BEEP - BEst DrivEr's License Performer: A CALAMITA Challenge

Fabio Mercorio[1,3], Daniele Potertì[2], Antonio Serino[2] and Andrea Seveso[1,3,*]

[1]Dept of Statistics and Quantitative Methods, University of Milano Bicocca, Italy
[2]Dept of Economics, Management and Statistics, University of Milano Bicocca, Italy
[3]CRISP Research Centre crispresearch.eu, University of Milano Bicocca, Italy

Abstract

We present BEEP (BEst DrivEr's License Performer), a benchmark challenge to evaluate large language models in the context of a simulated Italian driver's license exam. This challenge tests the models' ability to understand and apply traffic laws, road safety regulations, and vehicle-related knowledge through a series of true/false questions. The dataset is derived from official ministerial materials used in the Italian licensing process, specifically targeting Category B licenses. We evaluate models such as LLaMA and Mixtral across multiple categories. In addition, we simulate a driving license test to assess the models' real-world applicability, where the pass rate is determined based on the number of errors allowed. While scaling up model size improved performance, even larger models struggled to pass the exam consistently. The challenge demonstrates the capabilities and limitations of LLMs in handling real-world, high-stakes scenarios, providing insights into their practical use and areas for further improvement.

Keywords

Large Language Models, Benchmarks, CALAMITA, CLiC-it

1. Challenge: Introduction and Motivation

In recent years, Large Language Models (LLMs) have become a significant breakthrough in Natural Language Processing (NLP) and Artificial Intelligence (AI) [1]. Assessing model performance is crucial yet challenging, involving multiple critical attributes: models must be precise, resilient, fair, and efficient, among other characteristics [2].

Developing effective models in underrepresented languages such as Italian is a continuing challenge [3]. This disparity arises from limited and lower-quality data [4] and a development process often prioritising Anglocentric perspectives [5]. Recently, there has been a surge in research aimed at making LLMs more culturally inclusive, moving beyond mere multilingualism to address deeper cultural contexts [6]. For instance, a structured benchmark utilising the INVALSI tests—well-established assessments measuring educational competencies across Italy—represents one such effort to embed culturally relevant content in model evaluation [7].

This work is part of CALAMITA [8] (Challenge the Abilities of LAnguage Models in ITAlian), an initiative launched by AILC, the Italian Association for Computational Linguistics. CALAMITA aims to develop a comprehensive and evolving benchmark for evaluating the capabilities of LLMs in Italian. The goal is to establish a shared platform with a suite of tasks and a live leaderboard, allowing for ongoing assessments of Italian and multilingual LLMs. CALAMITA seeks to build this benchmark through community-driven challenges, inviting researchers to propose tasks and datasets that evaluate specific aspects of LLMs' performance in Italian. This paper contributes to this collaborative effort by presenting a benchmark that assesses LLMs' ability to comprehend and apply Italian driving regulations, forming one of the initial tasks in this evolving benchmark.

This challenge evaluates LLM's ability to comprehend and apply knowledge in a practical, real-world scenario. While LLMs have shown remarkable capabilities in understanding and generating human language, their effectiveness in real-world decision-making scenarios remains underexplored, especially in languages such as Italian. This challenge tests whether these models can perform effectively in a linguistically demanding and contextually rich domain. Success in this challenge would demonstrate the model's ability to generalise language understanding to practical tasks, a crucial step towards their broader application in everyday life.

2. Challenge: Description

BEst Driv**Er**'s License **P**erformer (BEEP) is a challenge benchmark that focuses on assessing LLMs through a

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
andrea.seveso@unimib.it (A. Seveso)
0000-0001-6864-2702 (F. Mercorio); 0009-0006-6525-4492 (D. Potertì); 0009-0008-0737-8547 (A. Serino); 0000-0001-7132-7703 (A. Seveso)

simulated driver's license exam in Italian. This task requires a deep understanding of traffic laws and reasoning through driving situations.

In Italy, obtaining a driver's license is a structured process involving theoretical and practical assessments to ensure drivers are well-versed in road safety, traffic regulations, and practical driving skills. The Italian driver's license process is governed by strict rules set forth by the Ministero delle Infrastrutture e dei Trasporti (Ministry of Infrastructure and Transport), and the license is recognised across the European Union.

Italy offers several categories of driver's licenses, depending on the type of vehicle a person wishes to operate. We focus on Category B, which is required for cars (up to 3.5 tons) and vehicles with up to 8 seats.

The theoretical exam is crucial to obtaining a driver's license in Italy, and it is required, along with the practical exam. It assesses the applicant's knowledge of traffic laws, road signs, and driving regulations. It consists of multiple-choice questions and is typically administered electronically. The candidate must understand traffic regulations, road signs, driving behaviour, and vehicle maintenance. A Category B license test typically consists of 30 questions; a candidate can pass up to 3 errors.

The licensing process is not just about learning the rules; it requires candidates to internalise and apply them practically. BEEP reflects this focus on real-world application and safety. The Italian driving system also emphasises road etiquette and the ability to navigate complex traffic situations, particularly in high-density urban areas. Consequently, the challenge aims to mirror this complexity in evaluating LLMs.

3. Data description

3.1. Origin of data

BEEP is derived from the publicly accessible PDF "Listato A e B", which includes all quiz questions related to Italian driver's license examinations provided by the official ministerial listing[1]. The quizzes consist of true or false questions for driving license categories A and B, with data updated as of 01/07/2020.

We extracted the data from the official PDF file. The text is segmented by identifying distinct patterns indicating the start of new questions and sections. These segments are classified into predefined categories and sub-categories. For each text segment, relevant metadata, question types (e.g., true/false) and related image numbers are extracted and compiled into a structured format. The final dataset is exported, offering a well-organised collection of questions for the evaluation.

[1]Visit ListatoAB for more information at https://www.neca.it/assets/pdf/ListatoAB.pdf.

3.2. Data format

The dataset is formatted with the following columns:

- **Categorisation Structure** - Each question in the dataset is organised within a hierarchical categorisation system consisting of **Major Categories**, **Minor Categories**, and **Subcategories** to ensure precise classification. For example, the Major Category *"Road Signage"* includes Minor Categories like *"Warning Signs"* and *"Prohibition Signs"*, which further break down into Subcategories detailing specific signs such as *"Speed Limit Signs"*;
- **Question Text** - The actual content of the question;
- **True Answer** - Can be either true or false;
- **Figure** - A reference for the accompanying figure, if present.

3.3. Example of prompts used

Question

The road can be divided into lanes.

Options

[**A. True**, **B. False**]

Options

Instructions:
You must return the letter corresponding to the correct answer in square brackets.
Answer format: [letter]

Answer

[**A**]

Figure 1: An example question, with instructions and a correct answer highlighted.

We exclusively employed the zero-shot setting in our evaluation process, where no prior examples were provided. An illustrative example of a prompt used in this setting is shown in Figure 1, which demonstrates the structure and input format supplied to the model. The decision to have the language model answer with '[letter]' rather than simply 'letter' or 'True/False' is due to our use of pattern matching for response extraction. By enforcing a consistent answer format with brackets, we

Table 1
An overview of the dataset categorised by major and minor traffic-related topics. The columns display the number of entries, the percentage of those entries containing figures, and the proportion of correct answers for each category.

Category				
Major	**Minor**	**Rows**	**Percent with Figures**	**True Answer (%)**
DOCUMENTS	MANDATORY DOCUMENTS, AGENTS AND LICENSE PLATES	261	—	129/261 (49.4%)
VEHICLE EQUIPMENT	VISUAL SIGNAL DEVICES AND LIGHTING	98	—	53/98 (54.1%)
	STATIONARY VEHICLE SIGNALS AND ROAD OBSTRUCTIONS	54	—	26/54 (48.1%)
VEHICLES	CLASSIFICATION OF VEHICLES	106	—	48/106 (45.3%)
MOTOR VEHICLE	VEHICLE COMPONENTS	119	—	63/119 (52.9%)
	TIRES, ADHERENCE AND STABILITY	134	—	68/134 (50.7%)
	WARNING LIGHTS AND SYMBOLS	61	100.00	28/61 (45.9%)
ACCIDENTS AND INSURANCE	CAUSES OF ACCIDENTS	566	—	303/566 (53.5%)
	CIVIL AND CRIMINAL LIABILITY AND INSURANCE	123	—	53/123 (43.1%)
ROAD	ROAD AND TRAFFIC DEFINITIONS	203	—	102/203 (50.2%)
TRAFFIC REGULATIONS	STOPPING AND SAFE DISTANCE	129	—	62/129 (48.1%)
	STOP, STANDING AND PARKING	208	—	121/208 (58.2%)
	DRIVING ON HIGHWAYS	59	—	31/59 (52.5%)
	SPEED LIMITS	81	—	45/81 (55.6%)
	RIGHT-OF-WAY RULES AND PROCESSIONS	457	86.87	235/457 (51.4%)
	POSITION ON ROADWAY, DIRECTION CHANGE AND LANE	27	70.37	13/27 (48.1%)
	SPEED REGULATION	96	—	56/96 (58.3%)
	OVERTAKING	156	—	82/156 (52.6%)
	TRANSPORT OF PEOPLE, LOAD ARRANGEMENT, PANELS AND TOWING	110	—	55/110 (50.0%)
FIRST AID	FIRST AID TO INJURED PEOPLE	96	—	48/96 (50.0%)
TRAFFIC SIGNS	SUPPLEMENTARY PANELS	59	100.00	27/59 (45.8%)
	TRAFFIC LIGHT SIGNALS AND POLICEMAN	218	96.33	105/218 (48.2%)
	PROHIBITION SIGNS	409	100.00	198/409 (48.4%)
	INFORMATION SIGNS	536	100.00	253/536 (47.2%)
	MANDATORY SIGNS	402	100.00	190/402 (47.3%)
	WARNING SIGNS	473	100.00	228/473 (48.2%)
	PRIORITY SIGNS	201	100.00	99/201 (49.3%)
	ROAD MARKINGS	147	100.00	73/147 (49.7%)
	TEMPORARY AND SUPPLEMENTARY SIGNS	189	100.00	89/189 (47.1%)
SAFETY AND POLLUTION	SEAT BELTS, AIRBAG AND PROTECTIVE HELMET	135	—	70/135 (51.9%)
	ENVIRONMENTAL AND NOISE POLLUTION	110	—	64/110 (58.2%)

Table 2
Overall accuracy of different models across major dataset categories, allowing for comparison of their effectiveness within these distinct areas.

Category	llama-3-8b	llama-3-70b	mixtral-8x7b	mixtral-8x22b
DOCUMENTS	53.26%	66.28%	67.43%	79.69%
VEHICLE EQUIPMENT	51.97%	66.45%	71.71%	75.00%
VEHICLES	51.89%	77.36%	82.08%	84.91%
THE MOTOR VEHICLE	56.13%	82.61%	82.21%	86.56%
ACCIDENTS AND INSURANCE	59.22%	85.78%	85.49%	91.15%
THE ROAD	51.72%	70.94%	71.92%	81.77%
RULES OF CONDUCT	54.36%	71.11%	70.34%	76.85%
FIRST AID	61.46%	90.62%	86.46%	88.54%
ROAD SIGNAGE	37.50%	75.00%	100.00%	100.00%
SAFETY AND POLLUTION	65.31%	88.57%	85.71%	88.57%

can reliably parse responses, reducing ambiguity and ensuring that variations in phrasing or formatting do not interfere with accurate evaluation.

3.4. Detailed data statistics

The questions are organised into the categories described in Tab. 1. This table summarises statistics across various road safety and vehicle regulation categories, providing detailed insight into major and minor classifications. Each entry in the table is categorised into broad Major Categories such as "DOCUMENTS," "Vehicle Equipment," and "Road Signage," which are further subdivided into more specific Minor Categories. For example, the major category "DOCUMENTS" includes the minor category "Mandatory Documents, Agents, and License Plates," highlighting different aspects of document requirements and administrative details.

We also include figures associated with specific questions, particularly those addressing traffic signals, road signs, and right-of-way scenarios. These visual elements provide additional context and enhance the comprehension of complex traffic situations. However, for the CALAMITA challenge, we opted not to include questions containing figures, focusing solely on text-based questions. This decision ensured that the evaluation of LLMs remains centred on their language comprehension, knowledge and reasoning abilities rather than visual processing capabilities. Including images would limit participation to multimodal models, excluding many language models that cannot process visual information. By using only text, we maintain a broader, more accessible benchmark.

4. Metrics

Since the dataset comprises questions that can only be answered with true and false, we involved the *Overall Accuracy* to evaluate the models' answers in our task. Overall accuracy is commonly used in classification tasks, particularly in true-false or binary decision evaluations [9]. It measures the proportion of all correct predictions (true positives and negatives) out of the total number of predictions made. In other words, it quantifies how well a binary classification system performs by indicating the fraction of correctly classified instances (both positive and negative classes) relative to the total number of instances evaluated.

Table 3
Overall accuracy of selected models, ranging from LLaMA to Mixtral, demonstrating their performance on the dataset.

Model	Overall Accuracy
llama-3-8b-instruct	56.27%
llama-3-70b-instruct	77.23%
mixtral-8x7b-instruct	77.19%
mixtral-8x22b-instruct	83.29%

Table 3 shows the Overall Accuracy obtained by *LLAMA3 8B - Instruct*[2] and others State of the Art models. We evaluate the metrics on the portion of our dataset that does not require image processing operations. The scaling laws hold as it is observed that performance increases with the number of parameters.

Table 2 shows the Overall Accuracy stratified by Major Category for each tested model. Models perform better in the "SAFETY AND POLLUTION", "FIRST AID", and "ACCIDENTS AND INSURANCE" categories. This may be possible given the generality of these major categories, as opposed to more niche categories such as 'DOCUMENTS' or 'VEHICLE EQUIPMENT', where the performance is worse.

4.1. Simulated Driving License Test

We also test the models by simulating a proper driving licence exam, following the appropriate official guidelines

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

and creating a new indicator. We sampled 1000 samples of 30 questions from the dataset, ensuring each sample was unique. We then counted the correct and incorrect answers for each sample and each evaluated model. The guidelines state that the test is passed if the number of wrong answers is less than or equal to 3. Therefore, we built an indicator for each model that considered the percentage of driving licence exams passed, related to the number of examinations attempted. The results are shown in Tab. 4. As expected, smaller models made many mistakes on average (around 13), which was fatal as it never passed the test in any of the attempts. Even larger models like Mixtral-8x22b did not perform well in most cases. However, we believe more advanced models, such as GPT-4, might succeed more reliably.

Table 4
Driving license Metrics of the Selected Models

Model	Total Tests Passed (%)	Avg Errors (Std.)
llama-3-8b-instruct	0/1000 (0%)	13.17 (±2.71)
llama-3-70b-instruct	64/1000 (6.4%)	6.88 (±2.65)
mixtral-8x7b-instruct	61/1000 (6.1%)	6.79 (±2.24)
mixtral-8x22b-instruct	258/1000 (25.8%)	5.01 (±2.09)

It is important to note that this simulated test is not integral to the CALAMITA benchmark. While it provides additional insights into the models' performance in a high-stakes, applied setting, the official evaluation metric focuses solely on overall accuracy.

5. Limitations

Considering state-of-the-art LLMs, it is possible that one's training sets are contaminated with examples from the U.S. driving licence test and that these may influence performance on our benchmark. Furthermore, although the benchmark allows the real driving licence test to be reproduced, it can only assess true-or-false binary answers and not dialogue or reasoning ability.

6. Ethical issues

Although the models may demonstrate positive performance in this benchmark, it is crucial to recognise that such results do not equate to an actual ability to drive or navigate safely in real-world environments. The benchmark assesses the models' ability to process and understand driving-related questions, a far cry from the complex task of driving a vehicle, which requires perception, decision-making and real-time motor control.

7. Data license and copyright issues

The data are publicly available online and not subject to copyright restrictions.

Acknowledgments

We thank Thomas Passera for providing the initial code for the dataset's extraction. Evaluation of the open-source models was conducted on Leonardo supercomputer with the support of CINECA-Italian Super Computing Resource Allocation, class C project IsCb7_LLM-EVAL (HP10CIO7T9).

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 2023. URL: http://arxiv.org/abs/2307.03109. arXiv:2307.03109.

[2] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2022).

[3] S. Ruder, N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, D. Garrette, G. Neubig, et al., Xtreme-r: Towards more challenging and nuanced multilingual evaluation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021.

[4] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, et al., Quality at a glance: An audit of web-crawled multilingual datasets, Transactions of the Association for Computational Linguistics 10 (2022) 50–72.

[5] Z. Talat, A. Névéol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev, et al., You reap what you sow: On the challenges of bias evaluation under multilingual settings, in: Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models, 2022, pp. 26–41.

[6] S. Pawar, J. Park, J. Jin, A. Arora, J. Myung, S. Yadav, F. G. Haznitrama, I. Song, A. Oh, I. Augenstein, Survey of cultural awareness in language models: Text and beyond (2024).

[7] F. Mercorio, M. Mezzanzanica, D. Potertì, A. Serino, A. Seveso, Disce aut deficere: Evaluating llms proficiency on the invalsi italian benchmark, arXiv preprint arXiv:2406.17535 (2024).

[8] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[9] C. M. Bishop, Pattern recognition and machine learning, Springer google schola 2 (2006) 1122–1128.