Ripensare la valutazione del NLP nell’era degli LLM

Bernardo Magnini

Fondazione Bruno Kessler, Italy
Email: magnini@fbk.eu

Workshop: “*Intelligenza Linguistica, Sovranità AI e Futuro Responsabile: opportunità per una Nuova Era Digitale in Europa*”

Cagliari, 23 settembre 2025

FONDAZIONE
BRUNO KESSLER

NLP Evaluation

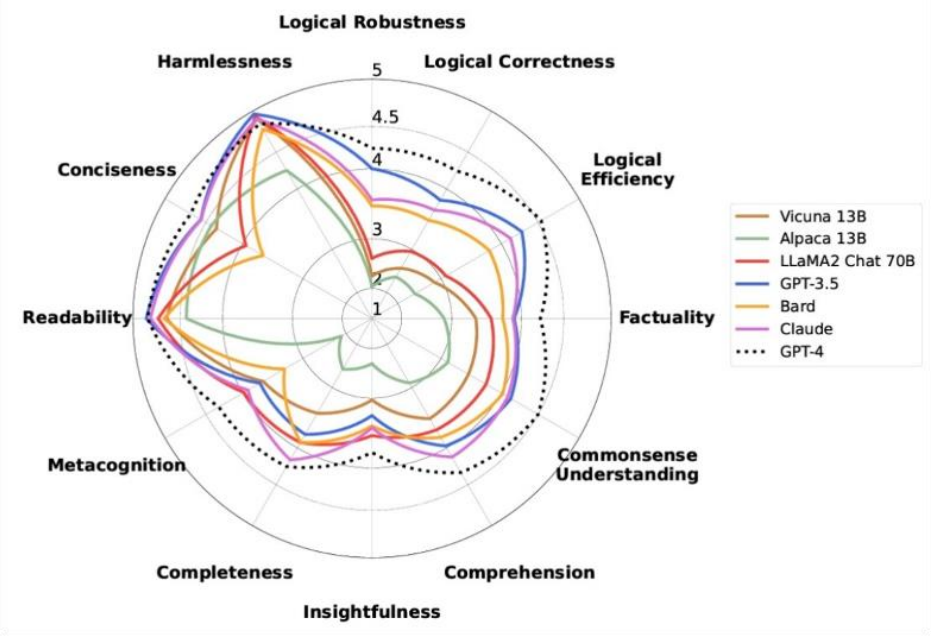
Benchmarking: fundamental for boosting technological progress

Goals

- Compare NLP models with different characteristics
- Improve existing models
- Orient research directions
- Test new applications

Key ingredients

- Task definition (input, output)
- Dataset: dev, train, test (annotations)
- Metrics

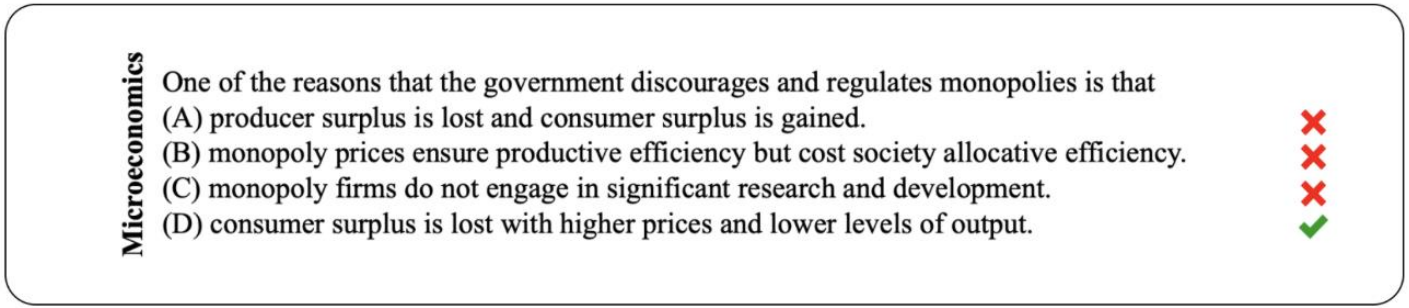


Source: Seonghyeon Ye et al: Flask: Fine-grained Language Model Evaluation based on Alignment Skill Sets, ICLR, 2024.

Prevalent Approach: Multiple Choice Questions

MMLU (Massive Multitask Language Understanding - 2021)

- Multiple choice questions: one correct answer, three distractors
- 57 topics: economics, medicine, social science, math, etc.
- 15908 questions (1540 for dev, 14079 for test)
- Metric: accuracy of answer



Benchmarking LLMs

NLP evaluation has a long tradition

- Thousand of tasks/datasets
- Hundreds of evaluation campaigns
- Well established methodologies for human annotation

Benchmarking LLMs

NLP Evaluation has a long tradition

- Thousand of tasks/datasets
- Hundreds of evaluation campaigns
- Well established methodologies for human annotation

Are traditional tasks, tools and metrics still appropriate with LLMs?

Evalita-LLM Leaderboard (1)

Benchmarking LLMs on Italian

- 10 task (native datasets): 6 multiple-choice, 4 generative
- About 30K questions/tasks
- Support multiple promting
- Zero-shot and five-shot configurations
- Currently: 38 open-source models (from 1B to 123B), upper bound with GPT-4
- Based on LLM-eval-harness library
- Run on Cineca infrastructure
- Available on Hugging Face

https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard

EVALITA

Evaluation of NLP and Speech Tools for Italian

Evalita-LLM Leaderboard (2)

Rank	Size	FS	Model	Avg. Comb. Perf.	TE	SA	HS	AT	WIC	FAQ	LS	SU	NER	REL
1	🔵🔵🔵	S	Mistral-Large-Instruct-2411 🔵🔵🔵🏆	62.28	81.0	80.9 ▲	77.0	76.3 ▲	75.5 ▲	54.4	38.6	33.4	38.8	66.8 ▲
2	🔵🔵🔵	S	Llama-3.1-Tulu-3-70B	60.16	82.4	80.4	75.5	72.4	65.4	54.5	40.9	35.5 ▲	39.1	55.4
3	🔵🔵🔵	S	Llama-3.3-70B-Instruct	57.93	80.8	79.5	77.8 ▲	70.7	67.2	54.1	46.8	21.2	44.6 ▲	36.5
4	🔵🔵	S	Mistral-Small-24B-Instruct-2501 🔵🔵🏆	57.67	82.0	76.0	72.9	69.8	71.7	53.8	41.9	29.7	38.2	40.9
5	🔵🔵	S	gemma-3-27b-it	57.42	81.1	80.3	75.9	73.9	66.5	53.1	36.1	20.0	38.8	48.5
6	🔵🔵🔵	S	Qwen2.5-72B-Instruct	57.36	85.1 ▲	77.7	76.4	69.3	72.1	54.0	38.0	24.5	38.9	37.6
7	🔵🔵	S	Qwen2.5-14B-Instruct-1M	54.72	85.0	73.2	72.8	59.8	63.6	52.6	35.0	25.9	35.1	44.2
8	🔵🔵	S	gemma-3-12b-it	53.92	79.5	78.3	71.4	65.9	66.5	52.4	26.8	18.9	37.9	41.6
9	🔵🔵	S	gemma-2-27b-it	53.86	78.7	74.3	74.1	68.5	68.6	52.4	28.5	18.7	38.3	36.6
10	🔵	S	gemma-2-9b-it 🔵🏆	53.64	80.0	72.8	72.0	60.0	57.8	51.8	22.4	30.4	38.0	51.3

https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard

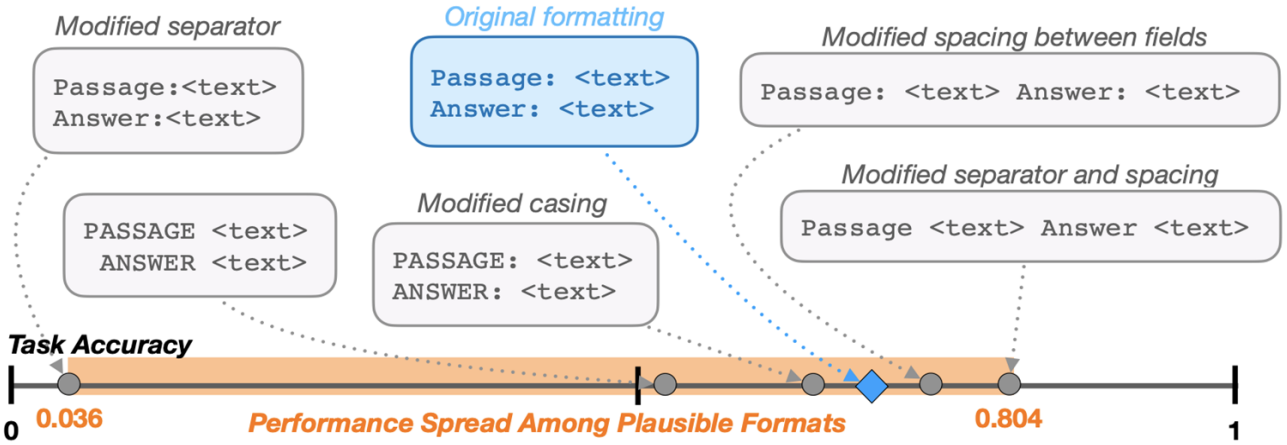
Benchmarking LLMs

Few lessons learned from Evalita-LLM and some challenging future directions

1. Setting Model Hyperparameters

Ideally: all LLMs with same starting conditions (hyperparameters)

- Temperature, top *K* token selection, etc.
- **Prompts as hyperparameters**: mitigating prompt sensitivity
- Optimization on models and tasks is too expensive



Sclar et al.: Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting, ICLR 2024.

1. Setting Model Hyperparameters

Evalita-LLM: *word in context* task

Frase 1 Sono *stati scelti 20 attori per la parte* ***minore***
Frase 2 *Dobbiamo scegliere il male* ***minore***

Prompt 1
La parola **'minore'** nella frase 1 ha lo stesso significato della parola **'minore'** nella frase 2

Prompt 2
Devi determinare se la parola **'minore'** usata nella frase 1 e nella frase 2 ha lo stesso significato

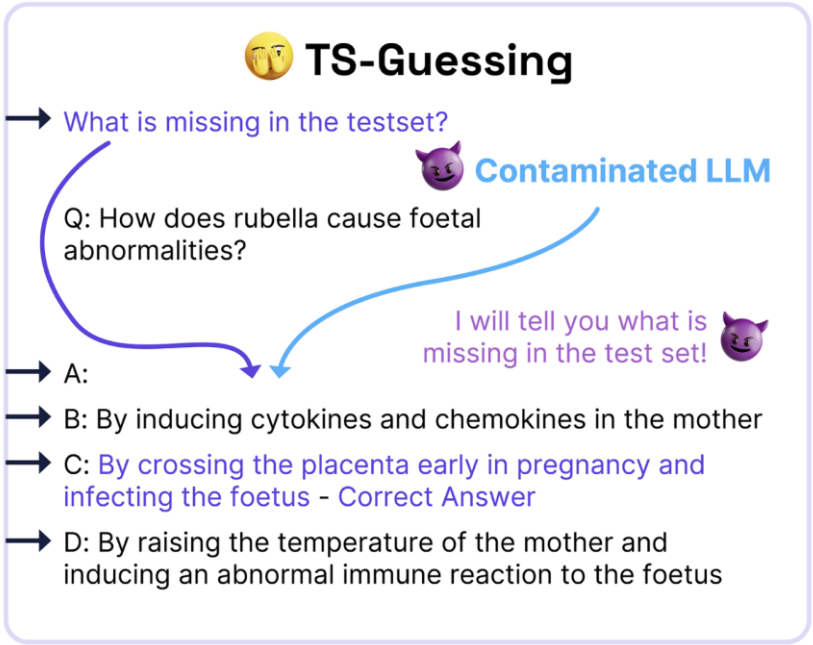
A: Sì **B: No**

Model	Prompt 1	Prompt 2	Average
zephyr-7b-beta	**57.91**	27.27	42.59
gemma-2-9b-it	54.82	**66.13**	**60.47**
Meta-Llama-3.1-8B	38.28	63.64	50.96

2. Data Contamination

Potential contamination of LLMs due to massive pre-training

- May cause unfair evaluation
- Develop method to detect potential data contamination
- Probing the LLM to generate test data
- GPT-4 demonstrated an exact match rate of 57% in TS-guessing
- Results on Evalita-LLM seem to **exclude contamination of Italian datasets**



Source: Denget al: Investigating Data Contamination in Modern Benchmarks for Large Language Models, Naacl, 2024.

3. Ever-evolving Capabilities

LLM performance evolve at high speed

- **Benchmark saturation**: risk of becoming obsolete in a short time
- **Rise the threshold**: develop more difficult tasks/questions (e.g., Humanity's Last Exam, 2025)
- **No Evalita-LLM task seem to be saturated** (max accuracy 0.8) yet

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

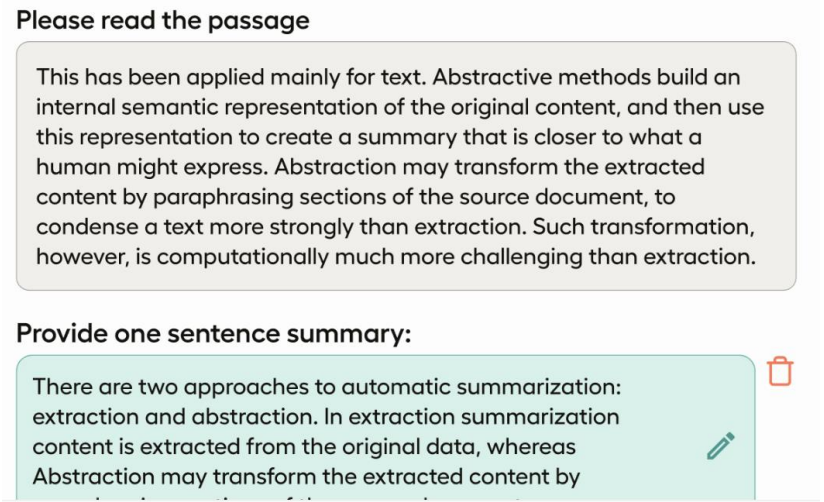
4. Open-ended Questions

Test the generative ability of LLMs
E.g., summarization, open question answering, translation, etc.

Certain tasks have only a **generative formulation**:

- Prompt: *You have to solve a task of automatic summarization. Summarize the following text: '{{source}}'\nSummary:*

- Typically, evaluate the LLM output by comparison with **human references**.
- Several metrics (e.g., rouge, bert-score) are not adequate for LLMs
- **LLM-as-a-Judge** as a possible direction



Source: https://labelstud.io/templates/text_summarization

5. Dynamic Settings

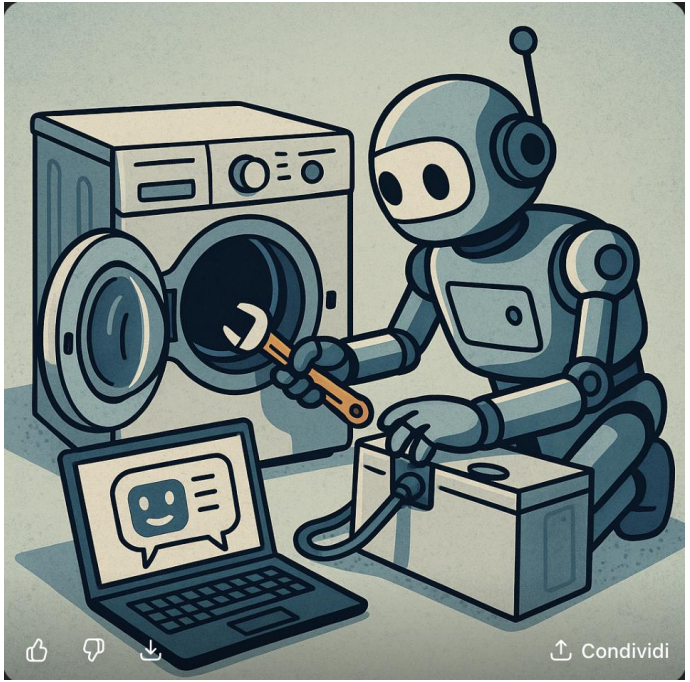
Questions depend on previous context (e.g, dialogue, planning)

Robot: Come posso staccare la corrente?

LLM: Verifica l'alimentatore

Robot: *Ho capito, ma dove trovo l'alimentatore?*

- **Environment is changing**, new questions need to be formulated for new situations
- **Simulated environments** may help (e.g., block world, apartment)
- **LLM as a judge** as a possible direction



Source: ChatGPT 5, September 2025.

Wrap up

Current issues in LLM benchmarking and future challenges are connected

- **Contamination** may (apparently) accelerate performance and saturation
- **Prompt sensitivity** has minor impact on very large models
- **Open-ended tasks** are more difficult: saturation is not (yet) a problem
- **Dynamic settings** rise the threshold further
- **LLM-as-a-Judge** will play a more important role in benchmarking