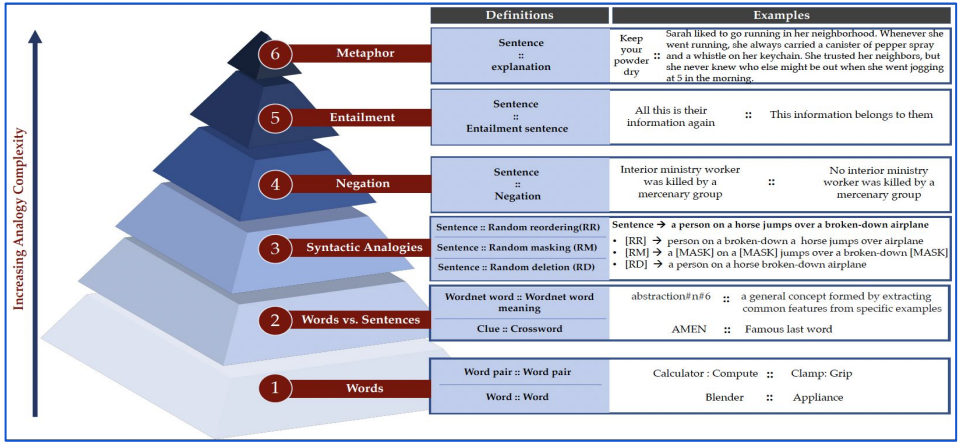
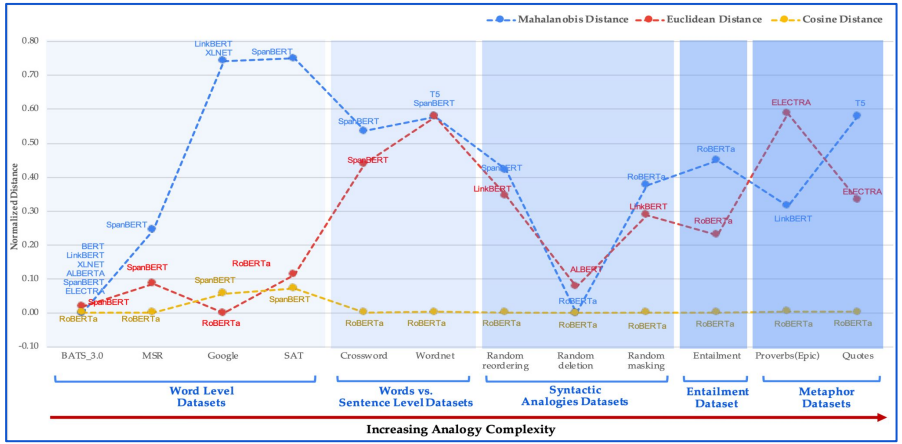
ANALOGICAL

A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models

Thilini Wijesiriwardene[1], Ruwan Wickramarachchi[1], Bimal G. Gajera[2], Shreeyash Mukul Gowaikar[3], Chandan Gupta[4], Aman Chadha[5,6,*], Aishwarya Naresh Reganti[7,*], Amit Sheth[1], Amitava Das[1]

[1]AI Institute, University of South Carolina, USA, [2]Nirma University, India, [3]BITS Pilani, Goa, India, [4]IIIT Delhi, India, [5]Amazon AI, USA, [6]Stanford, USA, [7]Amazon, USA





Word-level analogies intrinsically measure the quality of word embedding methods.

Large Language Models are primarily evaluated on extrinsic measures (e.g., GLUE, SuperGLUE) and there are only a few investigations on LLMs' ability to draw analogies between long texts.

ANALOGICAL, intrinsically evaluate LLMs across a six-level analogy taxonomy: (i) word, (ii) word vs. sentence, (iii) syntactic, (iv) negation, (v) entailment, and (vi) metaphor.

Abilities of 8 LLMs in identifying analogies in this taxonomy was evaluated (using 13 datasets and 3 different distance measures). We find that it is increasingly challenging for LLMs to identify analogies when going up the analogy taxonomy.

*Work does not relate to position at Amazon.

Acknowledgement: Dr. Krishnaprasad Thirunarayan
National Science Foundation (Grant #2133842)

For more information or feedback,
Contact Thilini at thilini@sc.edu