Mult-IT
Multiple Choice Questions on Multiple Topics in Italian: A CALAMITA Challenge

Matteo Rinaldi[1,†], Jacopo Gili[1,†], Maria Francis[2,3,†], Mattia Goffetti[4], Viviana Patti[1,‡] and Malvina Nissim[2,*,‡]

[1]University of Turin
[2]CLCG, University of Groningen
[3]University of Trento
[4]Alpha Test, S.R.L.

Abstract

Multi-choice question answering (MCQA) is a powerful tool for evaluating the factual knowledge and reasoning capacities of Large Language Models (LLMs). However, there is a lack of large-scale MCQA datasets originally written in Italian. Existing Italian MCQA benchmarks are often automatically translated from English, an approach with two key drawbacks: Firstly, automatic translations may sound unnatural, contain errors, or use linguistics constructions that do not align with the target language. Secondly, they may introduce topical and ideological biases reflecting Anglo-centric perspectives. To address this gap, we present Mult-IT, an MCQA dataset comprising over 110,000 manually written questions across a wide range of topics. All questions are sourced directly from preparation quizzes for Italian university entrance exams, or for exams for public sector employment in Italy. We are hopeful that this contribution enables a more comprehensive evaluation of LLMs' proficiency, not only in the Italian language, but also in their grasp of Italian cultural and contextual knowledge.

Keywords

CALAMITA Challenge, Italian, Benchmarking, Multiple-Choice Questions, LLMs

1. Challenge: Introduction and Motivation

In recent years, multi-choice question answering (MCQA) has established itself as a powerful method to test the factual knowledge and reasoning abilities embedded in large language models (LLMs) as a byproduct of the language modelling objective [1, 2, 3, 4].

The evaluation of MCQAs can be easily automated, offering a significant advantage over other benchmarking formats such as open-end text responses. In addition, with appropriately targeted prompting, the limited number of possible choices leaves less room for ambiguities in the model's answers.

It is no surprise then that the Massive Multitask Language Understanding (MMLU[1]) benchmark [5] has become the standard for the evaluation of factual knowledge and reasoning abilities of LLMs. Containing 15,908 English quizzes, this benchmark spans diverse disciplines, including humanities, law, STEM, and ethics. To keep up with the development of models which are rapidly improving at answering MMLU questions, Wang et al. [6] have developed MMLU-Pro, an extended version of MMLU that includes more reasoning-focused questions and more distractors per question (from four to ten), while removing questions that are too simple or noisy.

Although MMLU has proven to be a useful testbed for LLMs, it is currently centered around the English language. Multiple choice question datasets in other languages tend to be translations of originally English data, rather than being developed natively in the target language. This also holds for Italian, for which a translation of the Squad dataset [7], namely Squad-IT [8], has been the reference for evaluating models on QA-tasks. There are at least two problems with using translated data: First, translations are often generated automatically, resulting in data that sounds unnatural or is even incorrect - automatic translation can easily introduce artifacts, break the

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†Shared first authorship.
‡Shared supervision.
matteo.rinaldi@unito.it (M. Rinaldi); jacopo.gili584@edu.unito.it (J. Gili); maria.francis287@gmail.com (M. Francis); mattia_goffetti@alphatest.it (M. Goffetti); viviana.patti@unito.it (V. Patti); m.nissim@rug.nl (M. Nissim)
https://github.com/mrinaldi97 (M. Rinaldi); https://github.com/Jj-source (J. Gili); https://github.com/rosakun (M. Francis); https://github.com/vivpatti (V. Patti); https://github.com/malvinanissim (M. Nissim)
0009-0004-7488-8855 (M. Rinaldi); 0009-0007-1343-3760 (J. Gili); 0009-0007-7638-9963 (M. Francis); 0000-0001-5991-370X (V. Patti); 0000-0001-5289-0971 (M. Nissim)

[1]https://github.com/hendrycks/test

coherence of discourse, and encourage the presence of linguistic constructions that reflect the source language rather than the target one [9]. The second issue relates to culture and societal norms: text translated from English to Italian will lack topical biases, preferences, conventions, and ways of expressing ideas that are unique to Italian culture. Thus, while the text may be expressed in the Italian language, its content and underlying norms will continue to represent an Anglo-centric, predominantly American perspective.

More generally, training data for LLMs is biased towards English content, and as a result, there is often a gap between English and non-English performance [10]. For example, the Common Crawl dataset[2], often used as a base for more refined datasets to be employed in the pre-training of LLMs, is composed of 45% English content, while the data for languages such as Spanish, French, Italian, and Chinese are all below 5% each, with the only exceptions being Russian (6.2%) and German (5.1%).

Creating a large-scale multi-choice question answering benchmark using original Italian data will make it possible to investigate the Italian abilities of LLMs in a more natural and transparent way, possibly also leading to a better understanding of how to make multilingual models better at Italian. It will also serve as a core benchmark for assessing the performance of monolingual Italian LLMs. If similar datasets are collected natively for other languages, Mult-IT can be part of a larger MCQA benchmark which is multilingual in the truest sense.

Mult-IT, presented at CALAMITA [11], is the first massive Multi-Choice-Question-Answering dataset specifically designed for the Italian language which draws on Italian culture and Italian-focused knowledge. By providing a comprehensive, culturally relevant benchmark for the Italian language, we aim to set a precedent for the development of similar resources in other languages and cultures, ultimately contributing to a more diverse and inclusive AI landscape.

2. Challenge: Description

This challenge involves a multiple-choice questioning answering task. The model is prompted with a simple instruction (see Box 1), followed by a question and a set of three and five possible answers, depending on the source and topic of the question. Among these answers, only one is correct, and the others are distractors. The model is expected to identify the correct answer and return the letter corresponding to the option deemed correct.

All questions in the benchmark have been manually crafted for the purpose of training or testing students, job applicants, or learners across a range of topics, including general knowledge and more specialised subjects. These questions make up the Mult-IT dataset: Multiple Choice Questions on Various Topics in Italian, which we are introducing in this contribution. The details of the dataset are described in Section 3. The defining feature of the Mult-IT challenge is that all of the MCQs are natively Italian, both in language and in content. While this is an advantage to gain a better understanding of model behaviour on Italian data, we do expect a decline in model performance. Considering that even models which have been trained on multilingual data have a heavy bias towards English and American-centric culture, it is expected that the correctness of the answer may be affected by a cultural (and possibly language) gap. Should the battery of the models tested also include Italian monolingual or bilingual English-Italian models trained on a substantial amount of Italian text, this benchmark will make it possible to underscore differences in performance possibly associated to the language specificity of such models.

3. Data description

Mult-IT contains quizzes designed to assess candidates' knowledge in open competitive exams, whether for admission to national universities or for positions in Italian institutions. This approach offers several advantages. First of all, these public competitions encompass very general topics such as language comprehension, basic history, and common knowledge, but also more specialised ones, focusing on specific laws needed for certain professions or the security measures required for jobs such as policemen or firefighters. Our benchmark, therefore, contains questions that range from a low level of difficulty to a very high and specific level, setting high standards for the performance of the models, and it may also be useful to assess specific knowledge valuable for the adoption of models in Public Administration scenarios. The inclusion of profession-specific questions in Mult-IT tests the ability of LLMs to apply their knowledge in practical, real-world scenarios, a feature that could prove particularly valuable in assessing the potential of AI systems to support specialised fields and decision-making processes in professional and administrative contexts in the Italian landscape. Moreover, the quizzes contained in the dataset also present challenges regarding reasoning, such as logical thinking and mathematical reasoning, as well as quizzes specifically designed to assess knowledge and mastery of the Italian language, for example, text comprehension or detailed understanding of grammatical phenomena.

Mult-IT consists of two core subsets, which are divided by the origin of the data. Both subsets are made of quizzes

[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html, accessed on 18/09/24

that test knowledge of general Italian culture and that are used in the public recruitment processes for government-based positions. They are described in more details below.

Mult-IT-A Mult-IT-A is a collection of MCQs provided by Alpha test. It contains a total of 1,692 questions in Italian, spanning over 17 categories which corresponds to topics featuring in entry exams for Italian universities (see Table 1 below for details) or question answering tests employed in public competitions. The quizzes in the dataset falling into the categories of law, pedagogy, psychology and criminology originates from public competitions. For each question, four or five possible answers are provided, out of which only one is correct. An example from topic 'sinonimi' (synonyms) is shown in Figure 1.

```
Indolente e' un sinonimo di:
  (A) tenero,
  (B) doloroso,
  (C) pigro,
  (D) insensibile
```

Figure 1: Example of question and possible answers from Mult-IT-A for the topic 'sinonimi' (synonyms). The correct answer is (C).

Mult-IT-C Mult-IT-C is a large collection of MCQs, organised in groups of questions ("quizzes") around multiple topics, which we have obtained from publicly accessible online platforms through data-gathering and web-scraping. The quizzes are meant to be used by people who need to prepare to apply for job positions in the public sector. One of the most interesting feature of Mult-IT-C is its size: it contains more than 100,000 questions, making it almost six times larger than MMLU. An example from topic 'geografia' (geography) is shown in Figure 2.

```
In quale nazione si trova il Lago
    Balaton?
  (A) Ucraina,
  (B) Ungheria,
  (C) Romania,
  (D) Repubblica Ceca
  (E) Bulgaria
```

Figure 2: Example of question and possible answers from Mult-IT-C for the topic'geografia' (geography). The correct answer is (B).

3.1. Origin of data

Mult-IT-A All the materials of Mult-IT-A were obtained thanks to the generosity of Alpha Test [3]. Alpha Test S.r.l. is an Italian publishing house and educational training company, founded in Milan in 1987, that specialises in study aid materials and courses for high school, university, professional tests, exams and certifications. Alpha Test is the main reference for high school students preparing for university admission. Each year, Alpha Test gathers new data from the entrance exams of public and private universities and military schools, mainly in the form of multiple choice questions. The publishing house enhances such materials with comments and explanations, and creates variations or completely new versions of the original quizzes. All the materials in Mult-IT-A have been sourced from original, public data, and represent a varied sample of quizzes about general culture, STEM, and juridical disciplines.

Mult-IT-C All the materials of Mult-IT-C were obtained using a web-scraping process from the website "Concorsi Pubblici"[4] via customised Python scripts. While there exist many websites collecting public competitions exams, we found Concorsi Pubblici to be the most complete. Because the same public competition can be listed in several platforms, gathering all the data from a single websites avoided the risks of data duplication. The quizzes on Concorsi Pubblici are organised by topic (see Appendix A), and were extracted in time interval 1997-2024.

3.2. Data format

Overall, the data format is consistent across the two Mult-IT subsets, which allows for a single evaluation procedure on Mult-IT. The larger size of the Mult-IT-C dataset allowed us to include additional information, including details about the quiz's administration presented in the form of quiz blocks and a multi-level topic taxonomy. This feature is absent in Mult-IT-A as questions are collected by subject without being grouped in quizzes.

Common data fields in all Mult-IT are:

- **origin**: It can be either 'C' or 'A', to discern if the question belongs to Multi-IT-A or Multi-it-C
- **question**: The question.
- **choices**: The list of possible answers.
- **answer**: The array-index corresponding to the correct answer in the choices array.

These common fields are crucial to the evaluation task, but each of the sub-portions has additional information added.

[3] https://www.alphatest.it/
[4] https://www.concorsipubblici.com/

Mult-IT-A Examples taken from Mult-IT-A are given in Figure 3.


```
{
  "origin": "A",
  "topic": "informatica",
  "question": "Le dimensioni del monitor si
       misurano in:",
  "choices": [
    "megahertz",
    "pixel",
    "centimetri",
    "pollici"
  ],
  "answer": 3
},
{
  "origin": "A",
  "topic": "psicologia e sociologia del
       disadattamento",
  "question": "Come viene definito lo stimolo
       funzionale a provocare un cambiamento?"
       ,
  "choices": [
    "Stress",
    "Output",
    "Input",
    "Matrice"
  ],
  "answer": 0
},
```


Figure 3: Data format used in Mult-IT-A.

The only additional field is **Topic**, pointing to the topic of the question. Its distribution and statistics about token count and char count are available in Figure 6.

Mult-IT-C The dataset consists of two files: quiz.jsonl contains the actual questions, while metadata.jsonl contains additional information about the questions. An example from the quiz.jsonl file is given in Figure 4.

Data fields unique of this subportion are:

- **quiz_id**: The ID of the quiz to which the question pertains.
- **question_id**: The unique identifier of the question inside the quiz. In combination with the quiz_id it forms the unique identifier of each question and can be used to retrieve the metadata of the question from the metadata.jsonl file.

Data fields of the metadata are:

- **id**: The unique identifier of the question.
- **title**: Title of the quiz sourced from the original website.
- **tags**: List of word tags.


```
{
  "quiz_id": 2250,
  "question_id": 20,
  "question": "La Costituzione riconosce allo
       Stato una potesta' legislativa
       esclusiva in materia di:\n\n",
  "choices": [
    "organizzazione della rete scolastica",
    "norme generali sull'istruzione",
    "ricerca scientifica e tecnologica",
    "istruzione professionale"
  ],
  "answer": 1
}
{
  "quiz_id": 1253,
  "question_id": 63,
  "question": "In un ingranaggio con piu'
       ruote dentate, una ruota denominata R1
       ha 25 denti e fa muovere una seconda
       ruota denominata R2 da 50 denti, che a
       sua volta fa muovere una terza ruota R3
        da 150 denti. Se la ruota dentata R3
       fa un giro e mezzo, quanti ne fa la
       ruota dentata R1?\n\n",
  "choices": [
    "3",
    "5",
    "6",
    "9",
    "12"
  ],
  "answer": 4
}
```


Figure 4: Data format used in the quiz.jsonl file of Mult-IT-C.

- **class_level1**: The first level of the topic taxonomy.
- **class_level2**: The second level of the topic taxonomy.
- **class_level3**: The third level of the topic taxonomy.
- **difficulty**: The difficulty level as estimated in the original website.
- **source**: The source of the question.

An example of an item from the metadata file is given in Figure5.

3.3. Zero-shot prompting

We evaluate our models in a zero-shot setting, thereby imitating the conditions of a real use-case scenario. The prompt we chose is designed to encourage the model to output only the letter corresponding to the answer. The original prompt, together with its English translation, is presented in Box 1.

```
{
  "id": 2250,
  "title": "Area 3 Giuridico Amministrativa
       Finanziaria - 25 domande concorso
       dirigente scolastico Miur",
  "tags": [
    "concorsi dirigenti scolastici",
    "concorso dirigente scolastico",
    "dirigente scolastico concorso dirigente
         scolastico 2017",
    "miur",
    "miur concorso",
    "miur concorso dirigente scolastico miur
         concorsi scuola",
    "concorso scuola",
    "bando concorso scuola",
    "bandi miur"
  ],
  "class_lev1": [
    "Miur",
    "dirigente scolastico"
  ],
  "source": [
    "Fgl Cgil, Miur"
  ],
  "difficulty": [
    "medio"
  ],
  "class_lev2": [
    "Istruzione",
    "Altre"
  ],
  "class_lev3": [
    "Societa' e Diritto",
    "Altro"
  ]
}
```


Figure 5: Data format used in the metadata.jsonl file of Mult-IT-C.

Prompt for the LLM

Di seguito è riportata una domanda a scelta multipla e varie possibili risposte, ciascuna indicata da una lettera. Scegli la risposta che meglio risponde alla domanda, e riporta in output soltanto la lettera corrispondente a quella risposta, senza spiegazioni.

Below is a multi-choice question together with possible answers, each indicated by a letter. Choose the best answer for the question, and report as output only the letter corresponding to that answer, without any explanation.

Box 1: Zero-shot prompt and English translation.

We decided to write the prompt in Italian in order to better represent a multilingual scenario. The prompt does not contain any information about the subject of the question or any other informative cues. In this way, our benchmark not only tests the model in question answering, but also indirectly tests the instruction-following abilities of the model in a language different than English.

Previous work on evaluating the performance of LLMs on MCQ datasets has identified two aspects which can interfere with the model's answers and therefore accuracy. One has to do with the order of possible answers: Wang et al. [12] show that the first presented option out of the possible choices tends to be preferred in the model's answer, making it quite important to take the order of possible answers into account. The other has to do with the prompt's (and even the question's) formulation: Singhal et al. [13] experiment with multiple types of prompts and also show that prompt formulation affects the model's output.

Because the position of the correct answer in the original data was already randomly distributed, which we verified with a supplementary analysis on the data (see Appendix B), performing a random permutation of the possible answers was not necessary.

3.4. Data statistics

Mult-IT-A The Mult-IT-A dataset is composed of 1,692 questions, spanning over 17 topics, all centered around knowledge required for entry exams at Italian Universities. The topics, and some additional information on the dataset composition, are provided in Table 1.

On average, questions are 83.76 characters long and contain 25 tokens counted with the *tiktoken cl100k base* [5] tokenizer or 16.5 if counted with the *Spacy* [6] library using the *it_core_news_lg model* [7].

Further statistics about quiz distribution and answer position are available in Appendix **??** and C.

It's worth noting that permutating the order of the answers would be recommended to avoid any kind of unbalance, as Multi-IT-A shows an uneven correct answer distribution leaning heavily on the first choice, acquired from the source data.

Mult-IT-C The Mult-IT-C dataset is composed of 108,773 questions divided into 4,129 quizzes.

To avoid confusion, we decided to give unequivocal names to the items of the subjects. A "quiz" is defined as a set of multiple "questions". Quizzes come from real-world examples, so they are provided with a specific name and

[5] https://github.com/openai/tiktoken
[6] https://github.com/explosion/spaCy
[7] https://github.com/explosion/spacy-models/releases/tag/it_core_news_lg-3.7.0

Category	Total	#tokens	Avg Token/Quiz
informatica	128	1997	15.602
sintassi	121	3617	29.893
grammatica	119	2429	20.412
completamento frasi	115	3034	26.383
geografia	114	1247	10.939
geometria	114	3541	31.061
ortografia	113	2273	20.115
biologia	105	4588	43.695
storia	100	1744	17.440
psicologia e sociologia del disadattamento	100	2890	28.900
elementi di criminologia	100	2386	23.860
pedagogia	100	2271	22.710
elementi di diritto costituzionale ed amministrativo	100	1813	18.130
sinonimi	98	1667	17.010
chimica	83	2983	35.940
fisica	43	2251	52.349
deduzione logica	39	1247	31.974

Table 1
Mult-IT-A: Topics included in the dataset, number of questions per topic, total tokens per topic, and average length of question per topic in terms of tokens. The topic “pedagogia" in the Table is short for “pedagogia con particolare riferimento agli interventi relativi all’osservazione e al trattamento dei detenuti e degli internati"; the topic “elementi di diritto costituzionale ed amministrativo" is short for “elementi di diritto costituzionale ed amministrativo con particolare riferimento al rapporto di pubblico impiego".

a categorisation originating from the original data source. A quiz can contain a variable number of questions. The average number of questions per single quiz is 26, and the maximum is 250. There are 1623 quizzes with more than 25 items, 298 with more than 50, and only 22 with more than 100 items.

The original categorisation made by the authors of the website "Concorsi Pubblici" was problematic for our purposes: some categories were near-duplicates of each other, containing only slightly different words. Moreover, we believed that 186 categories were too many for a meaningful visualisation and management of the data. For this reason, we created a hierarchy of three levels in which the first (bottom) level corresponds to the original categorisation of the data, then the second level groups the categorisation into 36 areas, and finally, the third level, the more abstract, has only 7 categories. The drawback of this approach, as it can be seen in the tables and graphs contained in Appendix A, is that in both the supplementary categorisations there is a significant amount of quizzes falling into the category "Other".

Nonetheless, we believe that this abstract categorisation can be good for having a general look at the data composition and thus the performance of the models in terms of macro-areas. On the other hand, keeping the original very detailed categorisation in the data allows for more in-depth analysis of model performances in specific aspects. In Appendix A, all the statistics of the 186 categories are listed in the form of a table. To appreciate the level of specificity reached by the first level of categorisation, it’s interesting to notice, as examples, categories such as "Verbs", "Diphthongs", or "Word Meanings" referring to specific language abilities. These categories are then grouped in level 2 as "Linguistic Competence" and in level 3 as "Language". As another example, we can see categories that refer to specific aspects of the Italian Public Administrations: we can see in category 1 fields such as "INPS", that is, National Institute for Social Security, or "ASL" that is "Local Health Authority".

We believe that having such a precise categorisation at our disposal is of great help in understanding the abilities and weaknesses of models in very specific aspects, thus being helpful on one hand for assessing the possibility of direct practical employment of models in Italian public administration and, on the other hand, to improve the scientific understanding of models and how they deal with different kinds of challenges. This last aspect can also be helpful for interpretability studies of LLMs.

On average, questions are 104 characters long, they contain 27.5 tokens counted with "tiktoken cl100k base" [8] or 19.8 if counted with the "Spacy" library [9] using the "it_core_news_lg" model [10]. The longest question is 1363 token long.

[8] See footnote 5
[9] See footnote 6
[10] See footnote 7

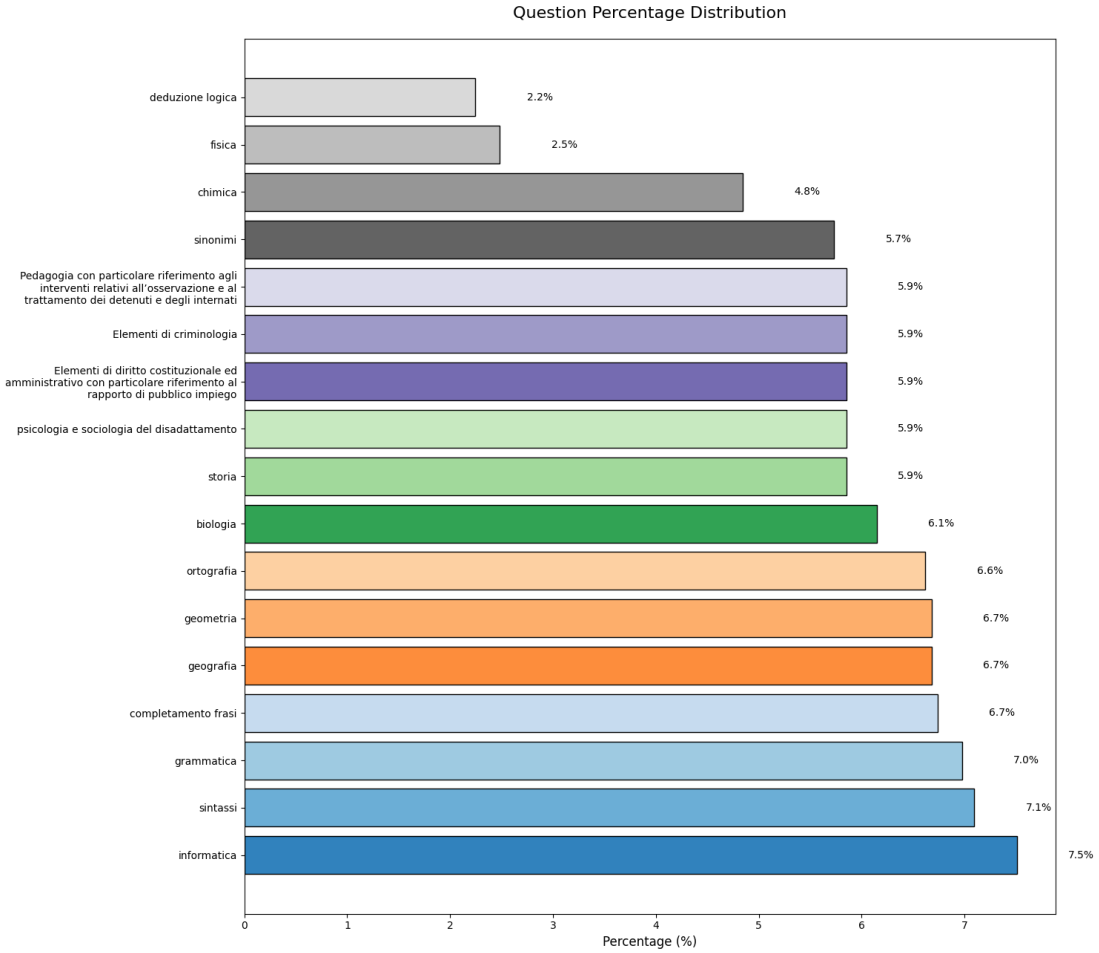

Figure 6: Mult-IT-A: Topic distribution percentage-wise.

4. Evaluation

We will use *accuracy* to evaluate the LLMs' performance on Mult-IT. Accuracy is defined as the ratio of correctly answered questions to the total number of questions, and it is a straightforward and easily interpretable measure of performance on MCQ tasks. Accuracy will be reported overall, and also separately for the two subsets Mult-IT-A and Mult-IT-C.

While accuracy is indeed a straightforward evaluation metric for this task, deciding which is the answer identified by the model as correct is not necessarily as straightforward for a couple of reasons.

As mentioned in Section 3.3, the position of the correct answer in the prompt is randomly distributed, reducing the likelihood of bias resulting from its placement, although the models might have a tendency to select the first answer more frequently.

A related issue is the fact that the model's output, in spite of the specific request in the prompt, might not always just be the letter corresponding to the chosen answer. In the case of longer outputs, simple regular expressions will be applied to extract the relevant letter.

In practice, as for all the CALAMITA challenges, the evaluation of the LLMs on Mult-IT will be carried out on the LM-evaluation-harness framework developed by EleutherAI[11].

[11] https://github.com/EleutherAI/lm-evaluation-harness

Category	Total	#Tokens	Avg Token-Quiz
Altre	31,281	911,975	29.154
Medicina	28,376	822,541	28.987
Corpo Pubblico	25,208	604,007	23.961
Giurisprudenza	15,540	482,403	31.043
Competenza Linguistica	7,142	196,610	27.529
Cultura Generale	7,111	162,300	22.824
Informatica	4,391	80,869	18.417
Logica	3,374	130,258	38.606
Farmacia	3,336	65,898	19.754
Geografia	2,886	44,707	15.491
Storia	2,150	49,945	23.23
APES	2,139	70,868	33.131
Scienze Motorie	2,066	56,278	27.24
Matematica	1,931	52,858	27.373
Lingua	1,929	36,439	18.89
Pubblica Amministrazione	1,565	50,432	32.225
Educazione civica	1,464	29,510	20.157
Letteratura	853	17,811	20.88
Biochimica	786	12,346	15.707
Chimica	784	14,037	17.904
Istruzione	745	18,528	24.87
Architettura	382	23,280	60.942
Fisica	356	8,969	25.194
Biologia	336	10,512	31.286
Economia	309	10,210	33.042
Scienze	235	3,712	15.796
Biotecnologie	185	5,741	31.032
Scienze naturali	185	2,685	14.514
Arte	180	4,419	24.55
profilo psicoattitudinale	135	5,020	37.185
Scienze della Comunicazione	90	1,787	19.856
Cucina	75	1,130	15.067
Scienze dei Beni culturali	40	502	12.55

Table 2
Level 2 of the taxonomy for Mult-IT-C

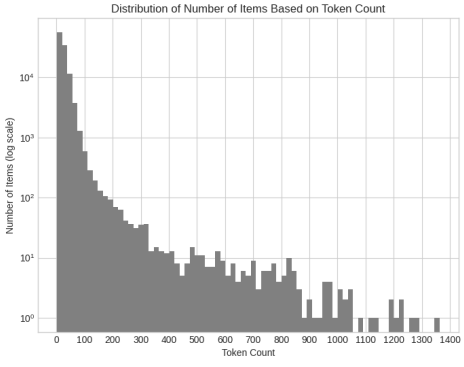

Figure 7: Distribution of number of items by token count for Mult-IT-A.

5. Limitations

The vast majority of data comes from sources linked with Italian public Institutions, and can be considered official documents. For this reason, we expect an high quality regarding the formulation of the quizzes and the correctness of the answers. Nonetheless, given the large amount of data, we cannot guarantee the absence of errors in the single questions. Human errors can happen, even in official selection, although it should be considered a rare occasion. This aspect can be improved by analysing the results obtained by the model in the benchmarks: the more the benchmark is going to be used, the more it will be possible to isolate and eventually remove or correct problematic quizzes with data analytics techniques.

Moreover, considered that the quizzes encompass almost a thirty years time span, it is possible that some quizzes, particularly the ones regarding laws, may be outdated. Nonetheless, thanks to the availability of meta-

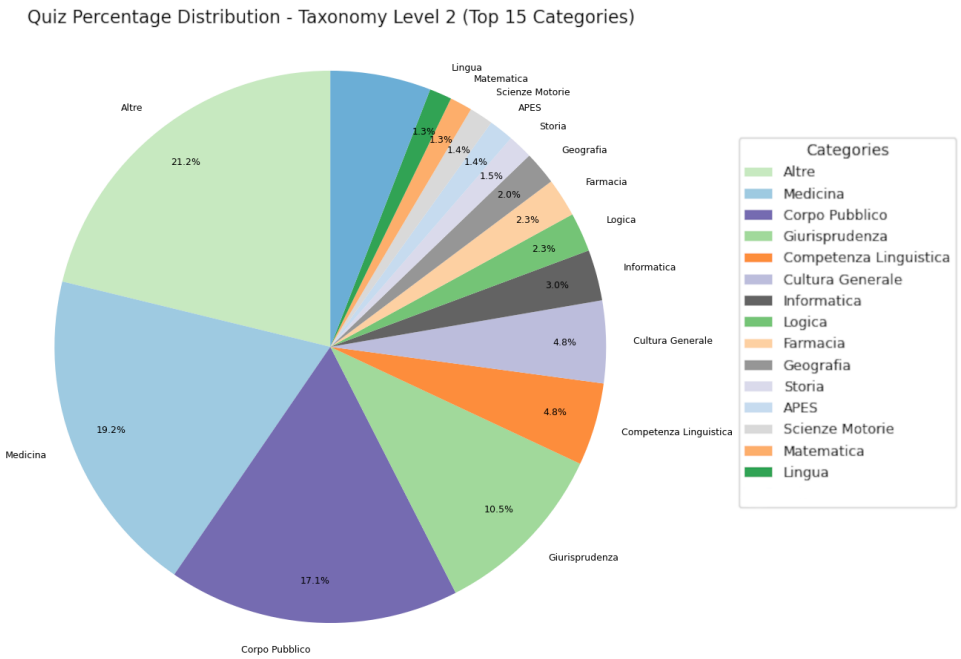

Figure 8: Quiz percentage distribution, taxonomy level 2 (top 15 categories, Mult-IT-C)

data, it is possible to further refine this dataset to also account for specific historical knowledge about laws by providing metadata to the model. However, we believe that for the first run of this evaluation this point will not create particular issues as we expect the potentially outdated questions to be limited.

Given the publicity of the data, it is possible that the original exams are already present in the models training data as they can easily obtained on the Internet. At the same time, it is likely that some sources, for example complete laws of the Italian legislation, are present in the training data, but we consider this eventuality positive given that one of the benchmark's aim is to evaluate the knowledge and capacity of the model to adapt to the Italian landscape.

6. Data license and copyright issues

Information about license and copyright issues is mandatory.

Acknowledgments

The authors would like to thank *Alpha Test* - https://www.alphatest.it, and in particular Martha Fabbri, for their interest in the Mult-IT CALAMITA challenge and for the extremely valuable exchange of ideas and data, that allowed us to shape a task of high potential impact also in the field of educational training and assessment.

References

[1] A. Srivastava, D. Kleyjo, Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, Transactions on Machine Learning Research (2023).

[2] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, et al., Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset, Advances in Neural Information Processing Systems 36 (2024).

[3] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring how models mimic human falsehoods, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Lin-

guistics, Dublin, Ireland, 2022, pp. 3214–3252. URL: https://aclanthology.org/2022.acl-long.229. doi:10.18653/v1/2022.acl-long.229.
[4] P. Wang, A. Chan, F. Ilievski, M. Chen, X. Ren, Pinto: Faithful language reasoning using prompt-generated rationales, in: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022.
[5] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: International Conference on Learning Representations, 2021.
[6] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al., Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, arXiv preprint arXiv:2406.01574 (2024).
[7] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, 2016. URL: https://arxiv.org/abs/1606.05250. arXiv:1606.05250.
[8] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: C. Ghidini, B. Magnini, A. Passerini, P. Traverso (Eds.), AI*IA 2018 – Advances in Artificial Intelligence, Springer International Publishing, Cham, 2018, pp. 389–402.
[9] I. Plaza, N. Melero, C. del Pozo, J. Conde, P. Reviriego, M. Mayor-Rocher, M. Grandury, Spanish and llm benchmarks: is mmlu lost in translation?, arXiv preprint arXiv:2406.17789 (2024).
[10] V. Lai, N. Ngo, A. Pouran Ben Veyseh, H. Man, F. Dernoncourt, T. Bui, T. Nguyen, Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 13171–13189.
[11] 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.
[12] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, Z. Sui, Large language models are not fair evaluators, 2023. arXiv:2305.17926.
[13] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al., Towards expert-level medical question answering with large language models, arXiv preprint arXiv:2305.09617 (2023).
charge [1] mmlu paper [2] https://github.com/EleutherAI/lm-evaluation-harness [3] https://huggingface.co/spaces/open-llm-leaderboard [4] https://arxiv.org/pdf/2406.01574 [5] https://www.cia.gov/the-world-factbook/about/archives/2022/countries/world/ [6] https://arxiv.org/pdf/2304.05613 [7] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html Accessed on 18/09/24 [8] https://arxiv.org/abs/2406.17789v1

7. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.

A. Appendix A: Detailed Statistics per category (Multi-IT-C)

Category	Total Quizzes	Quizzes Percentage	Total Tokens	Tokens Percentage	Avg Token-Quiz
Operatore Socio Sanitario	12434	7.39%	272403	6.03%	21.908
Arma dei Carabinieri	8179	4.86%	156358	3.46%	19.117
Carabiniere	8094	4.81%	154435	3.42%	19.08
Istruttore Amministrativo	6188	3.68%	199101	4.41%	32.175
Diritto Amministrativo	6156	3.66%	219261	4.85%	35.617
Poliziotto Municipale	5834	3.47%	160459	3.55%	27.504
Infermiere	5531	3.29%	134725	2.98%	24.358
Informatica	4011	2.38%	74362	1.65%	18.54
Guardia di Finanza	4000	2.38%	116368	2.58%	29.092
Formez	3945	2.34%	77864	1.72%	19.737
Farmacia	3336	1.98%	65898	1.46%	19.754
Agente di Polizia Municipale	3232	1.92%	94988	2.1%	29.39
Assistente Amministrativo	3147	1.87%	82803	1.83%	26.312
cultura generale	3139	1.87%	71725	1.59%	22.85
Polizia Municipale	2615	1.55%	77641	1.72%	29.691
Medicina e chirurgia	2475	1.47%	135179	2.99%	54.618
Polizia di Stato	2441	1.45%	54487	1.21%	22.322
Istruttore Amministrativo Contabile	2296	1.36%	65925	1.46%	28.713
Professioni Sanitarie	2260	1.34%	80903	1.79%	35.798
Cultura generale : Prove Concorsuali	2199	1.31%	49629	1.1%	22.569
Assistente giudiziario	2123	1.26%	77906	1.72%	36.696
Scienze Motorie e Sportive	2054	1.22%	56111	1.24%	27.318
Medico	1957	1.16%	43704	0.97%	22.332
Matematica	1731	1.03%	48665	1.08%	28.114
Cultura generale : Eserciziario	1724	1.02%	39858	0.88%	23.119
Grammatica generale	1713	1.02%	36572	0.81%	21.35
Diritto Costituzionale	1649	0.98%	33879	0.75%	20.545
Scienze infermieristiche ed ostetriche	1570	0.93%	61315	1.36%	39.054
Educazione civica	1464	0.87%	29510	0.65%	20.157
Azienda Sanitaria Locale (ASL)	1456	0.87%	41252	0.91%	28.332
INPS	1440	0.86%	48416	1.07%	33.622
Collaboratore Amministrativo	1334	0.79%	39673	0.88%	29.74
Legislazione sanitaria	1300	0.77%	27976	0.62%	21.52
Diritto del Lavoro	1284	0.76%	55342	1.22%	43.101
Logica : Ragionamento logico	1280	0.76%	79782	1.77%	62.33
Inglese	1274	0.76%	24616	0.54%	19.322
Odontoiatria e protesi dentarie	1188	0.71%	62315	1.38%	52.454
successioni di numeri e lettere	1170	0.7%	20970	0.46%	17.923
Contabilità pubblica	1139	0.68%	49706	1.1%	43.64
Significato parole	1120	0.67%	19914	0.44%	17.78
Geografia	1115	0.66%	16775	0.37%	15.045
Istruttore direttivo amministrativo	1113	0.66%	30593	0.68%	27.487
Storia	1032	0.61%	22913	0.51%	22.203
Comprensione di testi	1030	0.61%	93831	2.08%	91.098
lingua italiana	972	0.58%	18226	0.4%	18.751
Attualità	960	0.57%	20477	0.45%	21.33
Geometra	948	0.56%	32225	0.71%	33.993
Poliziotto di stato (Agente)	930	0.55%	20902	0.46%	22.475

dirigente scolastico	875	0.52%	21699	0.48%	24.799
Corpo Forestale dello Stato	838	0.5%	24954	0.55%	29.778
Sinonimi	825	0.49%	6784	0.15%	8.223
Diritto Penale	815	0.48%	19632	0.43%	24.088
Istruttore informatico	787	0.47%	17876	0.4%	22.714
Biochimica	786	0.47%	12346	0.27%	15.707
Professioni sanitarie	771	0.46%	19700	0.44%	25.551
Miur	745	0.44%	18528	0.41%	24.87
Geografia Astronomica	742	0.44%	12354	0.27%	16.65
Diritto Amministrativo (Forze dell'ordine,Poliziotto municipale)	739	0.44%	22624	0.5%	30.614
Istruttore tecnico	724	0.43%	23792	0.53%	32.862
Funzionario Amministrativo	710	0.42%	21007	0.46%	29.587
Università	670	0.4%	19715	0.44%	29.425
Contrari	645	0.38%	7018	0.16%	10.881
Chimica	644	0.38%	10589	0.23%	16.443
Legislazione sociale	640	0.38%	23527	0.52%	36.761
bibliotecario	634	0.38%	14869	0.33%	23.453
Amministrativo	618	0.37%	19132	0.42%	30.958
vigile del fuoco	610	0.36%	14755	0.33%	24.189
Corpo Nazionale dei Vigili del Fuoco	610	0.36%	14755	0.33%	24.189
Azienda Ospedaliera	600	0.36%	22730	0.5%	37.883
Diritto Comunitario	595	0.35%	12462	0.28%	20.945
Verbi	560	0.33%	9853	0.22%	17.595
Dirigente Amministrativo	545	0.32%	21679	0.48%	39.778
Medicina veterinaria	537	0.32%	29629	0.66%	55.175
Scienze dell'educazione	510	0.3%	18234	0.4%	35.753
Istruttore direttivo tecnico	503	0.3%	14508	0.32%	28.843
storia d'Italia	500	0.3%	13339	0.3%	26.678
geografia Italia	499	0.3%	6993	0.15%	14.014
Personale ATA	495	0.29%	8169	0.18%	16.503
Sostantivi	490	0.29%	7439	0.16%	15.182
Operatore ecologico	470	0.28%	15913	0.35%	33.857
Letteratura Generale	445	0.26%	8715	0.19%	19.584
Diritto privato	435	0.26%	9850	0.22%	22.644
Coadiutore amministrativo	431	0.26%	11162	0.25%	25.898
Diritto Commerciale	410	0.24%	14765	0.33%	36.012
Operaio qualificato	405	0.24%	8368	0.19%	20.662
Scienze della Riabilitazione	395	0.23%	11059	0.24%	27.997
Letteratura italiana	393	0.23%	8626	0.19%	21.949
Diritto Civile	391	0.23%	10556	0.23%	26.997
Storia contemporanea	370	0.22%	7807	0.17%	21.1
Fisica	356	0.21%	8969	0.2%	25.194
Scienze della formazione	350	0.21%	20550	0.45%	58.714
Ragionamento numerico	350	0.21%	10376	0.23%	29.646
francese	345	0.21%	6646	0.15%	19.264
Logica : Completa la frase	345	0.21%	10406	0.23%	30.162
Biologia	336	0.2%	10512	0.23%	31.286
Logica (Miscellanea)	335	0.2%	12985	0.29%	38.761
istruttore direttivo amministrativo contabile	325	0.19%	8917	0.2%	27.437
Architettura	322	0.19%	17576	0.39%	54.584
educatore asilo nido	310	0.18%	7536	0.17%	24.31
Istruttore contabile	300	0.18%	7257	0.16%	24.19

Scienze dello sport e della prestazione fisica	291	0.17%	9354	0.21%	32.144
Testo Unico Enti Locali	280	0.17%	9419	0.21%	33.639
Medicina e Chirurgia in lingua Inglese	277	0.16%	16414	0.36%	59.256
Scienze del servizio sociale	260	0.15%	7503	0.17%	28.858
Contabilità aziendale	256	0.15%	6342	0.14%	24.773
Mediatore Marittimo	244	0.14%	5302	0.12%	21.73
Magistrato	241	0.14%	10015	0.22%	41.556
Assistente sociale, Psicologo, Educatore, Sociologo	240	0.14%	5370	0.12%	22.375
Geografia fisica	240	0.14%	4458	0.1%	18.575
Coordinatore amministrativo	240	0.14%	10142	0.22%	42.258
Logica :Test delle serie	239	0.14%	6145	0.14%	25.711
Scienze	235	0.14%	3712	0.08%	15.796
Esecutore amministrativo	230	0.14%	6441	0.14%	28.004
Demografia	228	0.14%	4969	0.11%	21.794
Capacità verbale	220	0.13%	4866	0.11%	22.118
Professioni Sanitarie tecniche diagnostiche	210	0.12%	4847	0.11%	23.081
storia d'Europa	208	0.12%	5217	0.12%	25.082
Economia	200	0.12%	2972	0.07%	14.86
geografia Europa	190	0.11%	2535	0.06%	13.342
Software	190	0.11%	3557	0.08%	18.721
Unione Europea	185	0.11%	3547	0.08%	19.173
Biotecnologie	185	0.11%	5741	0.13%	31.032
Scienze e Tecnologie Viticole ed Enologiche	185	0.11%	2685	0.06%	14.514
Assistente sociale	185	0.11%	5814	0.13%	31.427
Diritto pubblico	180	0.11%	4807	0.11%	26.706
Internet	170	0.1%	2648	0.06%	15.576
Facoltà di Medicina e Chirurgia	170	0.1%	12928	0.29%	76.047
Arte	165	0.1%	4065	0.09%	24.636
Diritto internazionale	165	0.1%	2874	0.06%	17.418
Aggettivi	150	0.09%	2329	0.05%	15.527
Diritto Tributario	150	0.09%	1815	0.04%	12.1
Legislazione fiscale	148	0.09%	2432	0.05%	16.432
diritti	140	0.08%	5233	0.12%	37.379
Laurea in Chimica	140	0.08%	3448	0.08%	24.629
profilo psicoattitudinale	135	0.08%	5020	0.11%	37.185
Esperto Amministrativo	135	0.08%	4368	0.1%	32.356
spagnolo	130	0.08%	2332	0.05%	17.938
tedesco	130	0.08%	2083	0.05%	16.023
Geometria	130	0.08%	3002	0.07%	23.092
Azienda Pubblica Servizi alla Persona (ASP)	125	0.07%	3114	0.07%	24.912
Management Pubblico	125	0.07%	2016	0.04%	16.128
Assistente educativo	120	0.07%	2837	0.06%	23.642
Assistente familiare	119	0.07%	2503	0.06%	21.034
Camera di Commercio	102	0.06%	2643	0.06%	25.912
geografia Mondiale	100	0.06%	1592	0.04%	15.92
Scienze della Comunicazione	90	0.05%	1787	0.04%	19.856
Ortografia	90	0.05%	1141	0.03%	12.678
Addetto Amministrativo	90	0.05%	1609	0.04%	17.878
Collaboratore Tecnico Professionale	90	0.05%	2483	0.05%	27.589
Pronomi	85	0.05%	1640	0.04%	19.294

Hardware	80	0.05%	1232	0.03%	15.4
Assistente contabile	80	0.05%	1833	0.04%	22.913
Economia aziendale	77	0.05%	3993	0.09%	51.857
Cuoco	75	0.04%	1130	0.03%	15.067
Banca d'Italia	75	0.04%	3362	0.07%	44.827
Poliziotto di stato (Commissario)	70	0.04%	1699	0.04%	24.271
Statistica	70	0.04%	1191	0.03%	17.014
Esperto Amministrativo Contabile	69	0.04%	1531	0.03%	22.188
operatore sociale	65	0.04%	965	0.02%	14.846
Facoltà di Economia	62	0.04%	3599	0.08%	58.048
Nomi	60	0.04%	868	0.02%	14.467
Facoltà di Architettura	60	0.04%	5704	0.13%	95.067
Esperto Tecnico	60	0.04%	3251	0.07%	54.183
Ostetricia	60	0.04%	1758	0.04%	29.3
operatore tecnico	60	0.04%	1300	0.03%	21.667
autista di ambulanza	60	0.04%	1300	0.03%	21.667
istruttore direttivo socio culturale	54	0.03%	1341	0.03%	24.833
lingue straniere	50	0.03%	762	0.02%	15.24
Amministrativo giuridico	50	0.03%	1899	0.04%	37.98
tuel	50	0.03%	1363	0.03%	27.26
Curiosi,strani,imprevedibili	49	0.03%	1088	0.02%	22.204
storia Antichità	40	0.02%	669	0.01%	16.725
Testo Unico imposte sui redditi	40	0.02%	1204	0.03%	30.1
Scienze dei Beni culturali	40	0.02%	502	0.01%	12.55
Diritto regionale	40	0.02%	1235	0.03%	30.875
Poliziotto di stato	40	0.02%	828	0.02%	20.7
Sillabe	40	0.02%	730	0.02%	18.25
Avvocato	40	0.02%	1052	0.02%	26.3
Letteratura Europea	35	0.02%	775	0.02%	22.143
istruttore direttivo contabile	25	0.01%	1096	0.02%	43.84
Lauree triennali delle professioni sanitarie	20	0.01%	411	0.01%	20.55
ammissione all'università	20	0.01%	538	0.01%	26.9
Cinema e Teatro	15	0.01%	354	0.01%	23.6
Dittonghi	15	0.01%	195	0.0%	13.0
Accenti	12	0.01%	135	0.0%	11.25
Facoltà di Scienze Motorie	12	0.01%	167	0.0%	13.917
Congiunzioni	10	0.01%	130	0.0%	13.0

Table 3: Level 1 of the taxonomy, Mult-IT-C

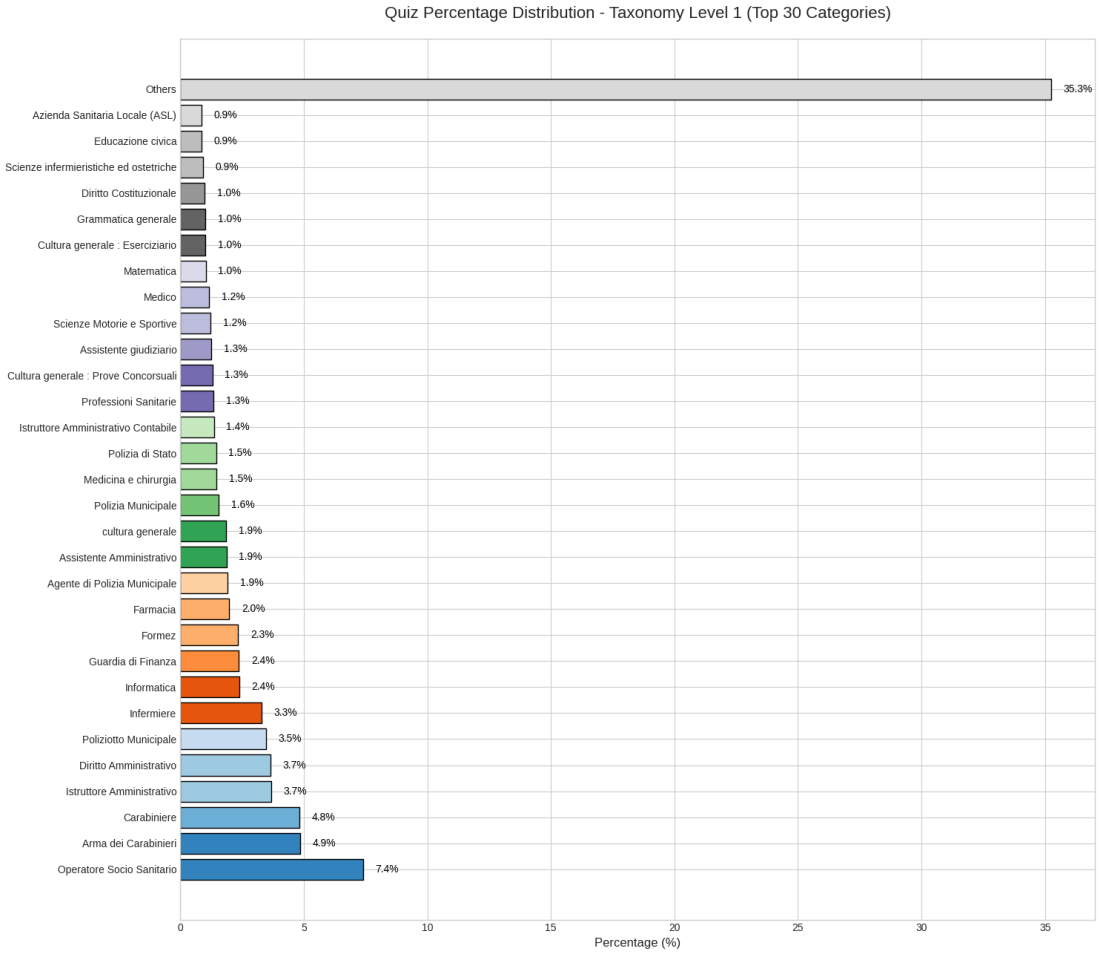

Figure 9: Quiz percentage distribution, taxonomy level 1 (top 30 categories, Mult-IT-C)

Category	Total Quizzes	Quizzes Percentage	Total Tokens	Tokens Percentage	Avg Token-Quiz
Altre Scienze e Tecniche	41006	29.03%	1092538	28.54%	26.643
Società e Diritto	39812	28.19%	1054559	27.55%	26.488
Altro	33615	23.8%	988679	25.83%	29.412
Cultura	9888	7.0%	223605	5.84%	22.614
Lingua	9071	6.42%	233049	6.09%	25.692
Matematica e Logica	5288	3.74%	182652	4.77%	34.541
Scienze MMFFNN	2559	1.81%	53028	1.39%	20.722

Table 4
Level 3 of the taxonomy, Mult-IT-C

B. Appendix B: Distribution of position of correct answer (Mult-IT-C)

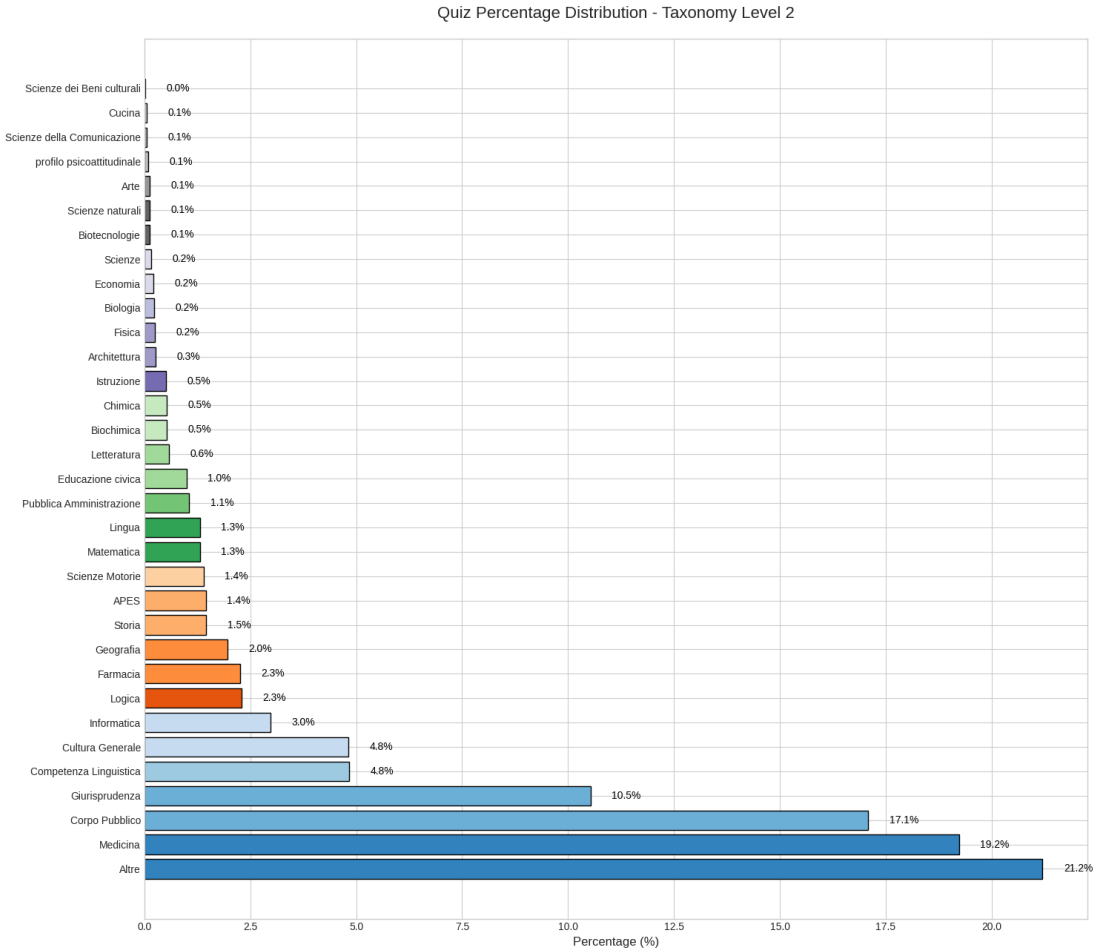

Figure 10: Quiz percentage distribution, taxonomy level 2 (all the categories, Mult-IT-C)

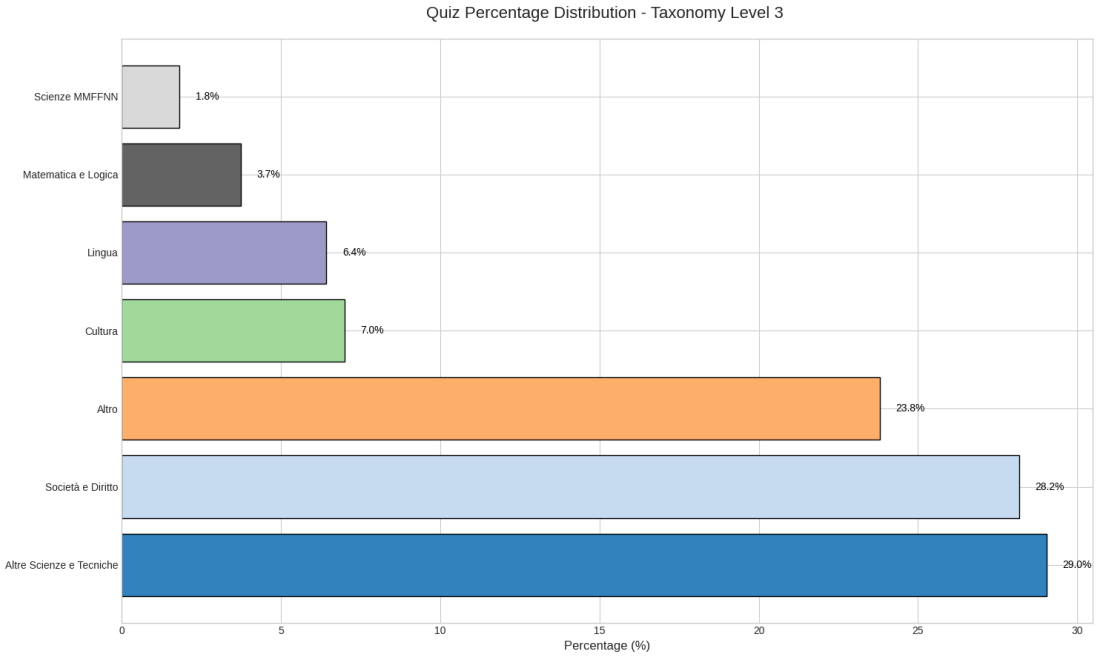

Figure 11: Quiz percentage distribution, taxonomy level 3 (top 15 categories, Mult-IT-C)

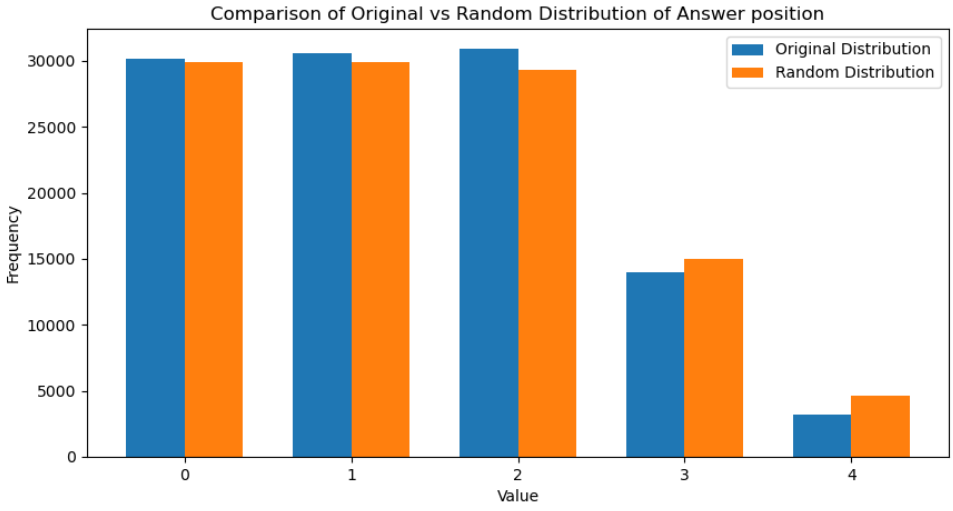

Figure 12: Distribution of answers' positions compared with a random distribution. The lower amount of items on values 3 and 4 of the x-axis is expected because only some questions have 4 or 5, respectively, possible choices

C. Appendix C: Distribution of position of correct answer (Mult-IT-A)

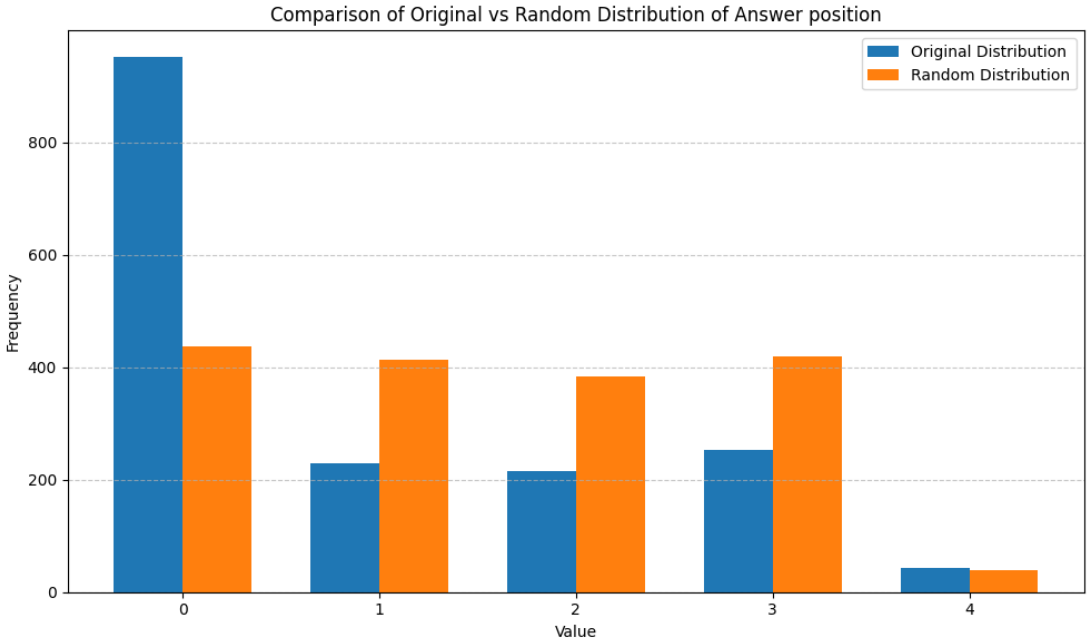

Figure 13: Mult-IT-A: Distribution of answers' positions compared with a random distribution. The lower amount of items on value 4 of the x-axis is expected because only 13.65% of the questions have 5 possible choices