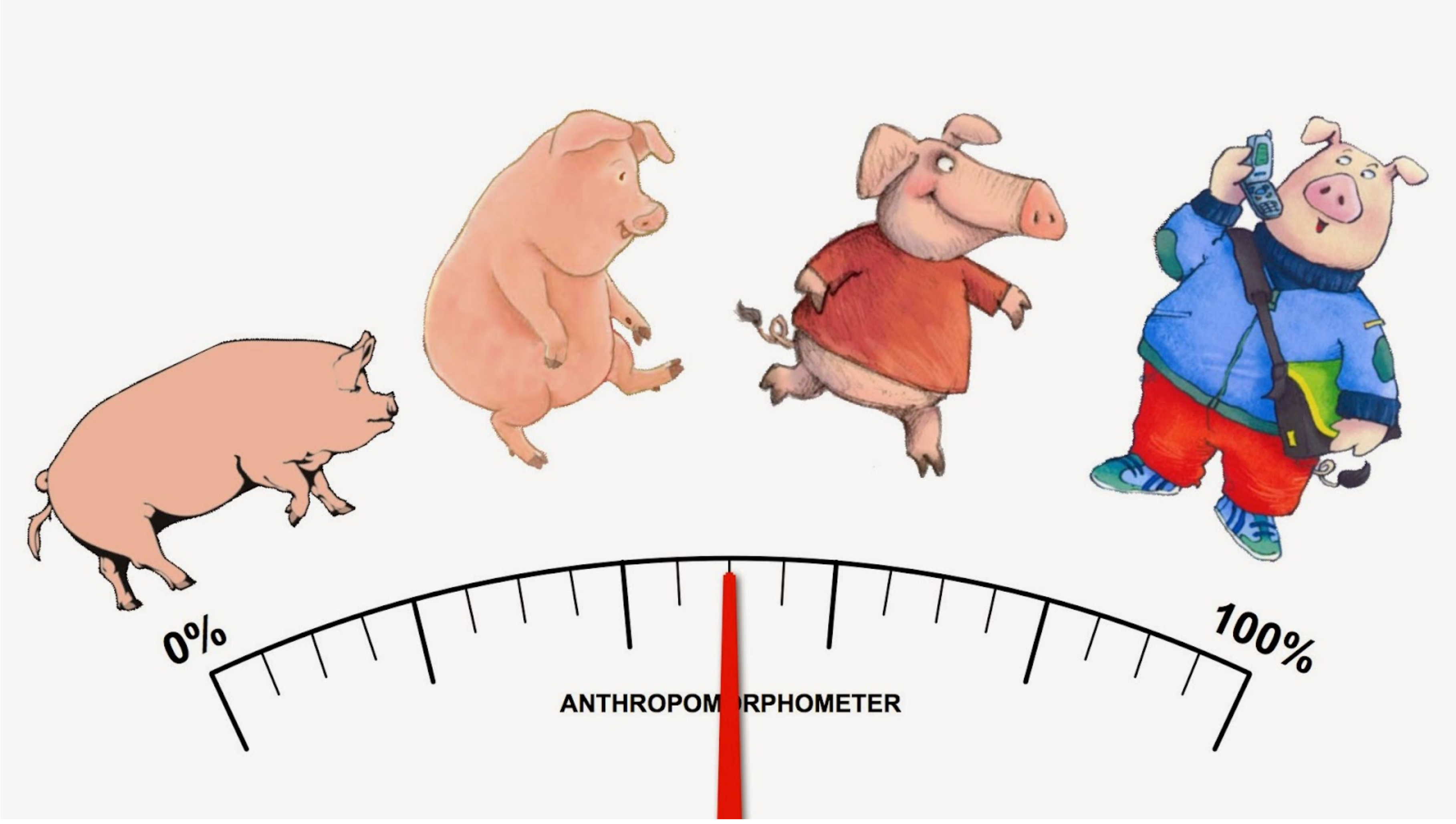
Caveat

I thought of giving a presentation that consisted exclusively of caveats

- I am not an AI expert
    - although I am better educated than most non-experts on the subject
- I am not a computational humor guy
    - Despite some skeletons in my closet
    - For a history of the field, see Taylor (2017), Amin & Burghardt (2020)
- When I say AI what I really mean is LLMs.
    - I have tried to differentiate between various flavors of LLMs ChatGPT 3, 3.5, 4; Bert, Bard, Llama, etc.,
        - but I do not promise I succeeded.
- This is admittedly unfair
    - There is plenty of AI in use that are not LLMs
        - E.g., Speech recognition, adaptive cruise control
- Given how rapidly the AI field is moving nowadays, what I will say is probably already out of date

Anthropomorphism

Attribution of human characteristics to objects



- AI has "hallucinations"
- AI says "I"
- AI "learns," "thinks," "recognizes,"
- Siri/Alexa "listen" [not "detect a sound wave"]
- Ai is "empathetic" (more than doctors!)
- The very term "artificial *intelligence*"
- etc.
- Salles et al. (2020)
- Placani (2024) "Anthropomorphic language is so prevalent in the discipline that it seems inescapable" (p. 692)

Understanding humor means being intelligent/human

Lt. Commander Data (left) and TARS (“a giant sarcastic robot”)

“What’s your humor setting?”
“100%”
“Bring it down to 75, please.”

Humor is AI complete

What evidence was there for this claim?

- The idea was floating around in the late 1990s
- And, full disclosure, I made that claim myself in print (2001, p. 208).
  - See Winters (2021) for discussion
- It was a guess, let’s be honest.
  - Not a terrible one, as it took almost 30 years to be proven wrong
- One of the things we’ve learned is that systems that are obviously not-AGI can handle significant aspects of humor processing/recognition/generation

Why would we even want humor in an AI?

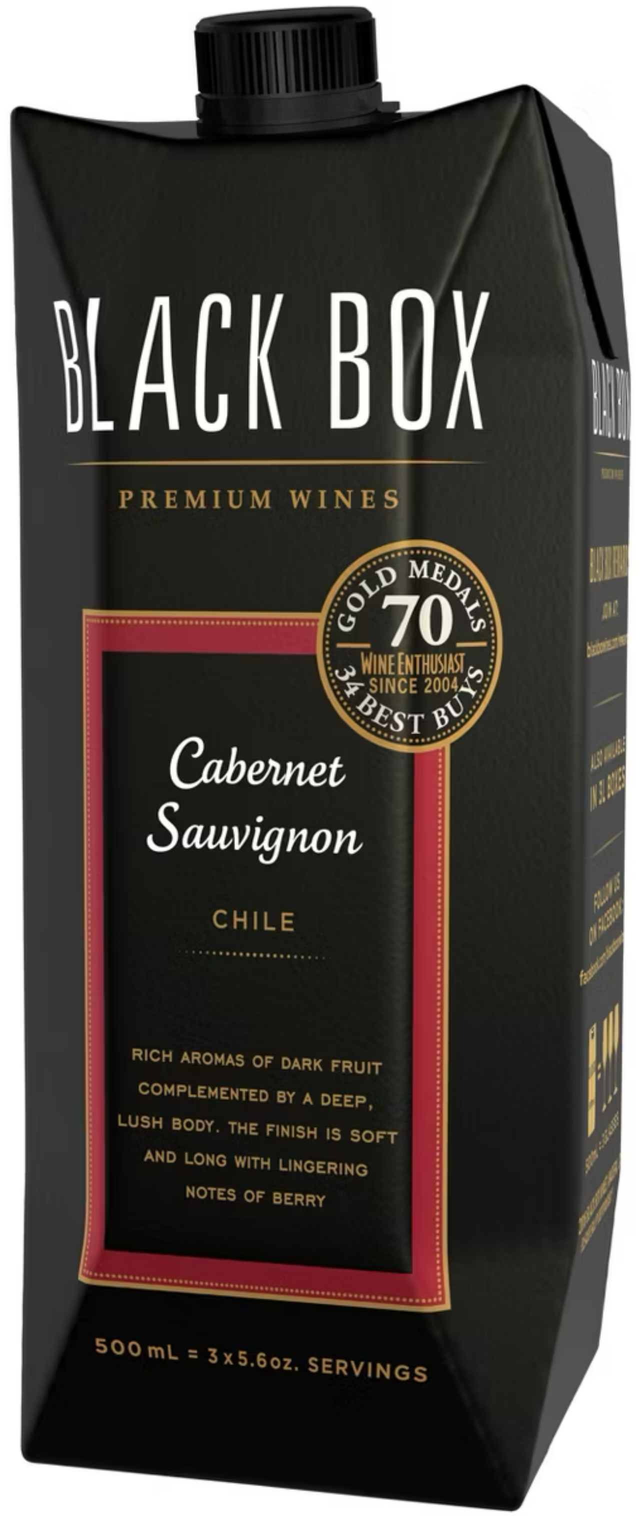
“My calculations are complete: the asteroid will impact the earth in ten days; I guess now is a bad time to ask for a raise?”

- Anthropomorphism
  - we like humor
- HCI
- The “having a beer” component (pro-social bias)
- It may be important to recognize humor (especially irony/sarcasm) for the purpose of sentiment classification
- There is evidence that it can help in teaching L2 (Zhai & Wibowo, 2022)
  - Students like humor

LLMs are black boxes

On several levels, to boot

- “The term "Black Box" in artificial intelligence (AI) and machine learning (ML) refers to systems where internal operations and decision-making processes remain opaque or inscrutable, regardless of the clarity or accuracy of their outputs.” (Wang, 2023)
    - We don’t always have access to the algorithms, training data, and resulting models
    - And even if we did, when the training data number in the billions they cannot be inspected manually
    - There is no way of knowing, unless you built it yourself, to what extent the LLMs are augmented or part of a more complex system
- External validation (e.g., “the system produces an output that users find funny”) is not an explanation



A black box is not an explanation

“obscurum per obscurius”

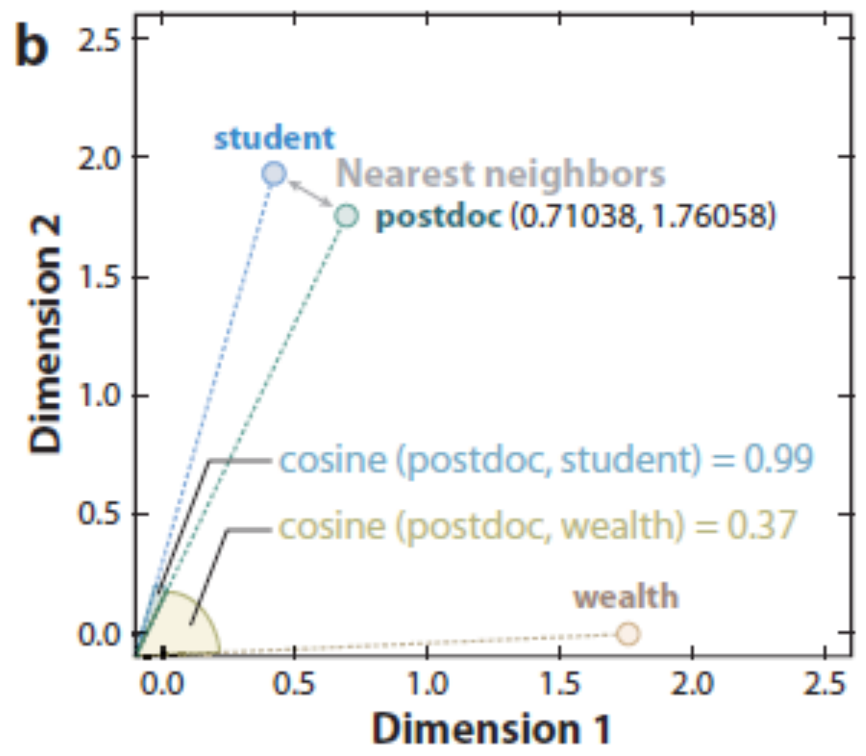
- Marcus & Teuwen (2024) AI models are in need of explanations
- For example, De Grave et al. (2021) and Maguolo & Nanni (2021) show that black box AI systems used spurious information not part of the lung x-rays to achieve accurate Covid diagnoses, such as the borders of the images and not the lungs.
- It is not impossible to extract explanation-directed information from black box systems (see a taxonomy in Bodria et al. 2023)
    - For example, Nugent & Cunningham (2005) show how case-based reasoning can be used to provide (some) explanations of black-box systems

So, does AI get humor?

TL;NR: Sort of

- Popova & Dadic (2023) “artificial intelligence can perform the more or less simple tasks such as pun detection and location with relative success, but more complicated tasks such as pun interpretation and translation still require a lot of improvement. So, does artificial intelligence have a sense of humor? So far not much but [it] is developing it”
- Jentzsch & Kersting (2023) “Over 90% of 1008 generated jokes were the same 25 Jokes. The system accurately explains valid jokes but also comes up with fictional explanations for invalid jokes. Joke-typical characteristics can mislead Chat-GPT [3] in the classification of jokes.
- Goes et al. 2023: many-shots with training and instructions produce a “weak correlation” between GPT 4 and human raters
- Hessel et al. (2023) “human-authored explanations [of New Yorker cartoons] are preferred head-to-head over the best machine-authored ones (few-shot GPT-4) in more than 2/3 of cases.” (p. 688)
- Gorenz & Schwarz (2024) “ChatGPT 3.5-produced jokes were rated as equally funny or funnier than human-produced jokes”
- Suljic & Pervan (2024). “AI-generated texts [Bard/Gemini] in(…) meet the aim stipulated by the prompts—they are humorous, ironic, occasionally sarcastic, but also repetitive regarding some word combinations. The humour is based on the computational prediction that the addressee will recognise incongruencies in lexical choices or style based on their previous knowledge of language and culture.” (p. 262)
- Mirowski et al. (2024) “Generated outputs [various AIs] are generally of poor comedic quality. Many participants noted that they only used LLMs for setup and structure generation due to their inability to generate humorous outputs from the models: `the most bland, boring thing—I stopped reading it. It was so bad’” (p. 1626)
- Wu et al. (2024) [humor classification] “Llama 3-8B performed the best, achieving an accuracy of 89.68%.”
- Li et al. (2024) “cunning texts that are easy for humans to understand but difficult for models to grasp (…) mainly consist of the tricky, humorous, and misleading texts collected from the real internet” (p. 1)
    - Best performance of LLMs
        - detection: GPT4 Turbo w/ CoT = 82.73% vs. Human 93.35%;
        - classification in types ERNIE-Bot-4.0 w/ CoT 14.42% vs. human 63.69%

(Boleda, 2020, p. 214)



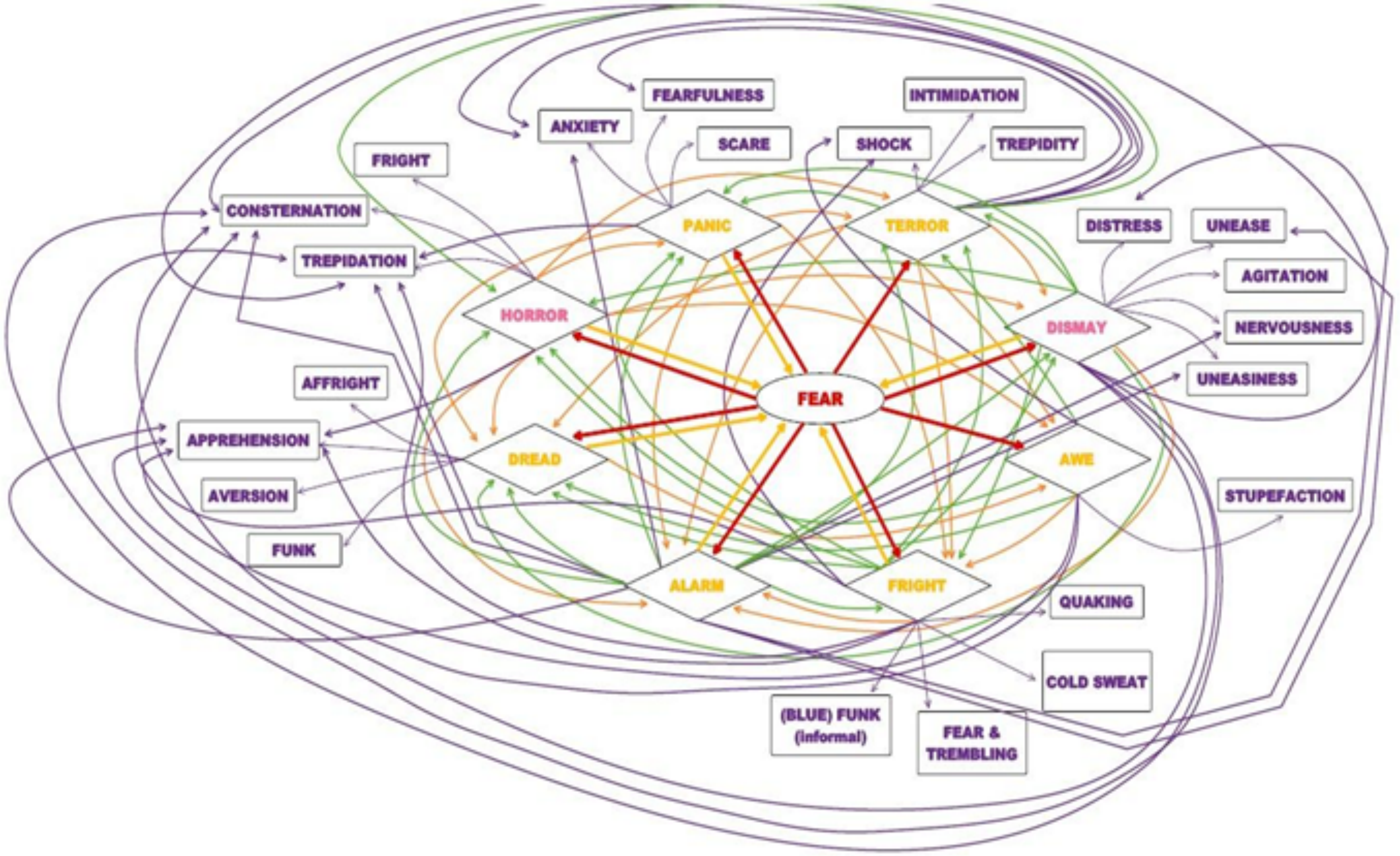
A different perspective

Unintended side effect of LLMs embeddings

- LLMs provides us with a massive scale implementation of semantic field theory (Trier, 1931) a.k.a., lexical field theory
  - Meaning equals the relationships between words in a lexical field
  - Saussurean differences
  - LLM-type semantics is now being called “distributional semantics” (Clark, 2015; Lenci, 2018; Boleda, 2020)
  - The connection between DS and semantic field theory is not always recognized.
  - The terminological mess is a traditional feature of this area of research (Peters, 1991)
- More broadly the unlimited n-dimensional semantic network postulated by frame semantics (Fillmore, Raskin), semiotics (Peirce, Eco) and many others.

Synonyms field for “fear” (Romaniuk, 2020)

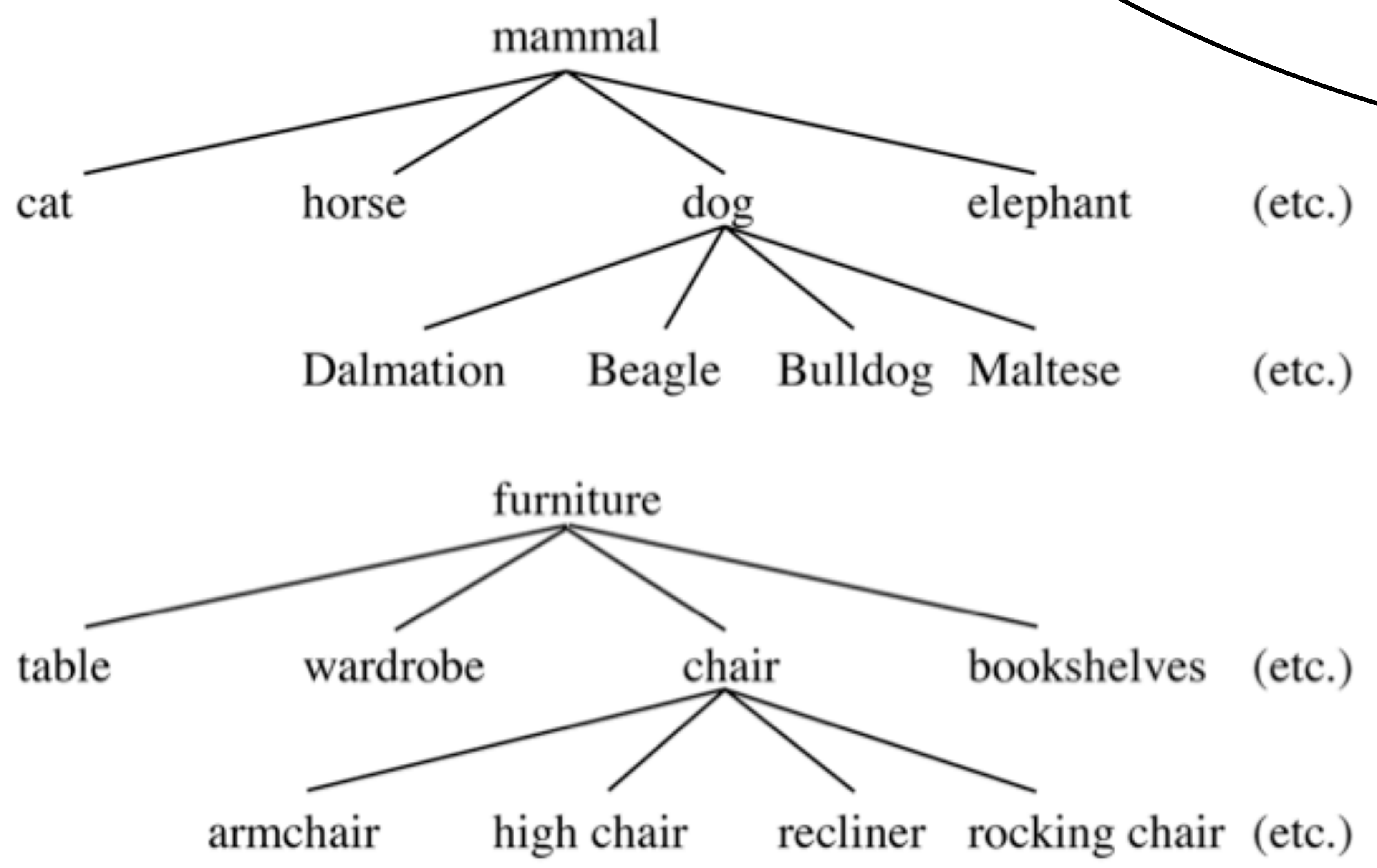
Colors indicate first level synonyms (red), second (purple), etc.

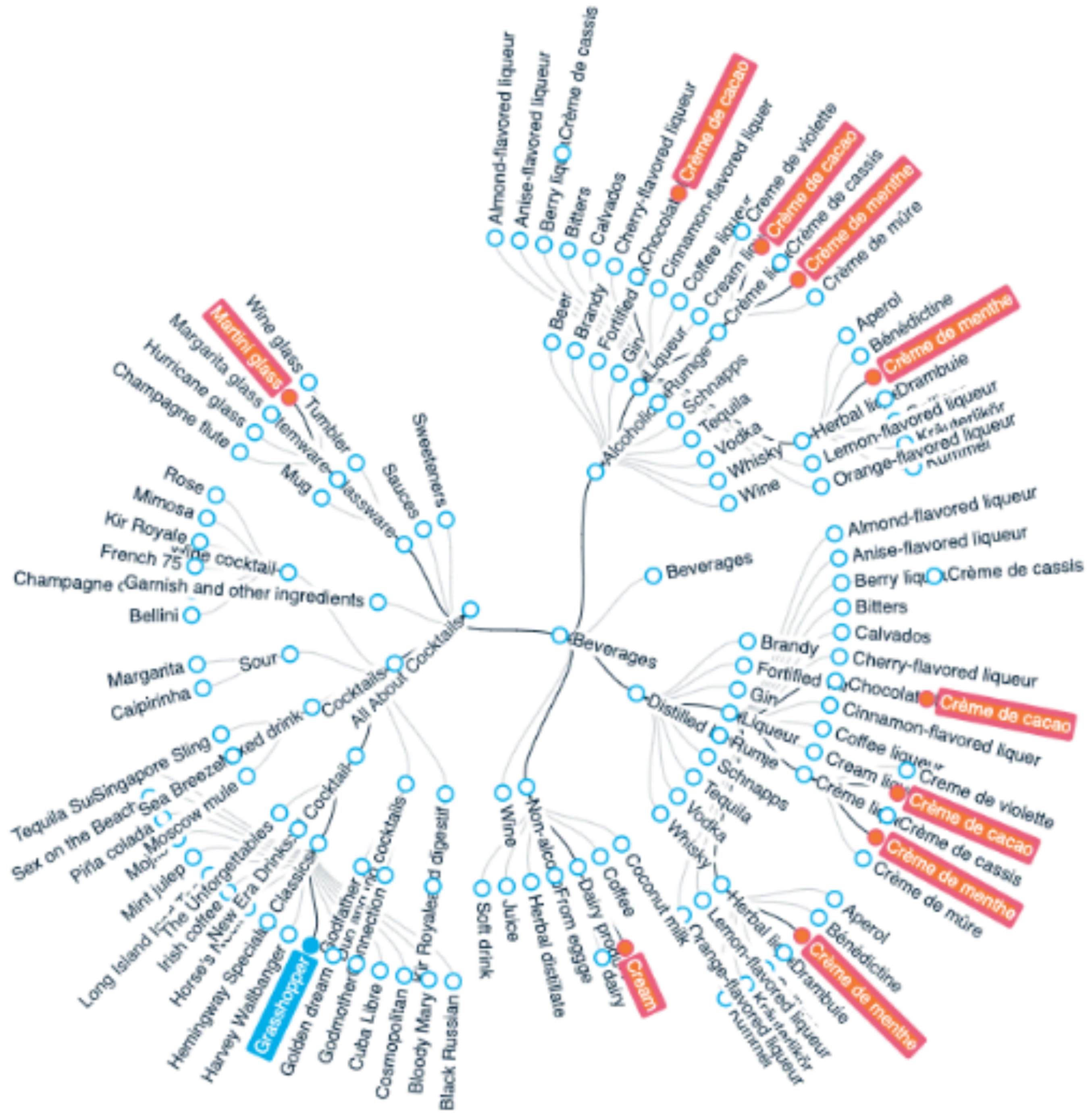


More relationships

Hyponymy (“isa” hierarchy) and hyperonymy

Ontology

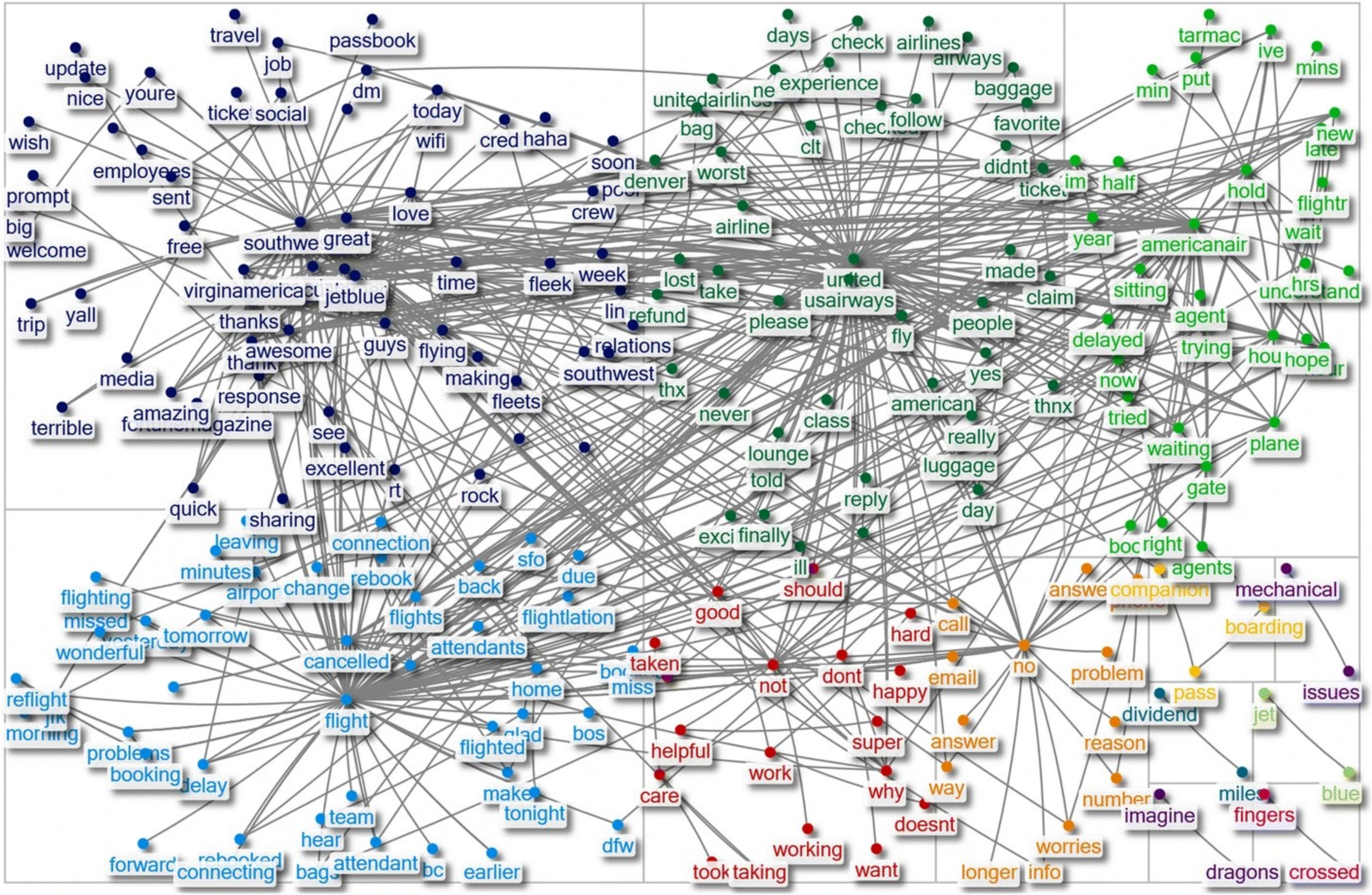




Semantic Roles	Example
Agent + action	*Mommy go*
Action +object	*Push toy*
Agent + object	*Mommy milk*
Action + location/locative	*Come here*
Entity + location/locative	*Toy box*
Possessor and possession	*Mommy hair*
Entity and attributive	*Cup full*
Demonstrative and entity	*This shoe*

Adapted from Brown (1973) *A first Language*. Cambridge, MA: Harvard University Press.

Semantic Network (Danowski et al. 2021)



Validation of Humor Theories through Computational humor

- West & Horvitz (2019) validate some claims form the SSHT and GTVH
  - Script Opposition, Logical Mechanisms, final position of the punch line, etc.
- Chen et al. (2023) validate Toplyn’s model and incongruity theory using GPT 3
- Skalicky & Attardo (2025) [this conference!]
  - Validate incongruity (SO) in puns using semantic distance, conceptualized as cosine distance between vectors

References

Amin, M., & Burghardt, M. (2020). A Survey on Approaches to Computational Humor Generation. In S. DeGaetano, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (pp. 29–41). International Committee on Computational Linguistics. https://aclanthology.org/2020.latechclfl-1.4/

Attardo, S. (2001). *Humorous Texts*. De Gruyter.

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. Data Mining and Knowledge Discovery, 37(5), 1719–1778. https://doi.org/10.1007/s10618-023-00933-9

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. Annual Review of Linguistics, 6(Volume 6, 2020), 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Chen, Y., Shi, B., & Si, M. (2023). Prompt to GPT-3: Step-by-Step Thinking Instructions for Humor Generation. arXiv preprint arXiv:2306.13195.

Clark S. (2015). Vector space models of lexical meaning. In The Handbook of Contemporary Semantic Theory, ed. S Lappin, C Fox, pp. 493–522. New York:Wiley

Danowski, J. A., Yan, B., & Riopelle, K. (2021). A semantic network approach to measuring sentiment. Quality & Quantity, 55(1), 221–255. https://doi.org/10.1007/s11135-020-01000-x

DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3(7), 610–619. https://doi.org/10.1038/s42256-021-00338-7

Góes, L. F., Sawicki, P., Grzes, M., Brown, D., & Volpe, M. (2023). Is gpt-4 good enough to evaluate jokes? https://figshare.le.ac.uk/articles/conference_contribution/Is_GPT-4_Good_Enough_to_Evaluate_Jokes_/24324415/1 [ICC 2023]

Gorenz, D., & Schwarz, N. (2024). How funny is ChatGPT? A comparison of human-and AI-produced jokes. Preprint.

Hessel, J., Marasović, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., ... & Choi, Y. (2022). Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. arXiv preprint arXiv:2209.06293.

Jentzsch, S., & Kersting, K. (2023). ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. arXiv preprint arXiv:2306.04563.

Lenci, A. (2018). Distributional Models of Word Meaning. Annual Review of Linguistics, 4(Volume 4, 2018), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Maguolo, G., & Nanni, L. (2021). A critic evaluation of methods for COVID-19 automatic detection from X-ray images. Information fusion, 76, 1-7.

Marcus, E., & Teuwen, J. (2024). Artificial intelligence and explanation: How, why, and when to explain black boxes. *European Journal of Radiology*, 173, 111393. https://doi.org/10.1016/j.ejrad.2024.111393

Mirowski, P., Love, J., Mathewson, K., & Mohamed, S. (2024, June). A Robot Walks into a Bar: Can Language Models Serve as Creativity SupportTools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1622-1636).

Nugent, C., & Cunningham, P. (2005). A Case-Based Explanation System for Black-Box Systems. Artificial Intelligence Review, 24(2), 163–178. https://doi.org/10.1007/s10462-005-4609-5

Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. AI and Ethics, 4(3), 691–698. https://doi.org/10.1007/s43681-024-00419-4

Peters, B. (1991). A few remarks on terminological insecurity in semantic field theory. Quaderni di semantica, 2, 335-343.

Popova, O., & Dadic, P. (2023). Does AI Have a Sense of Humor? CLEF 2023 JOKER Tasks 1, 2 and 3: Using BLOOM, GPT, SimpleT5, and More for Pun Detection, Location, Interpretation and Translation. CLEF (Working Notes), 1888–1908. https://ceur-ws.org/Vol-3497/paper-160.pdf

Romanyuk, M. Y. (2020). Comparative Modelling Of Lexico-Semantic Fields Of Fear In Russian And English Languages. In A. Pavlova (Ed.), Philological Readings, vol 83. European Proceedings of Social and Behavioural Sciences (pp. 575-582). European Publisher. https://doi.org/10.15405/epsbs.2020.04.02.66

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, 11(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Suljic, V., & Pervan, A. (2024). Creative writing in the hands of artificial intelligence: the analysis of humour in Bard-generated texts. The European Journal of Humour Research, 12(4), 251-268.

Taylor, J (2017). Computational treatments of humor. In S. Attardo (Ed.) The Routledge Handbook of Language and Humor. Taylor & Francis. 456-471.

Trier, J., 1931, Der Deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines sprachlichen Feldes I. Von den Anfangen bis zum Beginn des 13. Jhdts., Heidelberg: Winter.

Wang, Y. (2023). Deciphering the Enigma: A Deep Dive into Understanding and Interpreting LLM Outputs. https://doi.org/10.36227/techrxiv.24085833.v1

West, R., & Horvitz, E. (2019). Reverse-Engineering Satire, or "Paper on Computational Humor Accepted despite Making Serious Advances". Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 7265-7272. https://doi.org/10.1609/aaai.v33i01.33017265

Winters, T. (2021). Computers Learning Humor Is No Joke. Harvard Data Science Review, 3(2). https://doi.org/10.1162/99608f92.f13a2337

Wu, S. H., Huang, Y. F., & Lau, T. Y. (2024). Humour classification by fine-tuning LLMs: CYUT at CLEF 2024 JOKER Lab subtask humour classification according to genre and technique. In Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings (pp. 1933-1947).