

Measuring Generalization in Neural Language Models: Evaluation and Computational Creativity

Cristiano Ciaccio^{1,2}

¹Department of Computer Science
University of Pisa

²Institute of Computational Linguistics "Antonio Zampolli" (ILC)
CNR, Pisa

Mauriana Pesaresi - Seminars Series 2025/2026

Seminar Overview

The seminar will be structured into four parts:

- A brief introduction to Language Models
- Memorization Vs. Generalization, the problem of evaluation in Neural Language Models (NLMs)
- Computational Creativity in NLP as a tool to measure generalization
- Results, research directions and open problems

A Brief Introduction to Language Models

Language as a probabilistic system → the **Distributional Hypothesis** (Harris, 1954; Firth, 1957): linguistic items with similar **distributions** have similar meanings, in other words: *"you shall know a word by the company it keeps"*

One of the goals of NLP and CL became to develop a **probabilistic model** that could assign probabilities to sequences of natural language:

$$P(x_1, \dots, x_n) = \prod_{t=1}^n P(x_t \mid x_1, \dots, x_{t-1})$$

N-gram models (Markov, 1913; Shannon, 1948): count and memorize, **structurally unable to generalize**, no learned representations.

A Brief Introduction to Language Models

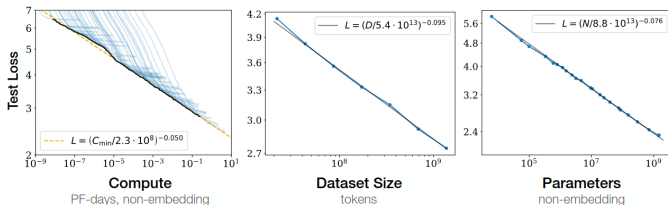
One of the key thrive of NLP/CL research → learning a meaningful and flexible representation of $P(x_t | x_1, \dots, x_{t-1})$ that's **expressive** and **generalize** → Representation Learning (Bengio et al., 2013)



N-Grams (Markov, 1913), Bag-of-Words (BoW) (Harris, 1954) with SVD compression, RNN (Elman, 1990), LSTM (Hochreiter and Schmidhuber, 1997), Neural Probabilistic Language Model (Bengio et al., 2003), Word2Vec (Mikolov et al., 2013), Attention Mechanism Bahdanau et al. (2014), Subword tokenization (Sennrich et al., 2016), Transformers (Vaswani et al., 2017)

A Brief Introduction to Language Models

Scaling laws $\rightarrow$ from GPT-1 (2018) with 117M parameters trained on $\approx$ 1B of tokens to, as an example, Qwen-3 (2025) with 235B parameters trained on $\approx$ 36T of tokens $\rightarrow$ stronger language modeling, emergent abilities and zero-shot task transfer (Radford et al., 2019)



- Models more complex and difficult to interpret/control
- Concerns about memorization and contamination
- Evaluation issues in generative tasks

The debate on Memorization and Generalization

The NLP/CL community was divided by a debate on whether these models understand or simply memorize:

- The stochastic parrots argument (Bender et al., 2021):
 - ▶ No understanding, just pattern matching
 - ▶ Massive scale hides memorization

Counter-positions and experiments in the opposite directions: linguistic creativity (McCoy et al., 2021), Grokking (Power et al., 2022) and Mechanistic Interpretability (Li et al., 2022)

The Problem of Evaluation

Traditional tasks – such as sentiment analysis, PoS tagging, word sense disambiguation, etc. – are mostly solved. The focus today is on higher level tasks, but:

- **Data contamination:** LLaMa-1 (Touvron et al., 2023) pretrain corpus has a 50% average contamination rate of common benchmarks (BBH, HumanEval, HellaSwag, MMLU...) (Singh et al., 2024)
- **Prompt variation and answer position bias** in Multiple-Choice Question Answering (Molfese et al., 2025)
- **Traditional** independent and identically distributed (i.i.d.) **train-test splits often not indicative of good performance** (Hupkes et al., 2023)

"Systematic generalization testing is not the status quo in the field of NLP"
(Hupkes et al., 2023)

Introducing Computational Creativity

Computational creativity is a multidisciplinary field that focuses on the application of computer systems to emulate human-like creative processes (Colton et al., 2012) with the aim of:

- Study the creative process through computational approaches
- Developing co-creative tools for artists, writers, game developers...
- Advancing AI since creativity is a backbone of general intelligence



Computational creativity deals with generative problems where usually there is no gold answer: poetry generation, computational humor, painting and music generation, linguistic creativity etc.

Introducing Computational Creativity

A very influential figure was Margaret Boden, who defined creativity as *the ability to come up with ideas or artifacts that are **new**, **surprising**, and **valuable*** (Boden, 2004).

Creativity is a fundamental feature of human intelligence, and a challenge. [...] AI must try to model creativity (Boden, 1998).

She defined three types of creativity.

- **Combinatorial**: recombining familiar elements in unfamiliar ways
- **Exploratory**: exploring the space of possibilities within explicit rules
- **Transformational**: changing the rules themselves to enable new possibilities

Moving Computational Creativity to NLP

Leverage concepts (tasks, evaluation) from Computational Creativity to assess the functioning and generalization of NLMs instead of benchmark proliferation and LLM-Popcorn (Hovy, ACL 2025).

Given the generative nature of these models, **creativity is an essential and grounded test-bed to advance NLMs reserach:**

- No gold answers → evaluation focus is a must
- Open-ended → tests compositional/exploratory capabilities
- Built-in novelty → bypasses contamination

The field is (implicitly) moving into this direction with novel tasks (Fein et al., 2026, LitBench), evaluation frameworks (Bellemare-Pepin et al., 2026, Divergent Creativity in LLMs) and surveys (Ismayilzada et al., 2024).

Experiments on Combinatorial Creativity

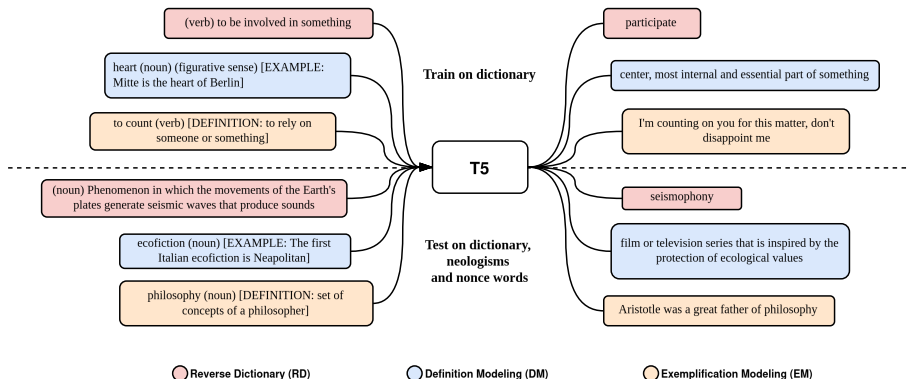
Evaluating Lexical Proficiency in Neural Language Models (Ciaccio et al., 2025a):

- Evaluation framework on lexical proficiency as a proxy to study **creativity** and **generalization** in NLMs
- Tasks: Reverse Dictionary, Definition Modeling, Exemplification Modeling on **existing words** extended to recent **neologisms** and **nonce words**

Conceptual blending (Fauconnier and Turner, 2003)

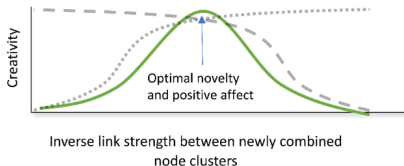
Tasks such as defining and generating novel words require complex **morphological understanding**, **linguistic creativity**, commonsense knowledge and the ability to **generalize over seen concepts**.

Combinatorial Creativity – Experimental Setup



Combinatorial Creativity – Evaluation

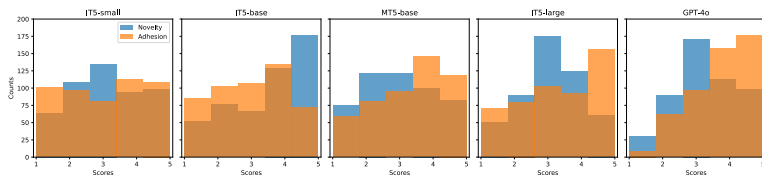
Optimal Innovation Hypothesis (Giora et al., 2004), pleasure ratings of creative artifacts are in an inverse U relationship with novelty (Schubert, 2021; Runco and Jaeger, 2012):



We conducted a human evaluation on the generated nonce words asking annotators to express a value of perceived **novelty** and **adhesion** to the definition. E.g., *The sound of earthquakes* → *seismophony*, novelty = ?, adhesion = ?.

Combinatorial Creativity – Findings

- Performance halves when dealing with neologism wrt existing words (in line with Zheng et al., 2024)
- Larger models can generate novel words that are coherent and aligned with the Optimal Innovation Hypothesis



- **LMs are capable of learning approximations of word formation rules**, rather than relying solely on memorization, thus **showing signs of generalization**

Combinatorial Creativity – Future Directions

A future direction → automate the evaluation process, i.e., develop an automatic metric of lexical creativity:

- Leverage the human annotation data to **train a predictor of adhesion and novelty values**
- Synthetically augment the dataset with the extremes of the two axis
- Preliminary results shows good correlations with human judgments (75% ρ adhesion and 43% ρ novelty)

Outperforms SOTA metrics on creativity (Organisciak et al., 2023, OCSAI).

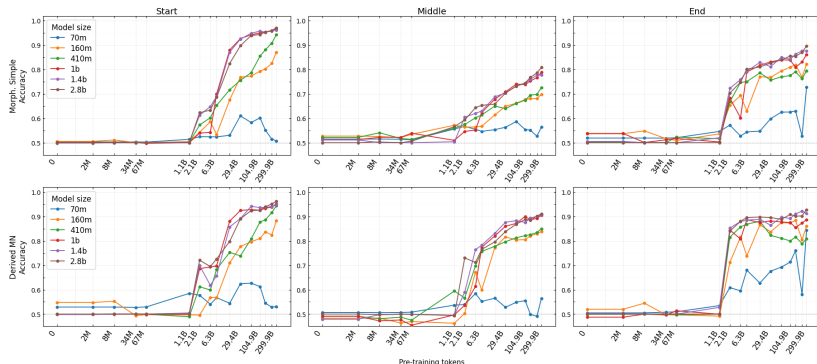
Combinatorial Creativity – Inspecting Substring Awareness

Current NLMs are **character blind** → they work on sequences of subtokens (Sennrich et al., 2016, BPE) that obscure the character composition of words. Despite so, they unexpectedly show some emergent spelling abilities: **Spelling Miracle** (Liu et al., 2023).

Beyond the Spelling Miracle: Investigating Substring Awareness in Character-Blind Language Models (Ciaccio et al., 2025b):

- We investigate NLMs across the space of parameters and pretraining data in a substring recognition task ("Is i in like?")
- Small controlled finetuning to simply teach the task and test on a held-out set (95% unseen tokens)
- We divide the analysis by substring position (start, middle, end) and morphological value ("Is ly in likely" vs. "Is li in likely")

Combinatorial Creativity – Inspecting Substring Awareness



Combinatorial Creativity – Inspecting Representations

Emergence of Property Norms in Language Models Representations: Generalization to Lexical Blends and Nonce Words (Ciaccio et al., work in progress...) → develop an interpretable framework to study how generative NLMs construct **representations** for both **existing words** and lexical innovations such as **blends** and **nonce words**.

The approach:

- Train a linear probing model to **learn a mapping between NLMs representation of existing words** (*rattlesnake*, *alligator*) and a **sparse set of cognitively motivated properties** (*has_a_tail*, *has_four_legs*)
- **Apply the same probing model to the representations of lexical blends** (*rattlegator*, *alligattle*, etc.) and **nonce words** (*berpen*, *connert*) with and without context.

Combinatorial Creativity – Inspecting Representations

Chimera	Blend	Nonce	Sample Passages (instantiated with the blend)	Properties
broccoli + spinach	<i>spinoli</i>	<i>anchet</i>	Raw salad leaves, <i>spinoli</i> [...] are [...] sources of fibre. "Because <i>spinoli</i> happens to be good for her [...]."	Comp. A: a_vegetable, eaten_by_cooking, eaten_with_cheese, [...] Comp. B: a_vegetable, assoc_with_Popeye, is_green, [...]
guitar + harpichord	<i>harpsitar</i>	<i>churd</i>	It 's led by gorgeously picked acoustic <i>harpsitar</i> [...]. This psaltery [...] was the forerunner to the <i>harpsitar</i> .	Comp. A: made_of_wood, has_a_hole, produces_sound, [...] Comp. B: is_large, has_strings, produces_music, [...]
alligator + rattlesnake	<i>rattligator</i>	<i>berpen</i>	[...]Albert reacted like any <i>rattligator</i> with live prey [...]. [...] go poking a stick at a <i>rattligator</i> if [...] want to get bitten.	Comp. A: a_reptile,is_dangerous, has_a_tail, is_scary, [...] Comp. B: a_reptile, eats_rodents, lives_in_deserts, [...]

The probe (PLSR) is trained to map NLMs representation of existing words to their corresponding set of sparse probabilities. Then testing is performed on lexical blends and nonce words. In terms of properties, we considered:

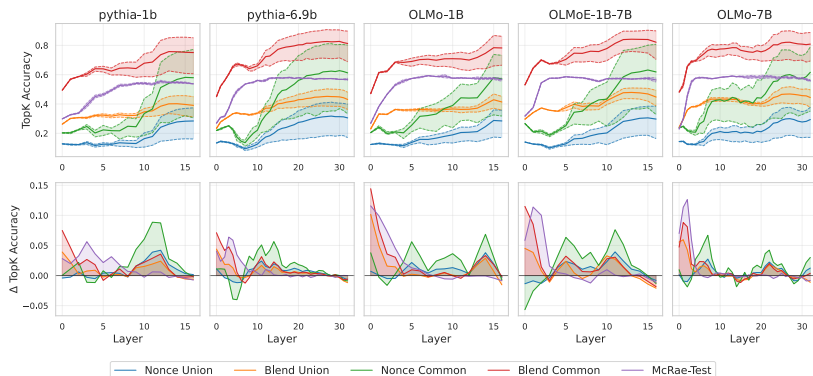
- The union ($\text{Prop A} \cup \text{Prop B}$)
- The intersection ($\text{Prop A} \cap \text{Prop B}$)
- Union without intersection ($\text{Prop A} \Delta \text{Prop B}$)

Combinatorial Creativity – Inspecting Representations

Chimera	Word	Passage	Top-20 Predicted
bomb_missile	bissile	During World War II, several windows, including the great east window, were damaged by a <i>bissile</i> blast. The USAF plans to organise its ground-launched <i>bissile</i> into “flights”, each carrying 16 weapons. The <i>bissile</i>	made_of_metal, used_for_war, a_weapon, used_for_killing, is_long, is_dangerous, inbeh_Explodes, is_fast, is_destructive, is_nuclear, is_expensive, is_long_range, has_warheads, launched_from_airplanes, is_launched, is_loud, is_large, used_by_the_army, beh_flies, is_lethal
alligator_rattlesnake	rattlegator	Animals such as capybara, jaguars, jacare (<i>rattlegator</i>) and hyacinth macaws are particularly vulnerable. Nadirpur stared at it as though it were a <i>rattlegator</i> , his face quite drained. The <i>rattlegator</i>	an_animal, is_long, is_green, is_large, is_dangerous, has_a_tail, has_scales, a_reptile, beh_swims, lives_in_water, has_a_tongue, has_fangs, is_scary, is_furry, beh_eats, beh_slithers, lives_in_deserts, beh_bites, has_4_legs, a_mammal
potato_turnip	turnotato	Collectively they use about 155,000 hectares of land and produce approximately 6,000,000 tonnes of <i>turnotato</i> each year. January 21st, 1846, Thomas Everett, a labourer of Cavendish, for stealing a quantity of <i>turnotato</i> from James Hickford of Cavendish. The <i>turnotato</i>	is_edible, a_vegetable, is_brown, eaten_by_cooking, grows_in_gardens, is_nutritious, is_round, grows_in_the_ground, tastes_good, eaten_in_salads, grows_underground, made_of_wood, is_yellow, eaten_with_butter, is_red, has_skin, eaten_by_baking, is_crunchy, a_food, is_inexpensive
train_bus	connert	Some people find it impossible to sleep a wink on a long night journey in an <i>connert</i> . Barcelona will be receiving 3 hydrogen-fuel <i>connert</i> as part of the pan-European testing of these vehicles in 2003. <i>Connert</i> are so fast that passengers complain about not being able to see any of the beautiful scenery. The <i>connert</i>	used_for_transportation, is_fast, made_of_metal, used_for_passengers, has_wheels, is_loud, is_large, has_an_engine, is_long, has_4_wheels, used_for_travel, requires_drivers, is_expensive, requires_engineers, used_for_freight, has_many_cars, requires_coal, has_a_caboose, used_on_rail_tracks, requires_a_conductor

- There is no representation collapse, the probing could reconstruct relevant properties, even for nonce words (but only in context)
- Blends yielded higher accuracy than nonce words
- Statistically significant correlations between accuracy and the morphological structure/passage informativeness

Combinatorial Creativity – Probing results



Conclusions – Take Home Messages

Computational creativity offers a **grounded evaluation framework** for NLMs generalization, backed by decades of theoretical work in cognitive science and AI (Boden, 2004; Colton et al., 2012).

- **Novel and controlled experimental setups** sidestep contamination by design
- **Creativity is an essential property** of generative NLMs: language is open-ended and so should be its models
- Creative tasks **expose the boundary between memorization and generalization**
- Focus on **tailored evaluation protocols**, involving humans – an interesting open problem in itself
- Investigating creative capabilities can **foster research into the advancing NLMs**

Open Questions and Future Directions

- **Automatic metrics of creativity:** how to develop reliable proxies that correlate with human judgments reducing reliance on expensive human evaluation?
- **Exploring other forms of creativity:** beyond combinatorial, what about exploratory and transformational creativity in NLMs?
- **Impacts of architecture and tokenization:** how do sparse architectures (MoE), agent interaction, tokenization schemes affect exploration of the possibility space and surface emergent creative behavior?
- **Cross-lingual creativity:** do creative capabilities transfer across languages, or are they tied to specific linguistic structures?

Thank you!

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279, 2026.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Margaret A Boden. Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2):347–356, 1998.

Margaret A Boden. *The creative mind: Myths and mechanisms*.
Routledge, 2004.

Cristiano Ciaccio, Alessio Miaschi, and Felice Dell'Orletta. Evaluating lexical proficiency in neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1286, 2025a.

Cristiano Ciaccio, Marta Sartor, Alessio Miaschi, and Felice Dell'Orletta. Beyond the spelling miracle: Investigating substring awareness in character-blind language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11361–11372, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.593. URL <https://aclanthology.org/2025.findings-acl.593/>.

Simon Colton, Geraint A Wiggins, et al. Computational creativity: The final frontier? In *Ecai*, volume 12, pages 21–26. Montpellier, 2012.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2): 179–211, 1990.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7740–7755, 2026.

John Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32, 1957.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, October 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00729-y. URL <http://dx.doi.org/10.1038/s42256-023-00729-y>.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*, 2024.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16270–16297, 2023.

Andrey Andreyevich Markov. An example of statistical investigation in the text of ‘eugene onyegin’ illustrating coupling of ‘tests’ in chains. *Proceedings of the Academy of Sciences of St. Petersburg*, 7:153–162, 1913.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training



data? evaluating linguistic novelty in text generation using RAVEN. *CoRR*, abs/2111.09509, 2021. URL <https://arxiv.org/abs/2111.09509>.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18477–18494, 2025.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.
- Emery Schubert. Creativity is optimal novelty and maximal positive affect: A new definition based on the spreading activation model. *Frontiers in Neuroscience*, 15:612379, 2021.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725, 2016.
- C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- Aaditya K Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. Evaluation data contamination in llms: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024. ▶

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jonathan Zheng, Alan Ritter, and Wei Xu. Neo-bench: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, 2024.