ITALIC:
An Italian Culture-Aware Natural Language Benchmark

Anonymous ACL submission

Abstract

We present ITALIC, a large-scale benchmark dataset designed to evaluate the natural language understanding of the Italian language and culture. ITALIC spans 12 domains, exploiting public tests to score domain experts in real-world scenarios. We detail our data collection process, stratification techniques, and selection strategies. ITALIC provides a comprehensive assessment suite that captures commonsense reasoning and linguistic proficiency in a morphologically rich language. We establish baseline performances using 15 state-of-the-art LLMs, revealing current limitations in Italian language understanding and highlighting significant linguistic complexity and cultural specificity challenges. ITALIC serves as a benchmark for evaluating existing models and as a roadmap for future research, encouraging the development of more sophisticated and culturally aware natural language systems.

1 Introduction and Motivation

The rapid advancement of Large Language Models (LLMs) has significantly transformed the field of natural language processing, with models now exhibiting impressive capabilities across a wide array of tasks (Chang et al., 2023). As these models approach near-human-level performance in various domains, there is a rising need for robust and comprehensive evaluation methods. Assessing model performance is crucial yet challenging, as multiple key factors must be considered. These include the model's accuracy, robustness, fairness, and computational efficiency, among others (Liang et al., 2022). The creation and assessment of language models capable of operating proficiently in multiple languages worldwide continue to pose substantial difficulties for researchers (Srivastava et al., 2022). The document by the Italian government on the strategy for AI for 2024 to 2026[1] highlights the risk of cultural homogenisation when using anglo-centric LLMs. The currently available models often perform highly in English but are lacking in under-represented languages (Ruder et al., 2021). This is due to factors such as the scarce and lower quality available data (Kreutzer et al., 2022), smaller contributing communities, and Anglo-centric cultural bias in development (Talat et al., 2022).

Evaluation datasets for Italian are often insufficient, as they are translated from English rather than properly localised. Moreover, specifically, Italian datasets tend to focus on simple tasks, lacking coverage of common sense and cultural knowledge. To evaluate whether an LLM aligns with Italian culture, we propose **ITALI**an **C**ultural Benchmark (ITALIC), a novel dataset using a set of cultural questions taken from exams for public competitions in Italy. These tests assess candidates' cultural knowledge, and applying them to LLMs would provide a useful tool to identify models not aligned with Italian culture.

Standardised aptitude and knowledge assessments are used in public domains such as university admissions, military recruitment, public sector positions, and medical licensure examinations, among others. The primary aim of these tests is to evaluate applicants' domain knowledge, general knowledge, and linguistic and analytical capabilities, which are deemed essential skills that public institutions and employers prioritise by Italian national standards requirements. These assessments are crucial for a merit-based selection process across diverse fields and positions (Ruffini et al., 2023). In addition to evaluating general knowledge and analytical skills, many standardised assessments include a dedicated section focused on language and culturally relevant topics, reflecting Italy's rich cultural heritage. It is an essential criterion for public and professional roles where cultural awareness and communication skills are paramount, and candidates must

[1] Italian AI Strategy Document (2024-2026)

be aligned with the cultural norms and values embedded within the Italian professional and public spheres.

Since these culturally relevant sections present a unique challenge for language models, our study seeks to create a standard evaluation suite for advanced large language models by utilising the structure of Italian public exams as a comprehensive testing framework. The tests' structured and standardised nature makes them an excellent benchmark for comparing different LLMs with questions that are contextually relevant to Italian culture.

1.1 Contributions

The main contributions of our work are three-fold:

1. We design and publicly release ITALIC, a benchmark dataset designed to evaluate LLMs' Italian cultural and linguistic understanding across diverse domains, including history, geography, literature, and civic knowledge.

2. We comprehensively categorise the dataset into cultural and linguistic domains, enabling a clear assessment of models' performance across specific tasks.

3. We empirically evaluate and discuss 14 different model configurations, showing the current limitations of LLMs in understanding Italian culture and language and highlighting areas where future improvements in culturally aware NLP are needed.

2 Related Work

Large Language Models are deep learning models trained on vast amounts of multilingual text data, capable of understanding and generating language in a sophisticated manner (Armengol-Estapé et al., 2021; Le Scao et al., 2023). While exhibiting emerging capabilities across multiple languages such as German, French, Spanish, and Italian, multilingual models do not perform as well as in the primary training language (Touvron et al., 2023; Jiang et al., 2024). Therefore, it is essential to evaluate their performance using language-specific metrics.

Available English Benchmarks. Several benchmarks have been developed to assess language models' general intelligence and reasoning capabilities.

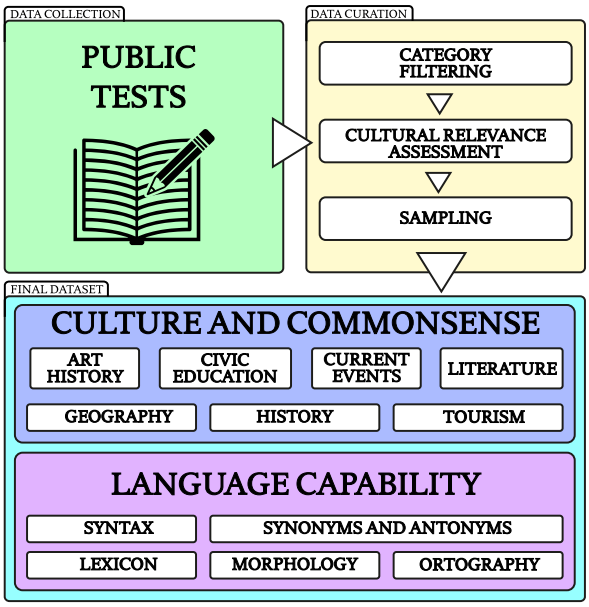

Figure 1: Overview of the data collection and curation process for the ITALIC benchmark.

Notably, MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2020) is a comprehensive benchmark that evaluates language models across 57 diverse subjects, including STEM, humanities, and social sciences. AGIEval (Zhong et al., 2023), on the other hand, is designed to assess the capabilities of models in artificial general intelligence across a broad range of tasks and skills.

Cultural Benchmarks. Cultural benchmarks have emerged as crucial tools in evaluating the adaptability and fairness of language models across different social and cultural contexts. Several studies have aimed to quantify how well these models perform across diverse demographic groups and cultural settings.

NORMAD (Norms and Adaptability of Language Models) (Rao et al., 2024) focuses on the cultural adaptability of LLMs. By utilizing 2,600 stories from 75 countries, NORMAD assesses how well these models understand and replicate social norms and etiquette as proxies for cultural comprehension. In a complementary effort, the BLEND benchmark (Myung et al., 2024) expands the scope by offering a dataset of 52.6k question-answer pairs spanning 13 languages and 16 countries/regions. This large-scale study reveals a significant imbalance in how LLMs process cultural knowledge, with models often performing better on questions related to highly represented cultures. BLEND highlights the need for more equitable training approaches that ensure underrepresented cultures

are not neglected in model performance. Lastly, CLICK (Cultural and Linguistic Intelligence in Korean) (Kim et al., 2024) provides a localised examination by testing models on their understanding of Korean culture and language. Based on 1,995 QA pairs from official exams and textbooks, the benchmark covers 11 categories related to Korean cultural knowledge, thereby addressing the linguistic and cultural nuances specific to the region.

Each benchmark focuses on a different aspect—social biases, adaptability to diverse social norms, cultural representation, or linguistic specificity—yet they all contribute to a broader understanding of the limitations and potentials of LLMs in culturally complex environments. While the above benchmarks have been instrumental in advancing the field, they primarily focus on English or specific cultural contexts. Some may include some tests in Italian but are not focused on Italian cultural commonsense and language capabilities.

Italian benchmarks. Compared to the English community, the Italian NLP community lacks the depth of original language evaluation benchmarks. (Basile et al., 2023) present a Unified Benchmark for Italian Natural Language Understanding, which includes tasks such as textual entailment, event detection and classification, factuality classification, sentiment polarity classification, irony detection, and hate speech detection. (Lai et al., 2023) propose a collaborative benchmark covering 13 tasks. Despite their contributions, both benchmarks mainly address classification-based tasks and do not examine LLM capabilities like commonsense reasoning. (Landro et al., 2022) provides an Italian benchmark for news text summarization. (Mercorio et al., 2024) introduces an adaptation of the INVALSI standardised tests for the automated evaluation of LLMs, aiming to improve the assessment of these models within diverse Italian linguistic and cultural contexts.

Our work aims to build upon these foundations by developing a specialised benchmark tailored for cultural knowledge assessment specific to Italian.

3 Dataset Curation

The dataset curation process overview is shown in Fig. 1. We have collected a corpus comprising 2,110,643 questions and answers. The data was sourced from various files in PDF, HTML, DOC, and other formats published by official bodies that announce individual competitive public examinations[2]. The corpus comprises questions and tasks from real-world exams, professional assessments, and domain-specific challenges. This ensures the dataset reflects the knowledge and skills required in Italy's various professional and academic contexts, like medical and military. Given that the data originates from institutional sources, it inherently maintains a high standard of quality and accuracy, as domain experts crafted it for public evaluations. The collected data, including the questions, answers, relevant metadata, and associated images, were normalised into JSON. Each entry follows a standard multiple-choice format with a single correct option, providing a clear structure for model evaluation.

3.1 Dataset Categorisation

We filtered out any questions that lacked answers or were incorrectly formatted, aiming to preserve only high-quality entries. We also removed any question that required reading and understanding a contextual paragraph and any question containing images. We conducted deduplication only at the question level without considering the answer options. As a result, some questions differ only in answer options, which we treat as distinct items for evaluation purposes. However, due to their similarity, these items will not appear together in the final dataset during subsequent processing steps. We applied simple preprocessing techniques to clean the dataset, such as removing question numbers and artefacts that appeared at the beginning of some questions. These artefacts included formatting characters, extraneous symbols, or other irrelevant text that could interfere with the clarity of the questions.

Using the collected metadata, the data was refined to include only questions from the dedicated cultural sections of the tests that fell into two main categories: *Culture and Commonsense* and *Language Capability*. The Culture and Commonsense category encompasses history, geography, civic education, current events, literature, art history, and tourism, reflecting Italy's cultural and social context. The Language Capability category includes tasks that assess knowledge of synonyms

[2]The dataset includes tests for admission to the Carabinieri, Penitentiary Police, Italian Army, State Police, Forestry Corps, Firefighters, Air Force, Navy, Guardia di Finanza, Italian ministries, teachers of the Italian school system of all levels, principals of the Italian school system of all levels, nurses of the national health system, and managers of the public administration from 2008 to 2024 available freely on the website of each institutional body

and antonyms, orthography, vocabulary, morphology, and syntax, which are crucial for evaluating linguistic proficiency in Italian. These subcategories assess a model's ability to understand Italian grammar rules, word formations, spelling conventions, and linguistic richness. Tasks related to Syntax determine sentence structure and grammatical correctness, while Morphology tests the model's ability to process word inflexions and conjugations, which are particularly complex in Italian. Lexicon addresses vocabulary knowledge, while Synonyms and Antonyms test semantic understanding and Orthography ensures models can accurately handle Italian spelling rules. After filtering, the dataset was reduced to 1,193,911 questions.

Assessing cultural relevance in the Italian context. To evaluate the cultural significance of the questions, we employ the LLM-as-a-Judge methodology, using LLMs to assess the quality of responses based on a given prompt. Previous work (Yuan et al., 2024) uses this technique to evaluate task difficulty, asking the LLM judge to determine the response using five additive criteria (relevance, coverage, usefulness, clarity and expertise), covering various quality aspects. We adopt this methodology, asking not one but three LLMs: GPT-4o-mini (Achiam et al., 2023), Gemini-flash-1.5 (Reid et al., 2024) and Llama 3.1 70b (Dubey et al., 2024). Our filtering criteria require that each question not be time-sensitive or dependent on current events and must be closely related to aspects of Italian culture, such as language, history, art, and other relevant cultural elements. The prompt used to rate the LLMs is:

Cultural relevance assessment prompt

You are an AI assistant. Your task is to evaluate the given question based on two criteria:
1. The question must NOT rely on current events or be something that could change over time. Example: Who is the president today?
2. The question must be related to the broad Italian culture, such as language, law, geography, history, art, or other related concepts.

Answer "True" or "False" for each criterion in JSON format.

Example JSON: { "1": "True", "2": "False" }

Now, think step-by-step, evaluate the given question based on the criteria and provide your answer.

Question: {question}
Options: {formatted options}

Based on these criteria, we retain only those items that all three models deem appropriate in every category. This filtering process removes 27.84% of questions.

Meaningfully reducing the sample size. Finally, to further reduce the size of the dataset and ensure the benchmark is not too expensive to run, we apply Farthest Point Sampling (FPS) (Eldar et al., 1997) to extract a sample of 10,000 questions. FPS iteratively selects the item farthest from the already selected items in the embedding space. This ensures the selected subset is diverse, removing items with similar questions and options, and captures a wide range of characteristics in the entire dataset. We first create word embeddings of each question in the dataset using the highest-performing model[3] on the MTEB leaderboard (Muennighoff et al., 2022). We then apply the FPS implementation proposed in (Li et al., 2022).

3.2 Dataset Description

The dataset comprises 10,000 questions, comprehensively evaluating models' abilities to understand and reason within the Italian context. Tab. 1 provides a detailed breakdown of the number of questions in each category, showcasing the distribution and variety within the dataset. Each question is formatted as a multiple-choice query with a median of 4 answer options. The minimum number of choices per question is 2, while the maximum is 5. The total number of tokens across the input data amounts to 499,963.

Category	Subcategory	#
Culture and Commonsense	Art History	980
	Civic Education	973
	Current Events	92
	Geography	979
	History	978
	Literature	984
	Tourism	980
Language Capability	Lexicon	979
	Morphology	140
	Orthography	971
	Synonyms and Antonyms	971
	Syntax	973
Total		10000

Table 1: Number of questions in ITALIC divided by *Culture and Commonsense* and *Language Capability*.

The following are two illustrative examples rep-

[3] https://huggingface.co/dunzhang/stella_en_1.5B_v5

resenting questions from the two main categories:

Example questions

Note: *every example shown is a direct translation; the original questions are in Italian. The correct option is marked by (✓).*

Culture and Commonsense

Dante imagines taking a journey through the Christian afterlife in his most famous work, the "Divine Comedy": in what year does this journey take place?
Options:
A) 1321 B) 1492 C) 1300 (✓) D) 1274

Language Capability

In the sentence "Dante was 56 years old when he died," "56 years old" is:
Options:
A) Age complement
B) Direct object complement (✓)
C) Time complement
D) Complement of abundance

3.3 Evaluation Strategy

Our approach is consistent with standard LLM evaluation frameworks like OpenAI's simple-evals[4]. We generate responses for open-source models and API-based LLMs and then compare these to the labelled correct answers. We use a decoding temperature of 0 to ensure deterministic output in this setting. Each evaluation has been run exactly once. We use zero-shot and five-shot prompts to ensure a fair assessment. We provide the models with no examples or demonstrations in the zero-shot scenario. In contrast, in the five-shot evaluation, in each question, we include in the prompt five high-quality examples that have been excluded from ITALIC during the previous filtering phase. The models are instructed to respond with only the corresponding letter (e.g., A, B, C, D) without providing any explanations or further reasoning. The evaluation is strictly based on identifying the correct answer letter, using a simple pattern-matching technique to find the first occurrence of the predicted letter in the response. The primary metric to measure performance is accuracy, reflecting the proportion of correctly answered questions.

Models employed. Our evaluation includes closed-source and open-source language models and models specifically fine-tuned for the Italian language. We include only instruction-tuned models, ensuring they are optimised for following prompts and generating responses in alignment with specific instructions. Among the closed-source models, we evaluate OpenAI's GPT-4o-mini (Achiam et al., 2023) and GPT-4o, along with Anthropic's Claude 3 Haiku and Claude 3.5 Sonnet. Additionally, we include Google's Gemini series, specifically Gemini Flash and Pro 1.5 (Reid et al., 2024). For open-source models, we assess Meta's Llama 3.1 series, including the 8b, 70b and 405b versions (Dubey et al., 2024), as well as Mistral's models, Mistral Nemo and Large (Jiang et al., 2024). To ensure a focus on Italian-specific capabilities, we also evaluate models fine-tuned for Italian, such as LLaMAntino-3 (Polignano et al., 2024) popular models Llama-3.1-8b-Ita[5] and maestrale-chat-v0.4[6], as well as the foundational model iGeniusAI's Italia 9B[7].

3.4 Manteinance

ITALIC provides a snapshot of language and cultural knowledge based on existing datasets and is designed to be robust and fully operational upon release, with no need for routine maintenance. However, as language and cultural norms evolve, periodic updates will be required to ensure the benchmark remains relevant. A new dataset version will be created and made available in such cases.

Previous versions of ITALIC will remain accessible and maintained through Zenodo, which offers a detailed version history to allow users to refer back to earlier iterations as needed. Researchers are encouraged to extend, augment, or contribute to the dataset. Those interested in making contributions can contact us with proposed additions or modifications, and after verification, the new data will be integrated into the existing dataset. A new version will then be released, ensuring all users can access the latest updates.

4 Results and Discussion

In this section, we present the performance of LLMs on the ITALIC benchmark, assessing their proficiency in Italian cultural and linguistic aspects. We evaluated the open-source models on an NVIDIA A100 80GB PCIe GPU. For this evalu-

[4] https://github.com/openai/simple-evals

[5] DeepMount00/Llama-3.1-8b-Ita

[6] mii-llm/maestrale-chat-v0.4-beta

[7] iGeniusAI/Italia-9B-Instruct-v0.1

	Culture and Commonsense							Language Capability					Total
Zero-shot Accuracy	Art (980)	Civic (973)	Eve (92)	Geo (979)	Hist (978)	Lit (984)	Tour (980)	Lex (979)	Morp (140)	Orth (971)	Syno (971)	Syn (973)	(10,000)
claude 3.5 sonnet	87.24	94.76	97.83	94.99	95.09	95.33	89.59	96.53	77.86	86.92	96.81	87.77	92.35
claude-3-haiku	76.63	82.53	84.78	85.70	85.89	83.03	78.67	90.40	67.86	71.16	87.64	72.25	81.24
gpt-4o-2024-08-06	85.71	93.11	97.83	94.79	94.79	94.11	88.98	96.83	80.00	83.42	96.91	85.92	91.36
gpt-4o-mini	76.73	82.53	88.04	88.36	84.46	82.72	78.27	90.50	70.00	71.88	93.72	74.20	82.22
gemini-pro-1.5	82.65	89.41	92.39	91.62	91.00	88.41	83.06	93.97	65.00	75.39	94.13	79.45	86.66
gemini-flash-1.5	76.33	82.73	86.96	86.62	83.74	81.10	77.45	90.81	62.14	73.53	92.17	74.41	81.66
llama-3.1-405b	86.22	90.65	96.74	93.77	94.17	93.19	86.84	94.18	62.14	78.06	95.98	78.83	88.89
llama-3.1-70b	81.12	84.58	93.48	88.56	90.90	86.08	82.04	90.91	61.43	69.52	92.58	71.94	83.61
llama-3.1-8b	66.22	69.48	77.17	77.02	74.95	64.13	69.39	74.97	40.00	49.33	70.44	50.46	66.38
mistral-large	82.96	88.39	94.57	91.32	91.82	88.72	83.37	93.67	68.57	74.36	94.75	75.33	86.30
mistral-nemo	66.33	69.27	82.61	76.00	74.34	68.39	67.76	79.37	52.86	55.61	74.25	54.78	68.53
LLaMAntino-3-8B	68.67	67.32	79.35	76.92	74.85	67.68	70.82	76.61	40.00	52.83	76.42	54.47	68.37
Llama-3.1-8b-It	70.10	71.22	82.61	79.26	77.40	67.17	71.73	81.51	52.14	53.04	81.15	53.65	70.49
maestrale-chat-v0.4	67.35	70.20	80.43	76.61	72.39	70.93	69.08	78.65	44.29	54.17	70.34	55.40	68.30
modello-italia-9b	55.20	52.00	65.22	62.82	57.16	54.47	59.18	59.14	27.14	32.13	43.46	31.35	50.53
Models Avg	75.30	79.21	86.67	84.29	82.86	79.03	77.08	85.87	58.10	65.42	84.05	66.68	77.79
Few-shot Accuracy	Art (980)	Civic (973)	Eve (92)	Geo (979)	Hist (978)	Lit (984)	Tour (980)	Lex (979)	Morp (140)	Orth (971)	Syno (971)	Syn (973)	(10,000)
claude-3.5-sonnet	89.18	96.20	96.74	96.02	96.11	95.53	91.94	98.06	81.43	88.57	97.12	89.72	93.70
claude-3-haiku	78.37	85.10	88.04	88.15	87.42	85.57	81.33	92.85	64.29	73.02	92.38	72.25	83.42
gpt-4o-2024-08-06	85.71	93.01	97.83	95.71	94.79	94.92	89.90	97.65	80.00	87.64	97.63	87.26	92.30
gpt-4o-mini	77.45	83.14	89.13	89.89	84.87	84.55	80.10	92.54	67.86	75.08	95.16	75.33	83.64
gemini-pro-1.5	84.39	91.57	94.57	93.36	93.46	91.36	86.63	96.32	63.57	79.30	96.09	84.89	89.42
gemini-flash-1.5	77.04	84.48	85.87	88.05	85.89	84.15	79.49	93.87	61.43	76.21	94.64	74.20	83.51
llama-3.1-405b	87.14	91.47	96.74	94.79	94.17	93.70	86.94	95.40	67.86	79.71	96.60	79.75	89.73
llama-3.1-70b	81.12	85.10	93.48	89.99	91.41	88.11	83.06	92.13	64.29	71.16	92.89	73.79	84.68
llama-3.1-8b	67.55	70.20	75.00	76.71	74.95	67.89	70.71	78.75	50.71	51.70	78.06	51.28	68.60
mistral-large	84.39	90.85	93.48	92.75	93.05	90.75	85.51	95.20	68.57	75.49	94.75	77.29	87.79
mistral-nemo	69.49	73.69	82.61	81.00	79.14	73.27	72.24	85.50	61.43	57.78	85.79	56.63	73.38
LLaMAntino-3-8B	70.41	69.58	82.61	76.30	74.54	68.80	70.92	80.18	48.57	52.32	81.57	54.68	69.76
Llama-3.1-8b-Ita	71.22	71.53	77.17	77.83	77.20	70.93	72.55	83.96	52.14	53.35	85.17	54.37	71.60
maestrale-chat-v0.4	66.22	71.74	82.61	75.18	72.29	71.24	70.31	79.98	49.29	53.04	77.55	54.37	69.05
modello-italia-9b	56.12	54.16	71.74	64.96	59.41	55.18	58.16	60.67	25.71	31.93	43.98	29.91	51.31
Models Avg	76.39	80.79	87.17	85.38	83.91	81.06	78.65	88.20	60.48	67.09	87.29	67.71	79.46

Table 2: Performance (accuracy %) comparison of AI models in zero-shot (top) and few-shot setting (bottom). The number of questions assessed is in parentheses next to the subcategory name. Categories are abbreviated as art history (Art), civic education (Civic), current events (Eve), geography (Geo), history (Hist), literature (Lit), tourism (Tour), lexicon (Lex), morphology (Morph), orthography (Orth), synonyms and antonyms (Syno), syntax (Synt).

ation, we utilised VLLM with its default OpenAI-compatible server, running models in bf16 format. The entire process, including setup and evaluation, took approximately 5 hours. Closed-source models and models with over 70 billion parameters were accessed from API services, costing around 100$. Tab. 2 shows the accuracy (%) of each model across several categories within the Italian cultural and linguistic context.

Model performance. Among the models tested, OpenAI's Claude 3.5 Sonnet demonstrates the highest overall accuracy across cultural and language tasks, closely followed by GPT-4o. The open-source models also perform well, with the 405b version of LLaMa standing out.

As expected, larger models tend to achieve higher scores, aligning with scaling laws (Kaplan et al., 2020) that indicate performance improves with increased training compute, dataset size and model parameters. When comparing Italian-specific models to multilingual models, we observe that although they are fine-tuned or trained on a large percentage of Italian data, they still struggle to outperform larger models trained with a greater computational budget and higher-quality multilingual data. Italian fine-tuning of models like Llama does result in a few percentage points of improvement over the multilingual baseline. However, the performance of the only foundational Italian-specific model evaluated appears unsatisfactory, suggesting room for significant enhancements in its training approach and data quality.

Qualitative study. Analysing the most frequently incorrectly answered questions by every model suggests the lack of specific linguistic knowledge and reasoning on language capabilities.

Linguistic knowledge intensive questions

A preantepenultimate word has the accent...
Options:
A) On the last syllable
B) On the second-to-last syllable
C) On the third-to-last syllable
D) On the fourth-to-last syllable (✓)

Most models consistently struggle to answer the above orthography question accurately, including variations of proparoxytone (a word with stress on the third last syllable). Models lack sufficient knowledge to handle this kind of very specific question effectively. Another type of frequently failed

question is:

Common language reasoning errors

"The emperor Tiberius chose the island of Capri as his personal residence for its beauty." In this sentence, you can find:
Options:
A) a predicative complement of the object and a complement of cause (✓)
B) a predicative complement of the subject and a complement of cause
C) a predicative complement of the object and a complement of quality
D) a predicative complement of the object and a complement of purpose

Many models struggle with performing language-based reasoning tasks, such as classic part-of-speech (POS) tagging. They also tend to falter in related tasks, such as identifying direct objects or time complements.

Artistic knowledge gaps

Which work by Lorenzo Lotto is located in the National Gallery of Ancient Art in Rome at Palazzo Barberini?
Options:
A) Mystical Marriage of Saint Catherine with Saints (✓)
B) Portrait of a Gentleman
C) Saint Jerome Penitent
D) Sacred Conversation

While analysing the most frequent wrong answers to cultural questions, we found that these focus on narrow geographical and cultural aspects. For example, the incorrect answers in the Art section were about specific details about lesser-known Italian art collections or technical elements like the architectural style of Greek temples in Italy or restoration techniques.

Geographic knowledge gaps

The highest mountain in Basilicata, standing at approximately 2000 meters, is:
Options:
A) Monte Vulture
B) Monte del Papa (✓)
C) None of the other answers

Another example of the most frequent wrong answers in the context of geographical questions requires in-depth knowledge of Italian history and human geography. Physical geography and mountain morphology are also explored, along with cultural and productive aspects such as viticulture and traditional craftsmanship. Certain questions focus on Italian geographic areas and institutions, including ports, Apennine mountain ranges, international connections, and UNESCO heritage sites.

Performance gaps within categories. Models consistently perform better in the culture and commonsense section than in language capability, likely due to the linguistic similarity of these tasks with their training data. Like many morphologically rich languages, Italian has intricate rules governing word formation, syntax, and orthographic structures. These complexities are difficult for models to grasp. Even when challenged in cultural questions, difficulties remain in nuanced tasks such as tourism, where even the best-performing models show relative weaknesses, especially in zero-shot evaluations. In particular, the substantial difference in Morphology and Orthography scores in both few-shot and zero-shot settings indicates that even the best models struggle with the subtleties of the Italian language structure, especially when examples are not provided.

One key observation from the results is the consistent difference in performance between the few-shot and zero-shot settings. The difference in the average score between the two tests is 1.27 percentage points in favour of the few-shot scenario for the culture and commonsense section and 2.12 for the language capability section (1.63 overall). In few-shot learning, models are given several task examples before making predictions. This allows the model to understand the structure, vocabulary, and nuances specific to the task or domain better. Questions in the language capability section are intertwined with language-specific rules, which may not be immediately obvious in a zero-shot setting.

4.1 Ethical Considerations

The dataset does not contain confidential information. It consists entirely of publicly available standardised test questions and does not include real data from individuals or non-public communications. Furthermore, all data and supplementary sources used in the collection process do not contain personally identifiable information or sensitive content. The dataset is also free from content that could be considered offensive, insulting, threatening, or distressing. Since it solely comprises data from standardised tests and does not involve human subjects or personal data, an ethical review process was not required. Potential risks of misuse include using the benchmark results to justify

or argue against the need to develop native LLMs specifically tailored for the Italian language. This possibility should be considered to avoid misinterpretations or unintended consequences when leveraging the evaluation outcomes.

5 Conclusion and Future Work

ITALIC represents a comprehensive and nuanced benchmark for evaluating Italian language understanding and cultural knowledge in AI models, providing diverse tasks across multiple domains and a robust tool for assessing and driving progress in Italian NLP.

Main findings. The main findings we found are:

1. Our evaluation of 15 state-of-the-art LLMs revealed that while models perform reasonably well in commonsense and culturally relevant tasks, they face significant challenges in mastering the linguistic intricacies of Italian, particularly in tasks involving morphology, syntax, and orthography.

2. We observed a consistent improvement in few-shot settings compared to zero-shot, demonstrating that providing task-specific examples greatly enhances model performance.

ITALIC highlights several areas for improvement in Italian language models, such as enhanced cultural training data to capture better Italian-specific knowledge and contexts and specialised fine-tuning for professional domains relevant to the Italian market. As the field continues to advance, ITALIC will serve as a valuable resource for researchers and practitioners working on Italian language AI, helping to ensure that these technologies can effectively serve the needs of Italian-speaking users and applications.

We envision several directions for expanding and improving ITALIC: (i) Periodic updates are needed to reflect evolving language use and cultural references in Italy, ensuring that benchmarks remain relevant for long-term evaluations. (ii) Integrating multimodal data (e.g., images, videos, audio) could challenge multimodal models. For example, combining visual art references or historical monuments with textual descriptions could improve cultural comprehension in models. (iii) Finally, future work could extend the framework used to build ITALIC to other underrepresented languages with rich cultural and linguistic diversity. This would help address the broader challenge of multilingual language models' ability to handle culturally specific tasks across different global contexts.

Resource Availability Statement. ITALIC is accessible on Zenodo[8] under the MIT license, with no IP-based or other restrictions. Additionally, the code used for the evaluation is available on GitHub[9].

6 Limitations

Scope of evaluation. The focus of ITALIC is exclusively on the Italian language and cultural context. While this is useful for evaluating models' performance in a specific language, the results may not generalise to other underrepresented languages or cultural settings.

Large-scale manual evaluation. Our dataset is based on a thoroughly validated source: standardised tests for public national assessments. Moreover, we conducted comprehensive human validation across all contexts to assess the dataset's validity. However, ideally, a human evaluation study would also require a large-scale assessment involving annotators from diverse cultural and background perspectives, ensuring an unbiased human effort.

[8] [AVAILABLE TO REVIEWERS VIA SUPPLEMENTARY MATERIALS]

[9] [AVAILABLE TO REVIEWERS VIA SUPPLEMENTARY MATERIALS]

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2021. On the multilingual capabilities of very large-scale english language models. *arXiv preprint arXiv:2108.13349*.

Valerio Basile, Livio Bioglio, Alessio Bosca, Cristina Bosco, and Viviana Patti. 2023. Uinauil: A unified benchmark for italian natural language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 348–356.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arxiv:2307.03109.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. 1997. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *arXiv preprint arXiv:2403.06412*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli, Giulia Venturi, et al. 2023. Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5):228.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jingtao Li, Jian Zhou, Yan Xiong, Xing Chen, and Chaitali Chakrabarti. 2022. An adjustable farthest point sampling method for approximately-sorted point cloud data. In *2022 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 1–6.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Fabio Mercorio, Mario Mezzanzanica, Daniele Potertì, Antonio Serino, and Andrea Seveso. 2024. Disce aut deficere: Evaluating llms proficiency on the invalsi italian benchmark. *arXiv preprint arXiv:2406.17535*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. 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.

Renato Ruffini, Marta Ingaggiati, et al. 2023. *Evoluzione dei concorsi pubblici in Italia: la valorizzazione delle competenze*, volume 33. Giuffrè Francis Lefebvre.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. 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*, pages 26–41.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.