Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Evaluating LLM Linguistic Abilities

Alessio Miaschi
Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa

Istituto di Linguistica Computazionale "Antonio Zampolli"
Consiglio Nazionale delle Ricerche

FAIR TP2: Vision, Language and Multimodal Challenges

- **Our contribution within TP2 focuses on:**
  - Linguistically-grounded evaluation of LLMs;
  - Building probes and benchmarks for morphology, lexicon, and formal linguistic constraints;
  - Studying generalization vs. memorization in Neural Language Models;
  - Specific focus on the Italian language;

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Motivation: Evaluating LLM Linguistic Abilities

- LLMs are evaluated via task-based benchmarks (MMLU, GSM8K, etc.) that measure surface performance, not the depth of linguistic knowledge
- Most benchmarks are English-centric and miss fine-grained linguistic competences
- A comprehensive evaluation of LLMs' linguistic abilities, independent of specific tasks and cross-cutting across them, is still missing

- **Our approach: linguistically-grounded probes and benchmarks to evaluate intrinsic LM capabilities**
  - Focus on morphology, lexicon, formal constraints, character-level awareness;

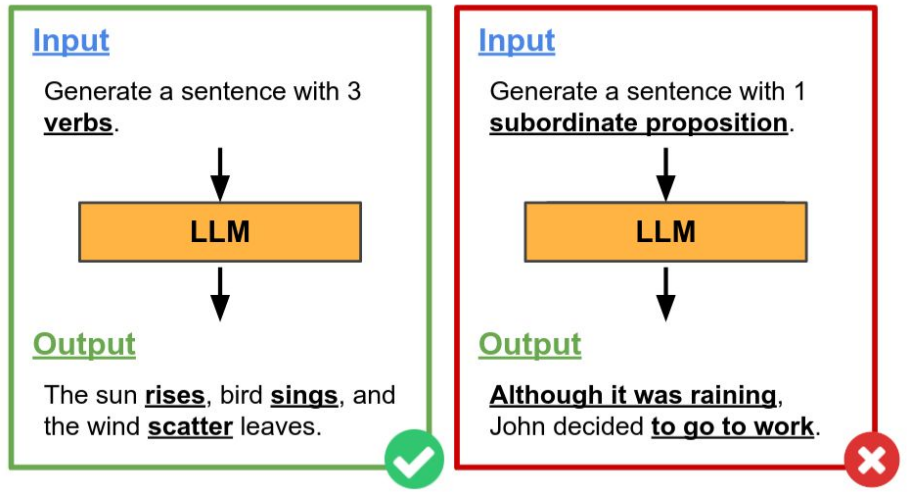
- **6 works presented in this talk, spanning EMNLP 2024, CLiC-it 2024, ACL 2025, and CLiC-it 2025**

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Evaluating LLMs via Linguistic Profiling

- We test LLMs' ability to generate text under specific linguistic constraints (morpho-syntactic and syntactic):
  - Drawing on the "Linguistic Profiling" approach → more than 130+ linguistic features extracted from ProfilingUD (Brunato et al., 2018)
  - 5 LLMs of varying sizes tested in zero-shot and few-shot scenarios

- **Results:**
  - Models struggle with explicit quantitative constraints (e.g., number of subordinates, word length)
  - Few-shot helps but significant gaps remain, especially for syntactic phenomena
  - Models tend to adhere more accurately to morpho-syntactic constraints rather than syntactic ones



Miaschi A., Dell'Orletta F., Venturi G. (2024). Evaluating Large Language Models via Linguistic Profiling. In Proc. of EMNLP 2024, pp. 2835-2848, Miami, Florida.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Controllable Text Generation for Italian LLMs

- Extension of the LLM Profiling approach to Italian LLMs
  - Testing Italian-specific models (LLaMAntino, Camoscio, Minerva) on their ability to generate text adhering to explicit linguistic constraints
  - Automatic verification of constraint adherence via ProfilingUD

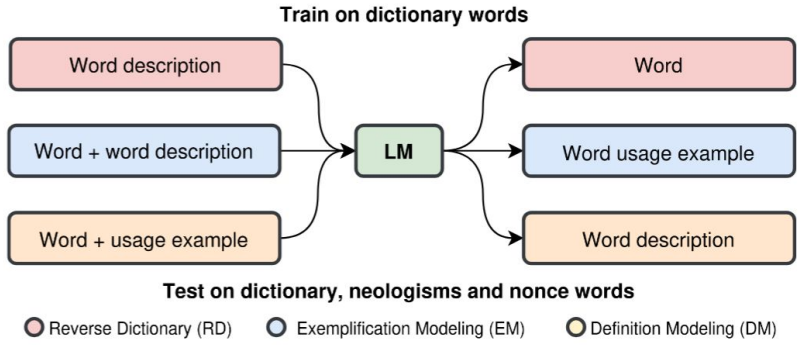
- **Results:**
  - Italian-specific LLMs show uneven performance across constraint types
  - Larger multilingual models outperform Italian-only ones on complex constraints
  - Two-step evaluation: generation + validation reveals gaps between producing text and controlling its properties

Ciaccio C., Dell'Orletta F., Miaschi A., Venturi G. (2024). Controllable Text Generation To Evaluate Linguistic Abilities of Italian LLMs. In Proc. of CLiC-it 2024, Pisa.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Evaluating Lexical Proficiency in Neural Language Models

- Novel framework assessing LMs' lexical creativity across three word categories:
  - Dictionary words (common Italian lexicon from Wikizionario)
  - Neologisms (recently coined words from ONLI)
  - Nonce words (novel/invented words — testing generalization)
- 3 Tasks: Reverse Dictionary, Definition Modeling, Example Modeling
  - Creativity measured via the Optimal Innovation Hypothesis



- **Results:**
  - Larger, monolingual models generally outperformed their multilingual counterparts
  - LMs are capable of learning approximations of word formation rules, rather than relying solely on memorization, thus showing signs of generalization

Ciaccio C., Miaschi A., Dell'Orletta F. (2025). Evaluating Lexical Proficiency in Neural Language Models. In Proc. of ACL 2025, pp. 1267-1286, Vienna.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Beyond the Spelling Miracle

- Most PLMs are “character-blind” (operate on subword tokens) yet acquire some character knowledge during pre-training (the “Spelling Miracle”)
- We systematically evaluate PLMs on a controlled binary substring identification task
  - “Is ‘*dis*’ in ‘*dislike*’?”

- **Three axes of investigation:**
  - Where? Layer-wise probing across Pythia models of different sizes
  - When? Training dynamics across checkpoints
  - How? Morphemic (prefixes, suffixes, roots) vs. random n-gram substrings

- **Key finding:**
  - Substantial amounts of data are necessary for character knowledge to emerge
  - Morphemic substrings are recognized significantly better than random n-grams — models develop partial morphological awareness through pre-training

Ciaccio C., Sartor M., Miaschi A., Dell’Orletta F. (2025). Beyond the Spelling Miracle: Investigating Substring Awareness in Character-Blind Language Models. In Findings of ACL 2025, pp. 11361-11372, Vienna.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

The OuLiBench Benchmark

- Inspired by the OuLiPo literary movement: constrained writing as a probe for LLM competence
  - Novel Italian benchmark with formal constraints: lipograms, palindromes, sentence anagrams, character length control, morpho-syntactic requirements
- Models tested: GPT-4o mini, Claude, Gemini, DeepSeek, Minerva, Anita, Velvet, Maestrale
  - Zero-shot and few-shot configurations
- **Results:**
  - Huge performance gaps between closed-source and open Italian LLMs
  - Performance gap between tasks involving quantitative constraints and those requiring more structural or stylistic control

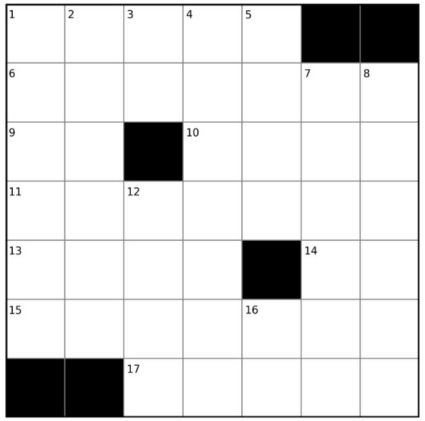
Calderaro S., Miaschi A., Dell'Orletta F. (2025). The OuLiBench Benchmark: Formal Constraints as a Lens into LLM Linguistic Competence. In Proc. of CLiC-it 2025, Cagliari.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Crossword Space: Latent Manifold Learning for Italian Crosswords

- We proposed a collection of siamese and asymmetric dual encoder architectures for the task of clue-answering in the context of Italian crosswords
  - Models trained on 20K+ clue-solution pairs from Italian crossword puzzles
  - Augmented with dictionary entries and neologisms

- **The learned manifold generalizes to:**
  - Definition matching and reverse dictionary tasks
  - Neologism understanding
  - Different linguistic settings that share the clue-solution inferential structure



Across:
(1) Il nome di Stern, il violinista,
(6) Isola dell'Arcipelago Toscano,
(9) Università Cattolica,
(10) Lo Zamorano calciatore,
(11) É un ottimo solvente,
(13) Il fiume che bagna Terni,
(14) Massini del teatro (iniz.),
(15) Molti abitano all'Asmara,
(17) La città con le contrade

Down:
(1) Grandi lucertole crestate,
(2) Così è detto il gioco del calcio negli Stati Uniti,
(3) Sono nel Garda e nel Lario,
(12) La dea della vendetta,
(4) Scossi dal nervosismo,
(5) Rifugio per animali,
(16) Se scappa, va in esilio,
(7) Tentò di raggiungere il Polo Nord con la nave Fram,
(8) Chi ne soffre, è smorto in viso

*Ciaccio C., Sarti G., Miaschi A., Dell'Orletta F. (2025). Crossword Space: Latent Manifold Learning for Italian Crosswords and Beyond. In Proc. of CLiC-it 2025, Cagliari. [🏆 **Best Student Paper** 🏆]*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Conclusion and Future Directions

- Linguistically-grounded evaluation reveals blind spots that standard benchmarks miss
- LLMs show gaps in morpho-syntactic constraint adherence, but shows signs of lexical generalization
- Italian-specific evaluation is crucial: open Italian LLMs still lag behind Larger Multilingual Models
- Formal and creative constraints (OuLiPo-inspired, crossword-based) offer novel evaluation lenses

- **Future Directions:**
  - Studying generalization of LLMs across different scenarios, domains, and languages
  - Towards a sustainable, community-driven Italian LLM evaluation framework

Thank you!

alessio.miaschi@ilc.cnr.it