Doctor, Is That You? Evaluating Large Language Models on Italy's Medical School Entrance Exams

Ruben Piperno[1,2,†], Agnese Bonfigli[1,2,†], Felice Dell'Orletta[2], Leandro Pecchia[1,3], Mario Merone[1,3] and Luca Bacco[1,2,*]

[1]*Research Unit of Intelligent Health-Technologies, Department of Engineering, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, 00128 Rome, Italy*

[2]*ItaliaNLP Lab, Institute of Computational Linguistics "Antonio Zampolli", National Research Council, Via Giuseppe Moruzzi 1, 56124 Pisa, Italy*

[3]*Fondazione Policlinico Universitario Campus Bio-Medico, Via Alvaro del Portillo 200, 00128 Rome, Italy*

Abstract

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities across a variety of linguistic and cognitive tasks. This study investigates whether such models can succeed in one of Europe's most selective academic assessments: the Italian medical school entrance exam. We evaluate a wide selection of open-weights LLMs, ranging from natively Italian-pretrained models to multilingual and Italian-specialised variants, on a benchmark dataset comprising over 3,300 real-world exam questions across five knowledge domains. Our experiments systematically explore the impact of language-specific pretraining, model size, prompt formulation and instruction tuning on exam performance. Results show that large multilingual models, particularly the Gemma-2-9B family, consistently outperform all other systems, surpassing the official admission threshold under all prompting settings. In contrast, models trained exclusively on Italian data fail to reach this threshold, even with larger architectures or instruction tuning. Additional analyses reveal that high-performing models display lower positional bias and greater inter-model consistency. These findings suggest that cross-domain reasoning and multilingual pretraining are key to handling multi-disciplinary educational tasks.

Keywords

Large Language Models, Italian Medical Admission Test, Instruction Tuning, Prompt Engineering, NLP in healthcare

1. Introduction

The Italian medical school entrance exam is widely regarded as one of the most competitive and demanding standardized tests in Europe. Each year, approximately 60-65,000 aspiring students face this rigorous assessment[1], which consists of 60 multiple-choice questions spanning biology, chemistry, physics, mathematics, and logical reasoning. Preparation typically begins as early as the penultimate year of high school, with students dedicating countless hours to theoretical study, targeted quizzes, and full-length simulated exams. Despite this intense effort, only a portion of students manage to be included in the national ranking: for example, in 2019 only 42.7% achieved the minimum score, while in 2020 this rose to 68.3%[2]. These figures highlight the exam's reputation as a formidable educational hurdle and a critical turning point in the academic lives of thousands of ambitious young individuals.

Against this backdrop, it is natural to ask what kind of cognitive skill set is truly necessary to succeed in such a highly selective process. Within this context, in an era increasingly shaped by Artificial Intelligence (AI), a provocative question arises:

To date, could a powerful Large Language Model (LLM), trained on vast data of human knowledge and capable of performing complex reasoning tasks, actually achieve what so many well-prepared students cannot? Could it earn a high enough score to gain admission to an Italian medical school?

LLMs represent a significant paradigm shift within Natural Language Processing (NLP), consistently demonstrating exceptional performance across diverse linguistic and cognitive tasks. Recent advancements have illustrated that these models frequently match or exceed traditional supervised methodologies and, in certain instances, surpass established human benchmarks [1, 2].

Complementary works in Italian have shown that GPT-style models can reach near-human scores on the national medical-specialty exam [3], introduced CLinkaRT

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author. Email: l.bacco@unicampus.it

†These authors contributed equally.

ruben.piperno@unicampus.it (R. Piperno); agnese.bonfigli@unicampus.it (A. Bonfigli); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); leandro.pecchia@unicampus.it (L. Pecchia); m.merone@unicampus.it (M. Merone); l.bacco@unicampus.it (L. Bacco)

0009-0007-7399-2636 (R. Piperno); 0009-0008-7092-2875 (A. Bonfigli); 0000-0003-3454-9387 (F. Dell'Orletta); 0000-0002-7900-5415 (L. Pecchia); 0000-0002-9406-2397 (M. Merone); 0000-0001-5462-2727 (L. Bacco)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]Report on 2024 Medicine test applicants

[2]Analysis of Medicine admission test scores

for clinical information extraction [4], and released native large-scale benchmarks such as INVALSI-MATE/ITA [5], Mult-IT [6] and the broader CALAMITA suite [7], laying the groundwork for systematic Italian-language evaluation.

With proven capabilities in natural language comprehension and logical reasoning, LLMs have exhibited substantial potential in educational contexts, offering instant personalized feedback, effectively summarizing intricate information, and even simulating complex human-like problem-solving processes.

However, despite their strong capabilities, previous studies have pointed out some limitations of LLMs. In particular, these models can be very sensitive to small changes in the prompt [8, 9]. One major issue is how the arrangement of elements within the prompt affects their performance, especially in tasks that require understanding and reasoning. For example, prior research has shown that LLMs are sensitive to both the specific few-shot examples provided and the order in which answer choices are presented [10, 11].

In this work, our key contribution is an in-depth analysis of how current LLMs, both Italian-specific and multilingual, perform on the multi-choice, multi-disciplinary Italian medical school entrance exam, investigating the following factors that may affect the performance:

Language-specific pre-training. We compare general multilingual models, both with multilingual pretraining and Italian specialization, and models specifically pre-trained in Italian, to assess the role of language-specific knowledge in a complex downstream task.

Model size. We evaluate models of different sizes to understand how parameter count influences performance.

Prompt design. We explore the impact of prompt formulation, including zero-shot vs. few-shot prompting, as well as the effects of prompt length and specificity.

Instruction tuning. We analyze how models that underwent instruction tuning (training on datasets designed to follow human-like task instructions) perform in comparison to base LLMs when faced with exam-style tasks.

2. Dataset

The employed corpus[3] consists of the official Italian medical school entrance exams administered in past years, collected from the public archive of the Ministry of Education, University and Research (MIUR)[4]. As such, it faithfully reproduces the exact wording, structure, and difficulty level encountered by real candidates.

[3]https://huggingface.co/datasets/room-b007/test-medicina
[4]https://www.miur.gov.it

Content and scale. The benchmark consists of $3,301$ high-quality items covering five domains (Table 1). Each item includes a question text (or stem) along with five multiple-choice answers, only one of which is correct. This structure supports two task formulations: a *classification* task, when the question is presented with the answer options, and a *generation* task, when only the question is provided and the model is expected to produce the correct answer. In our experiments, we adopt the classification setting, supplying both the question and the five candidate answers to the model.

Scoring Scheme. Each item is graded individually and then aggregated through a three-stage pipeline:

Per-Item Mark. A correct answer yields $+1.5$ points, an omission 0, and an incorrect answer -0.4. Negative marking discourages guessing and keeps the expected value of random choice below zero.

Per-Domain Average. Let s_{ij} be the mark obtained on the j-th question of domain $i \in \{\text{bio}, \text{chem}, \dots\}$ and n_i the number of items in that domain (Table 1). The mean score for the domain is

$$\overline{s}_i \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} s_{ij} \quad \in [-0.4,\, 1.5]. \tag{1}$$

Weighted Aggregation. Since domains contribute unequally to the final mark, mirroring both the weighting and question distribution of the actual exam, we adopt the official weights w_i shown in Table 1 to compute the overall average per item:

$$\overline{s} \;=\; \sum_i w_i\, \overline{s}_i \quad \in [-0.4,\, 1.5]. \tag{2}$$

Finally, the average is rescaled to the *admission-test scale* of $[-24,\, 90]$ by

$$S \;=\; 60\,\overline{s}. \tag{3}$$

Table 1
Number, distribution, and weights w_i of questions per domain, as used in Eq. (2).

Domain	# Questions	Distribution	Weight w_i
Biology	1 180	23/60	0.3833
Chemistry	1 009	15/60	0.2500
Mathematics & Physics	655	13/60	0.2167
Logic & Reasoning	212	5/60	0.0833
General Knowledge	245	4/60	0.0667
Total	3 301	60/60	1.0000

Hence a model (or a student) that answers everything correctly attains $S_{\max} = 90$, whereas one that is wrong on every question falls to $S_{\min} = -24$. Conversely, a purely random guesser (i.e., one that selects an answer

uniformly at random and is therefore correct with probability $1/5$) has an expected per-item score of

$$\bar{s} = \tfrac{1}{5} \cdot 1.5 + \tfrac{4}{5} \cdot (-0.4) = -0.02,$$

leading to an overall expected mark of

$$S_{\text{rand}} = 60 \times \bar{s} \approx -1.2$$

According to the official admission rules, only candidates who score at least **20 out of 90** are included in the national ranking. This threshold is fixed each year and represents the minimum requirement for consideration, although substantially higher scores are typically needed to secure a study place.

3. Large Language Models

Recent progress in open-weights LLMs has produced Italian-centric and Italian-specialised systems that still outperform much larger multilingual baselines on the EvalITA benchmark[5] [12]. In this study, we select from the EvalITA leaderboard the top-performing models with fewer than or equal to **9B parameters**, balancing state-of-the-art performance and computational feasibility, and we supplement them with four Italian-specialist models (DanteLLM [13], Cerbero [14], Loquace, Zefiro [15, 16]) that satisfy the same parameter budget but were not submitted to the leaderboard. This guarantees architectural diversity (LLaMA and Mistral families) while maintaining computational feasibility.

Selection Criteria Models were selected to facilitate the analysis of the factors outlined in Section1, while maintaining a constant computational budget. The selection criteria are summarised below:

Language of Pre-Training. We included (i) purely-Italian LLMs trained from scratch on Italian corpora, (ii) multilingual models that were later specialised to Italian and (iii) non-specialised multilingual models.

Model Size (Scaling). Families of LLMs offering several sizes in the 0.35 B - 9 B range, allowing us to gauge the effect of scale while holding architecture and linguistic coverage constant.

Instruction Tuning. Whenever a base and an instruction-tuned (or DPO-tuned) variant coexist, we included *both*.

Architectural Diversity. We cover the three dominant open-weights backbones available with an Italian specialisation under 9 B parameters: LLaMA / Gemma / Mistral [17, 18, 19].

Selected Models Table 2 lists every model considered in our experiments, organized by pre-training origin (Italian vs. multilingual) and instruction-tuning status. Each entry reports parameter count, original paper (if any) and the Hugging Face identifier.

This curated pool encompasses a wide range of model scales, pre-training strategies, instruction-tuning variants and backbone architectures, enabling us to rigorously evaluate how these factors affect each model's ability to tackle the Italian medical-school entrance test.

Data Leakage To the best of our knowledge, none of the questions included in the dataset were seen during the pre-training or fine-tuning of the evaluated models. The official model cards and papers explicitly exclude proprietary multiple-choice exam content, including the MIUR admission tests. While we cannot entirely rule out the possibility of indirect exposure (e.g., paraphrased content shared in online forums), we consider the risk of such leakage to be minimal.

4. Experiments

4.1. Experimental Setup

All experiments are performed on the dataset described in Section 2 and the models detailed in Section 3. No parameter is updated at any point: every model is used solely in inference mode. Unless otherwise specified in the original checkpoint, all models are queried with their default generation parameters (temperature = 1.0, top_p = 1.0, top_k = 50, repetition_penalty = 1.0); no hyperparameter tuning is performed.

Few-Shot Selection. For each topic of the dataset we randomly sample exactly two in-context demonstrations. These demonstrations are fixed once and reused across all models, prompts, and runs. In the **zero-shot** setting the demonstrations are omitted, while in the **few-shot** setting they are inserted directly into the prompt as fixed in-context examples.

Prompting Strategies. Instruction-tuned (`IT`) checkpoints are queried under two conditions:

plain — the prompt text in Table 3 is provided as a single user message, identical to the one used for base models;

chat-template — the same text is embedded in the model's native chat schema via `tokenizer.apply_chat_template`.

[5] https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard

Model	Base Architecture	Params	Instr. Tuned	Checkpoint and Reference
Non-Specialised Multilingual Models				
Gemma-2 [18]	Gemma	2 B	✗	`google/gemma-2-2b`
Gemma-2 [18]	Gemma	2 B	✓	`google/gemma-2-2b-it`
Gemma-2 [18]	Gemma	9 B	✗	`google/gemma-2-9b`
Gemma-2 [18]	Gemma	9 B	✓	`google/gemma-2-9b-it`
Multilingual Models Specialised in Italian				
DanteLLM [13]	LLaMA	7 B	✓	`rstless-research/DanteLLM-7B-Instruct-Italian-v0.1`
LLaMAntino-2 [15]	LLaMA	7 B	✓	`swap-uniba/LLaMAntino-2-7b-hf-dolly-ITA`
Cerbero [14]	Mistral	7 B	✓	`galatolo/cerbero-7b`
Loquace	Mistral	7 B	✗	`cosimoiaia/Loquace-7B`
Loquace	Mistral	7 B	✓	`cosimoiaia/Loquace-7B-Mistral`
Zefiro [15, 16]	Mistral	7 B	✓	`mii-community/zefiro-7b-dpo-ITA`
Pre-Trained Natively in Italian				
Minerva [20]	Mistral	350 M	✗	`sapienzanlp/Minerva-350M-base-v1.0`
Minerva [20]	Mistral	1 B	✗	`sapienzanlp/Minerva-1B-base-v1.0`
Minerva [20]	Mistral	3 B	✗	`sapienzanlp/Minerva-3B-base-v1.0`
Minerva [20]	Mistral	7 B	✗	`sapienzanlp/Minerva-7B-base-v1.0`
Minerva [20]	Mistral	7 B	✓	`sapienzanlp/Minerva-7B-instruct-v1.0`
Italia-9B	Mistral	9 B	✓	`iGeniusAI/Italia-9B-Instruct-v0.1`

Table 2
Overview of the LLMs considered in this work, grouped by type and listing base architecture, parameter count, instruction-tuning status, and checkpoint reference.

Hardware and Precision. All runs are executed on a single NVIDIA A100 80GB GPU, with `torch.float16` weights.

Evaluation Metrics. Model performance is assessed with four complementary metrics:

(i) Overall score S is computed by first averaging the per-item marks using the official domain weights w_i (Table 1) to obtain a weighted score $s \in [-0.4, 1.5]$, and then applying the linear rescaling $S = 60 \cdot s$, which maps the result to the standard entrance-exam range $[-24, 90]$ expressed in sixtieths, as explained in Section 2. Since our setup assumes that the model always selects an answer among the given options, we do not consider the possibility of no response. Consequently, each item is scored either $+1.5$ for a correct answer or -0.4 for an incorrect one.

(ii) Per-topic score S_t reports the same quantity computed separately for each domain (Biology, Chemistry, Mathematics & Physics, Logic, General Knowledge).

(iii) Overall Macro-averaged F_1 aggregates precision and recall uniformly across the five answer classes, making it robust to the pronounced class imbalance of the dataset, as shown in Table 1.

(iv) Per-topic macro-averaged F_1 applies the same statistic within each domain t, highlighting areas where a model may be disproportionately strong or weak despite similar global performance.

4.2. Prompt Design

The study adopts three system prompts that differ systematically in both length and semantic richness, allowing us to examine how sensitive each model is to the amount of contextual information it receives before attempting the task. The three system prompts are presented in Appendix A.

P1 is an *ultra-minimal* template that provides nothing more than the formal task instruction: the model is told that it will face a five-option multiple-choice question and must output *only* the index of the correct answer. It contains no role play, no mention of the entrance exam, and no hint about the underlying knowledge domains. This prompt therefore functions as a *lower bound* on instruction length.

P2 retains the same output constraint but introduces a concise role play: the model is asked to *"simulate a candidate who has studied intensively for the Italian medical admission test"*. This framing injects moderate priming about the exam context and about the desired mindset (efficiency and accuracy) while remaining compact.

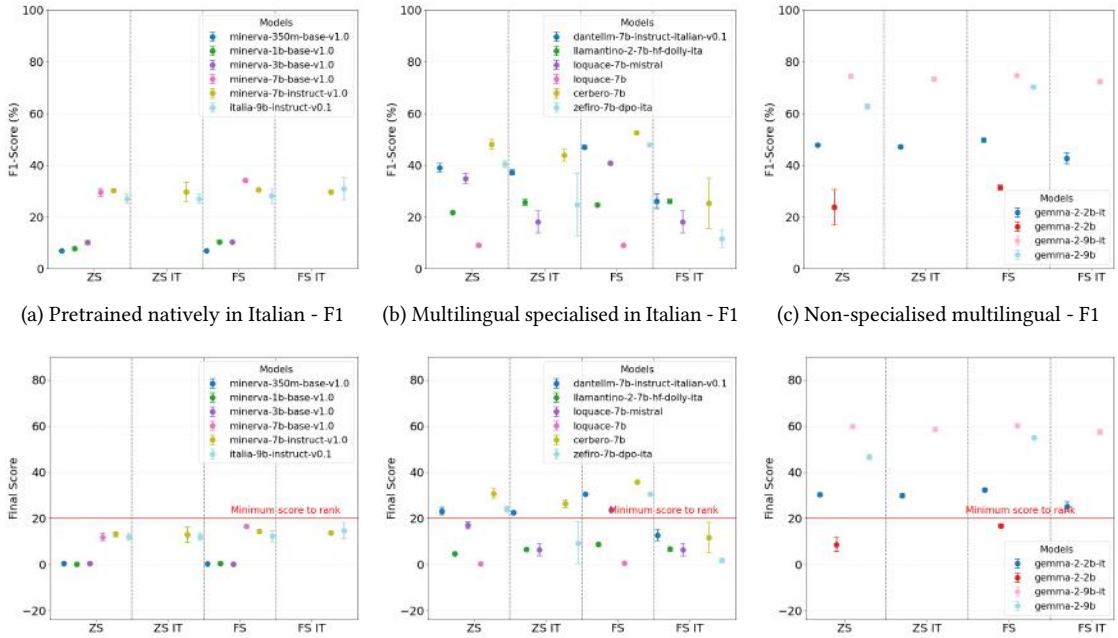
P3 is the most verbose instruction. It explicitly lists six knowledge areas (Logic, Biology, Chemistry, Mathematics, Physics, and General Culture), thereby grounding the task in the domains required by the real-world exam. The prompt also reiterates the number-only policy in boldface to maximise compliance.

Importantly, all three prompts prescribe the identical answer format: a single digit in $\{1, \ldots, 5\}$ with no accompanying text or explanation. Consequently, any variation in performance, positional bias, or inter-model agreement can be attributed to the incremental context rather than to differences in expected output style.

4.3. Qualitative Analysis

We complement the quantitative evaluation with a qualitative analysis aimed at assessing the *robustness* and *behavioural patterns* of the tested models.

First, we analyse **positional bias**, i.e., the tendency



(a) Pretrained natively in Italian - F1 (b) Multilingual specialised in Italian - F1 (c) Non-specialised multilingual - F1

(d) Pretrained natively in Italian - Final score (e) Multilingual specialised in Italian - Final score (f) Non-specialised multilingual - Final score

Figure 1: Performance comparison across model families and prompting setups. **Top row:** macro-averaged F_1 scores. **Bottom row:** final admission scores (red line = minimum threshold for national ranking). Prompting conditions: zero-shot (ZS), zero-shot with instruction formatting (ZS IT), few-shot (FS), few-shot with instruction formatting (FS IT).

of a model to overproduce certain answer indices (e.g., "1" or "3") regardless of the question. For each model and prompt, we compute how frequently each option (1-5) is selected. A uniform distribution would indicate an unbiased decision process, whereas strong deviations suggest systematic preferences unrelated to content [21].

Second, we investigate **inter-model agreement** to assess how similarly different models behave when prompted in the same way. For each prompt and setup, we compare the predicted answers across all model pairs and measure the percentage of matching responses. This reveals which models tend to converge on the same decisions and thus behave similarly, and which ones diverge more often.

Together, these two analyses provide insight into the internal consistency of each model and the structural similarity between them.

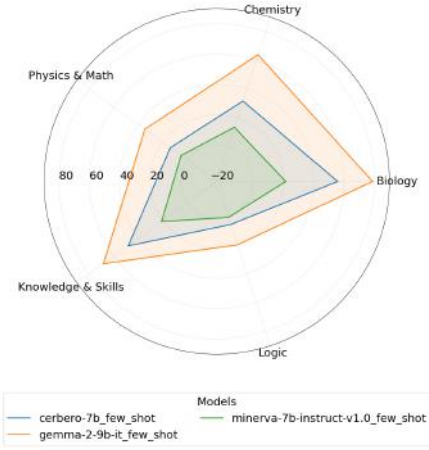


Figure 2: Final admission scores across exam disciplines (Prompt 3), comparing the best-performing model in each family under optimal prompting conditions

5. Results and Discussion

In this section, we present and analyse the performance of all evaluated models based on two key metrics: macro-averaged F1 score and final admission score (Figure 1). The reported values are computed by averaging results across three distinct prompt formulations, as we observed a high degree of consistency across prompts for both metrics.

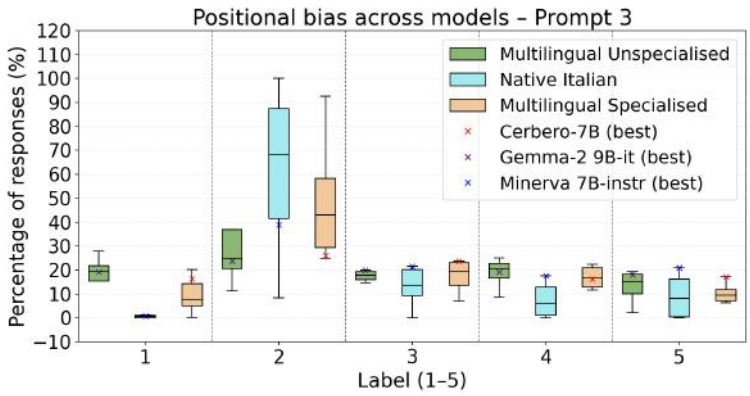


Figure 3: Distribution of selected answer positions on Prompt 3, shown separately for each model family. Crosses highlight the best model in each group.

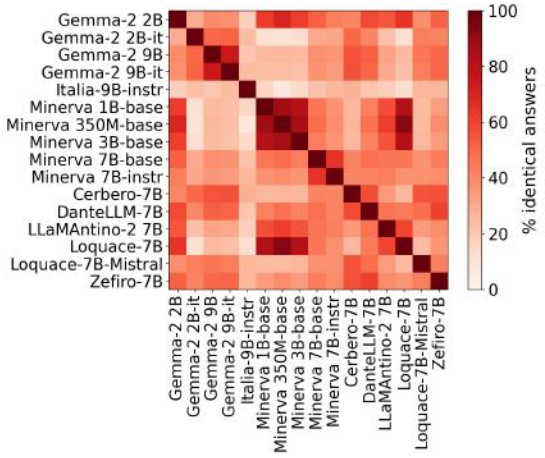


Figure 4: Pairwise answer-overlap on Prompt 3. Each cell reports the percentage of identical predictions between two models; darker shades signal stronger agreement. Models are grouped by family.

The analysis is structured around four main factors hypothesized to influence model performance: language-specific pre-training, model size, prompt design, and instruction tuning.

Language-Specific Pre-Training The results highlight a clear stratification based on language specialization. Non-specialised multilingual models, particularly gemma-2-9b-it and gemma-2-9b, consistently outperform other classes, achieving the highest F1 scores (≈74-76%) and final scores (≈58-60) across all settings. Notably, both models exceed the admission threshold of 20 in every configuration.

In contrast, natively Italian-pre-trained models, despite being trained from scratch on Italian corpora, perform significantly worse. Their F1 scores rarely exceed 33%, and none of them reach the admission threshold under any condition. Similarly, multilingual models specialised in Italian (e.g., dantellm, cerbero-7b) generally fall short of the top-performing multilingual baselines, though some (e.g., cerbero-7b) do surpass the admission threshold in specific few-shot setups. This suggests that pre-training solely on Italian data may not suffice for general-domain, multi-subject tasks like the medical entrance exam, which likely require both factual recall and cross-domain reasoning competencies that benefit from broader multilingual corpora. While multilingual models perform better, this advantage might reflect the greater scale and heterogeneity of their pretraining data, rather than the effect of multilinguality per se.

Model Size Across model groups defined by pre-training language origin, increasing model size generally correlates with improved performance, with only a few exceptions. In the Gemma series, for instance, the 9B models (Gemma-2-9b and Gemma-2-9b-it) significantly outperform their 2B counterparts, particularly in terms of F1 score. The difference is striking: Gemma-2-9b-it achieves 74% F1 in zero-shot settings, while Gemma-2-2b-it remains below 50%. This scaling effect, however, proves less predictable among models trained natively on Italian corpora or tailored to Italian. Within the Minerva family, performance increases modestly from 350M to 7B, though overall results remain limited. Moreover, Minerva-7B-instruct shows no substantial advantage over Minerva-3B-base, and Loquace-7B-Mistral underperforms relative to Cerbero-7B, despite similar model architecture and parameter count. Overall, larger models tend to perform

better, but these results suggest that size must be combined with effective training objectives and data coverage to yield consistent gains.

Prompt-Template Comparison Figure 1 reports the mean F1-score averaged across the three prompt templates; for nearly all models, the whiskers are tightly clustered, reflecting how little the specific wording shifts the central tendency. Only a few isolated exceptions emerge - e.g., Zefiro underperforms with P2, Cerbero shows higher variance in the FS IT setting, and gemma-2b displays slight sensitivity to prompt verbosity. When runs are examined separately, however, a small yet consistent ranking emerges: the minimalist **P1** systematically attains the highest scores, the verbose **P3** lands in the middle, and **P2** is invariably the weakest. Although the gap is only about 1-2 F1 points, its persistence across the entire model suite indicates that concise phrasing reduces ambiguity, whereas the intermediate framing of **P2** introduces just enough noise to dampen performance.

Prompt Design Prompt formulation plays a significant role in modulating model output. We evaluated instruction-tuned models under four prompting conditions: zero-shot (ZS), zero-shot with instruction-tuned formatting (ZS IT), few-shot (FS), and few-shot with instruction-tuned formatting (FS IT). All other non-instruction models were tested only in the ZS and FS settings.

Overall, few-shot prompting leads to improved F1 scores compared to zero-shot, particularly for mid-tier models such as DanteLLM and Cerbero, which show gains of approximately 5-10 points in F1. In contrast, high-performing models like gemma-2-9b-it achieve strong results even in zero-shot settings, indicating robustness to minimal context and reduced reliance on explicit examples.

Interestingly, zero-shot with instruction-tuned formatting often performs comparably to few-shot, especially for models with strong instruction-following capabilities. However, adding instructions to few-shot prompts does not consistently improve performance; for instance, Zefiro and Loquace exhibit a decline in F1 score compared to the few-shot setting without instructions, likely due to prompt verbosity introducing cognitive overload or disrupting the model's internal heuristics [22, 23]. These findings reinforce prior work on large language model sensitivity to prompt phrasing and structure [8, 9], and underscore the need for carefully tuned prompt engineering, particularly in lower-resource or lower-capacity models.

Instruction Tuning Instruction tuning provides consistent improvements across different model families. For example, the instruction-tuned gemma-2-2b-it outperforms its base counterpart, gemma-2-2b, by more than 20 F_1-score percentage points across all prompting conditions. Similar gains are observed for loquace-7b-mistral over the untuned loquace-7b, and for minerva-7b-instruct compared to minerva-7b-base. The impact of instruction tuning is particularly pronounced in smaller models. While the performance gap between gemma-2-9b and gemma-2-9b-it remains modest (typically around 2-3 F_1-score percentage points), tuning significantly enhances the usability of smaller variants, suggesting that instruction tuning complements model scaling and is especially valuable in resource-constrained contexts [24]. Nevertheless, instruction tuning alone is not sufficient to ensure competitive performance. Models such as zefiro-7b-dpo-ita and italia-9b-instruct, despite being instruction-tuned, still underperform relative to top-tier generalist models. This underscores the importance of tuning quality and alignment with the target domain.

Interestingly, instruction tuning appears to be most effective in the zero-shot setting, likely by helping the model better align with the intent of the prompt. However, when combined with few-shot exemplars, it can sometimes introduce redundancy or ambiguity, potentially hindering performance.

5.1. Per-Domain Performance

To complement the aggregate metrics discussed above, we conducted a topic-wise analysis of model performance, reporting final admission scores separately for each discipline in the entrance exam.

This additional evaluation aims to reveal domain-specific strengths and weaknesses that may be masked by overall scores, and to better understand how different model families handle the heterogeneous cognitive demands of the test.

For consistency, we selected the best-performing model within each family, prioritizing the few-shot setting whenever it led to superior results. The only exception is the family of non-specialised multilingual models, where the best performance was achieved in the zero-shot condition, though this setting proved competitively robust, even relative to few-shot prompting.

The selected models are:

`minerva-7b-instruct-v1.0` (natively Italian-pretrained family)

`Cerbero-7b` (Italian-tuned multilingual family)

`gemma-2-9b-it` (non-specialised multilingual family)

Given the consistency across prompts, we report results obtained with Prompt 3, which corresponds to the most verbose instruction. The results, summarized in Figure 2, show that gemma-2-9b-it achieves the highest final

admission scores across all five disciplines, with particularly strong margins in Biology and Knowledge & Skills. Cerbero-7b displays moderate performance overall but remains consistently below Gemma, with its best result also in Biology. Minerva-7b-instruct, despite instruction tuning, obtains markedly lower scores across the board, with final admission scores that remain below 40% in all subjects. The relative ranking of the models remains stable across domains, suggesting that global performance differences persist even when decomposed by topic.

Interestingly, all models achieve their highest marks in Biology and General Knowledge, two domains that largely reward factual recall, the ability to retrieve canonical facts memorised during pre-training (e.g., "mitochondria produce ATP") [25]. In sharp contrast, Mathematics & Physics and Logic & Reasoning are consistently the hardest areas, even for the best-performing Gemma checkpoint, because they demand multi-step quantitative or set-theoretic reasoning that current LLMs still struggle to perform reliably [26, 27]. Recent work further shows that simply scaling up parameters does not bridge this gap: effective reasoning requires mechanisms that disentangle memory retrieval from inference, rather than larger parametric memory alone [28].

The discipline-level analysis confirms the trends observed in the global scores, underscoring the persistent gap between non-specialized multilingual models and those trained exclusively on Italian data. These results highlight that cross-domain generalization remains a critical differentiator among models. They also reveal that even high-performing systems can display significant weaknesses in specific domains, an important consideration for real-world applications. Overall, the findings emphasize the crucial role of both model scale and pretraining diversity in developing LLMs with strong multidisciplinary capabilities.

5.2. Qualitative Analysis

Positional Bias. For every model we counted *how its answers are distributed across the five option slots*: the resulting percentages make up the box-plots in Figure 3[6].

Native-Italian Models The native-Italian models, cyan boxes, peak around 70% on option 2, and two systems select it in every single question. Such consistency betrays a positional shortcut: the model "trusts" the second slot more than the content it contains.

Italian-Specialised Multilingual Models Italian-specialised multilingual models, presented with the orange distributions, still favour label 2, but the median drops to roughly 45% and the whiskers now range from $\approx$ 25% to 90%. Extra Italian supervision therefore weakens, yet does not eliminate, the tendency to latch onto a preferred position.

General Multilingual Models General multilingual models scores, shown in green, cluster close to the 20% baseline expected from random choice, with no extreme outliers. These models appear to read the answers rather than the position, and they also lead our quantitative table, hinting at a link between genuine understanding and low positional bias.

Crosses mark the best model in each family: **Minerva 7B-instr** (blue), **Cerbero-7B** (red) and **Gemma-2 9B-it** (purple). Gemma and Cerbero stay comfortably inside their inter-quartile bands, whereas Minerva still predicts about 40% of its answers as label 2, illustrating that even the best native-Italian model has some residual bias.

Taken together, the figure draws a clear line: positional bias is most pronounced in smaller, language-specific models, softens with targeted fine-tuning, and is almost absent in large multilingual LLMs. The trend mirrors overall performance, suggesting that as models learn to solve the task they naturally stop relying on positional shortcuts. Monitoring this bias might offer a quick, model-agnostic check on whether apparent gains stem from real comprehension or from gaming the answer format. Concrete examples of typical model errors, including failures in numerical reasoning and logical minimization, are provided in Appendix B.

Inter-Model Agreement To gauge how closely the models behave, we compute for every pair the percentage of identical predictions on Prompt 3 and visualise these overlaps in Figure 4.[7]

General Overlap. Figure 4 reveals two compact blocks of high agreement. The first appears as a compact central block along the diagonal and involves the Minerva family: the four *base* checkpoints (1B, 350M, 3B, 7B) plus the instruction-tuned variant share $\geq 60\%$ identical answers, well above the $\approx 35\%$ background level observed between unrelated models, and, *in line with the positional-bias analysis*, this consensus largely reflects their tendency to pick the same (often incorrect) option. Interestingly, scaling Minerva from 350 M to 7 B parameters does little to break this uniformity: the 3 B - 7 B pair overlaps by $\approx$ 65%, only marginally higher than the 350 M - 1 B pair ($\approx$ 61%), suggesting that increased capacity amplifies the same bias instead of diversifying behaviours.

The second block, smaller but denser, occupies the upper-left portion of the diagonal and links Gemma-2 9B with its instruction-tuned sibling (Gemma-2 9B-it). Their overlap exceeds 75%; unlike Minerva, they agree mostly on *correct* answers, underscoring their stronger

[6] Shown for **Prompt 3**, the richest prompt; Prompts 1 and 2 lead to the same qualitative picture.

[7] Prompts 1 and 2 show the same qualitative pattern.

underlying capability. A size effect is evident here too: Gemma-2 2B and its instruction-tuned counterpart align at $\approx 55\%$, noticeably lower than the 9 B pair, hinting that larger multilingual backbones converge toward more stable (and more accurate) decision patterns.

Between these two extremes sit the *multilingual models specialised in Italian*, such as Cerbero-7B, DanteLLM-7B, and LLaMantino-2 7B. They form a looser band of mid-level agreement (45-60%), often acting as a bridge: they overlap moderately with Gemma while retaining some affinity with native-Italian systems. The pattern mirrors their performance table these models outperform Minerva yet trail the Gemma large pair, indicating that Italian-specific fine-tuning narrows the gap without fully matching the breadth of a high-capacity multilingual pre-training.

Outside the highlighted blocks agreement drops sharply, especially between native-Italian and general multilingual systems, supporting the idea that language-specific pre-training steers models toward distinct decision patterns.

Topic-Wise Agreement (see Appendix C). Topic-specific heat-maps paint a similar picture with nuanced shifts:

Biology and **Chemistry** closely reflect the global pattern: Minerva models cluster tightly, while Gemma leads a smaller high-accuracy duo, confirming that factual disciplines accentuate family-specific biases.

In **Logic** & **Reasoning**, the Minerva block tightens even further, with overlaps reaching $\geq 70\%$, implying that reasoning errors are strongly correlated across those checkpoints.

Mathematics & **Physics** show the widest dispersion: cross-family overlaps fall below 40% for most pairs, suggesting numerical items provoke model-specific heuristics rather than common patterns.

General Knowledge falls in between, exhibiting moderate agreement across the board.

Altogether, these observations confirm the main finding: models that share pre-training data and objectives tend to converge on the same answers while larger, broadly-trained multilingual baselines remain both accurate and mutually consistent. Model size amplifies these trends, and Italian-specialised multilingual checkpoints occupy an intermediate space, benefiting from targeted fine-tuning yet still trailing the strongest generalist pair.

6. Conclusions

Large multilingual LLMs have begun to clear the Italian medical-school admission bar, but they are still far from matching the level reached by human examinees. On the 3 301-question benchmark, the 9-billion-parameter *Gemma-2* family scored 58-60 / 90 with macro-F_1 around 75%, comfortably above the official ranking threshold of 20. A handful of Italian-tuned multilingual checkpoints (e.g. *Cerbero-7B*) also edged past the cut-off in favourable prompting conditions, whereas every natively Italian model remained well below it.

Detailed error analysis confirms that genuine reasoning remains an open challenge. Even top models stumble on Logic and on Mathematics & Physics and display residual positional shortcuts, signalling reliance on surface cues rather than deep understanding. Bridging this gap will demand progress in numerical and deductive reasoning, stronger defences against prompt variability, and tighter integration with external tools and retrieval.

In future work, we plan to extend the evaluation to a *cloze-style*, open-ended generation setting, where models must produce the correct answer without being shown the five multiple-choice options. This format may offer a more faithful assessment of their reasoning abilities and reduce positional biases. The dataset is already formatted to support this task. However, given that only a subset of LLMs currently achieves sufficient performance in the classification setting, such a shift could pose an even greater challenge. In addition, we plan to carry out a systematic exploration of decoding strategies and hyperparameters to quantify how sensitive exam performance and answer stability are to these settings. Such ablations might provide deeper insights into model robustness and optimal inference configurations.

Acknowledgments

Authors were supported by two projects: 1) the European Union under the Horizon Europe Programme through the Innovative Health Initiative Joint Undertaking (IHI JU) – Project GRACE (Project number: 101194778, Project name: bridGing gaps in caRdiAC health managEment). 2) the European Union - Next Generation EU - NRRP M6C2 - Investment 2.1 Enhancement and strengthening of biomedical research in the NHS - Project PNRR-MR1-2022-12376635 - "Early Detection of Rare Inherited Retinal Dystrophies and Cardiac Amyloidosis enhanced by Artificial Intelligence: the impact on the patient's pathway in Campania Region" (CUP: C83C22001540007)

References

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Alt-

man, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[3] S. Casola, T. Labruna, A. Lavelli, B. Magnini, Testing chatgpt for stability and reasoning: a case study using italian medical specialty tests (2023).

[4] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, Clinkart at evalita 2023: Overview of the task on linking a lab result to its test event in the clinical domain., EVALITA (2023).

[5] G. Puccetti, M. Cassese, A. Esuli, INVALSI - mathematical and language understanding in Italian: A CALAMITA challenge, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1168–1175. URL: https://aclanthology.org/2024.clicit-1.129/.

[6] M. Rinaldi, J. Gili, M. Francis, M. Goffetti, V. Patti, M. Nissim, Mult-IT multiple choice questions on multiple topics in Italian: A CALAMITA challenge, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1184–1201. URL: https://aclanthology.org/2024.clicit-1.131/.

[7] 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: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1054–1063. URL: https://aclanthology.org/2024.clicit-1.116/.

[8] Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, Calibrate before use: Improving few-shot performance of language models, in: International conference on machine learning, PMLR, 2021, pp. 12697–12706.

[9] J. Wang, Z. Liu, K. H. Park, Z. Jiang, Z. Zheng, Z. Wu, M. Chen, C. Xiao, Adversarial demonstration attacks on large language models, arXiv preprint arXiv:2305.14950 (2023).

[10] 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, arXiv preprint arXiv:2305.17926 (2023).

[11] P. Pezeshkpour, E. Hruschka, Large language models sensitivity to the order of options in multiple-choice questions, arXiv preprint arXiv:2308.11483 (2023).

[12] B. Magnini, R. Zanoli, M. Resta, M. Cimmino, P. Albano, M. Madeddu, V. Patti, Evalita-llm: Benchmarking large language models on italian, 2025. URL: https://arxiv.org/abs/2502.02289. arXiv:2502.02289.

[13] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let's push Italian LLM research forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: https://aclanthology.org/2024.lrec-main.388/.

[14] F. A. Galatolo, M. G. Cimino, Cerbero-7b: A leap forward in language-specific llms through enhanced chat corpus generation and evaluation, arXiv preprint arXiv:2311.15698 (2023).

[15] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[16] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, T. Wolf, Zephyr: Direct distillation of lm alignment, 2023. arXiv:2310.16944.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[18] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).

[19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[20] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[21] C. Zheng, H. Zhou, F. Meng, J. Zhou, M. Huang, Large language models are not robust multiple

choice selectors, arXiv preprint arXiv:2309.03882 (2023).
[22] B. Upadhayay, V. Behzadan, A. Karbasi, Cognitive overload attack: Prompt injection for long context, arXiv preprint arXiv:2410.11272 (2024).
[23] Y. Zhang, S. S. S. Das, R. Zhang, Verbosity $\neq$ veracity: Demystify verbosity compensation behavior of large language models, arXiv preprint arXiv:2411.07858 (2024).
[24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, Journal of Machine Learning Research 25 (2024) 1–53.
[25] Y. Wang, Y. Chen, W. Wen, Y. Sheng, L. Li, D. D. Zeng, Unveiling factual recall behaviors of large language models through knowledge neurons, arXiv preprint arXiv:2408.03247 (2024).
[26] M. Parmar, N. Patel, N. Varshney, M. Nakamura, M. Luo, S. Mashetty, A. Mitra, C. Baral, Logicbench: Towards systematic evaluation of logical reasoning ability of large language models, arXiv preprint arXiv:2404.15522 (2024).
[27] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, W. Yin, Large language models for mathematical reasoning: Progresses and challenges, arXiv preprint arXiv:2402.00157 (2024).
[28] M. Jin, W. Luo, S. Cheng, X. Wang, W. Hua, R. Tang, W. Y. Wang, Y. Zhang, Disentangling memory and reasoning ability in large language models, arXiv preprint arXiv:2411.13504 (2024).

A. Prompts

The study adopts three system prompts that differ systematically in both length and semantic richness, allowing us to examine how sensitive each model is to the amount of contextual information it receives before attempting the task. The three system prompts are presented in Table 3.

B. Concrete answer examples

To illustrate the kinds of mistakes made by the top-performing model (gemma-2-9B-it, prompt 3), we report two representative items: one from the Mathematics & Physics subset and one from the Logic & Reasoning subset, together with the label and the model's prediction. Each question is shown first in Italian and then in English.

Mathematics & Physics

Italian
Quanto vale il rapporto tra il volume e la superficie di un cilindro di raggio 6 cm e altezza 12 cm?
English
What is the ratio between the volume and the surface area of a cylinder with 6 cm radius and 12 cm height?
Options: (A) 2 cm (B) 1,5 cm (C) 1 cm (D) 0,5 cm (E) 4 cm
Correct answer: (A) 2 cm
gemma-2-9B-it answer: (B) 1,5 cm

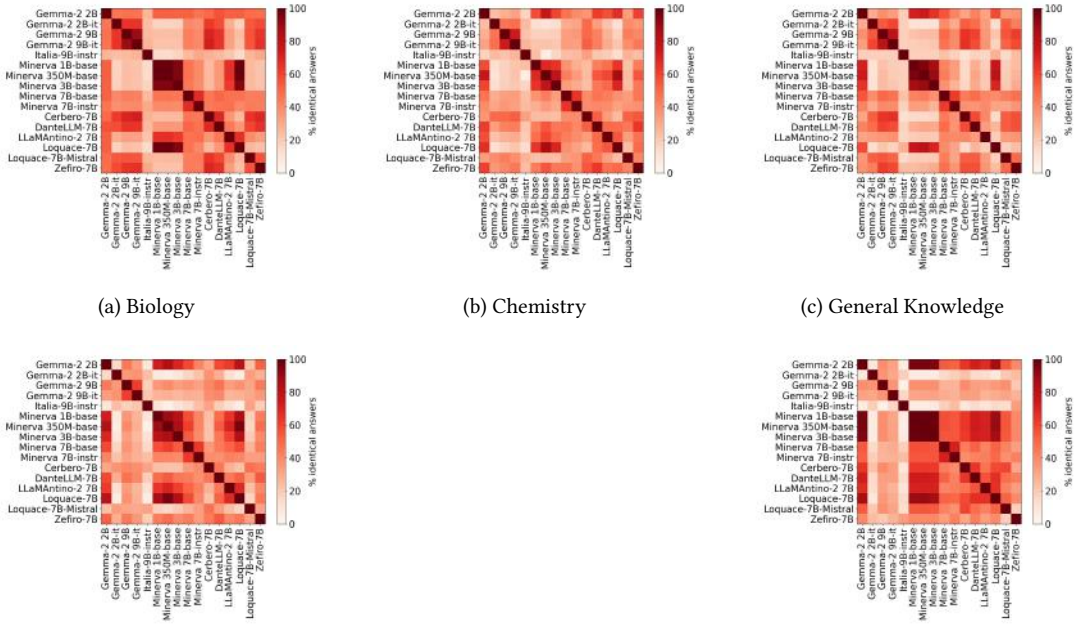
Logic & Reasoning

Italian
I partecipanti a una gara di corsa sono 150, di cui 98 maschi, 120 biondi e 90 destrorsi. Qual è il numero minimo di maschi, biondi e destrorsi che partecipano alla gara?
English
There are 150 participants in a running race: 98 are male, 120 are blond, and 90 are right-handed. What is the minimum possible number of participants who are simultaneously male, blond, and right-handed?
Options: (A) 8 (B) 10 (C) 20 (D) 12 (E) 18
Correct answer: (A) 8
gemma-2-9B-it answer: (B) 10

These snapshots show two common failure modes: a numerical miscalculation in the geometry element and an overestimation in the logical minimization problem, both of which may be influenced by the model's observed positional bias towards option (B) (as shown in Figure 3).

Table 3
System prompts used in the study.

Italian (original)	English (translation)
P1. Domanda con 5 risposte (0-4). Scrivi solo il numero della risposta corretta. Nessun altro output.	**P1.** Question with 5 answers (0-4). Write only the number of the correct answer. No other output.
P2. Simula il comportamento di un candidato che ha studiato intensamente per il test di Medicina in Italia. Per ciascuna domanda a risposta multipla con opzioni da 0 a 4, indica solo il numero della risposta che ritieni corretta. Non fornire alcuna motivazione. Concentrati sull'efficienza e sulla precisione.	**P2.** Simulate the behaviour of a candidate who has studied extensively for the Italian Medical School admission test. For each multiple-choice question with options 0-4, output only the number of the option you believe is correct. Provide no justification. Focus on efficiency and accuracy.
P3. Sei uno studente italiano con solide competenze in Logica e Ragionamento, Biologia, Chimica, Matematica, Fisica e Cultura Generale. Stai sostenendo il test di ingresso alla facoltà di Medicina. Ti verrà fornita una domanda a risposta multipla seguita da cinque opzioni numerate da 0 a 4. Il tuo compito è indicare **esclusivamente** il numero (0-4) corrispondente all'alternativa corretta. Non fornire spiegazioni.	**P3.** You are an Italian student with strong skills in Logic and Reasoning, Biology, Chemistry, Mathematics, Physics, and General Culture. You are taking the entrance exam for the Faculty of Medicine. You will be given a multiple-choice question followed by five options numbered 0 to 4. Your task is to output **only** the number (0-4) corresponding to the correct option. Do not provide any explanation.



(a) Biology (b) Chemistry (c) General Knowledge

(d) Mathematics & Physics (e) Logic & Reasoning

Figure 5: Pairwise answer-overlap heat-maps for the five exam domains. Each cell reports the percentage of identical predictions between two models when evaluated only on the subset of questions belonging to the indicated topic (Prompt 3 setting).

C. Heat-maps of Model Agreement

Figure 5 shows per-domain heatmaps of model agreement. Each cell reports the percentage of identical predictions on a given topic. The same trends seen in Figure 4 persist: (i) Minerva checkpoints are tightly aligned, mostly on wrong answers; (ii) Gemma-2 9B models remain the most consistent and accurate pair; (iii) unrelated models rarely exceed 40% overlap. Still, domain-specific traits emerge: Logic & Reasoning shows high Minerva coherence ($\geq 70\%$), suggesting shared shortcuts; Math & Physics shows the lowest cross-family overlap, likely due to numerical complexity. These results confirm that agreement varies by domain and should be interpreted accordingly.