A gentle push funziona benissimo: making instructed models in Italian via contrastive activation steering

Daniel Scalena[1,2,*], Elisabetta Fersini[1] and Malvina Nissim[2]

[1]University of Milano - Bicocca, Italy
[2]University of Groningen, CLCG, The Netherlands

Abstract
Adapting models to a language that was only partially present in the pre-training data requires fine-tuning, which is expensive in terms of both data and computational resources. As an alternative to fine-tuning, we explore the potential of activation steering-based techniques to enhance model performance on Italian tasks. Through our experiments we show that Italian steering (i) can be successfully applied to different models, (ii) achieves performances comparable to, or even better than, fine-tuned models for Italian, and (iii) yields higher quality and consistency in Italian generations. We also discuss the utility of steering and fine-tuning in the contemporary LLM landscape where models are anyway getting high Italian performances even if not explicitly trained in this language.

Keywords
Italian steering, Language adaptation, Activation steering, Instruction Tuning, Reasoning benchmarks

1. Introduction

The strong rise in capabilities of the latest large language models (LLMs) has brought significant improvements in a wide variety of downstream tasks. These abilities mainly derive from the instruction-tuning procedure (IT), i.e., model fine-tuning on instruction datasets, and enable the models to follow user-prompted instructions.

Most LLMs, however, are mainly pre-trained and fine-tuned in English, and while other high-resource languages are included in the training data, they are not present to the extent needed to achieve out-of-the-box performances comparable to English. A strategy to address this has been, in the past few years, to fine-tune models with language-specific instructions, such as the Stanford Alpaca dataset [1], which has been automatically translated in multiple languages – the Italian version of it has been used to train the Llama 2-based Camoscio model [2]. A combination of $\sim 240K$ training instances from three automatically translated instruction datasets was used to train the latest Llamantino [3], the most recent Llama 3-based instruction-tuned model for Italian.

This approach has proven effective, but using large amounts of machine-translated texts is far from optimal: although the translation is generally good for high-resource languages, the language's unique linguistic and cultural aspects are often not represented by the training data. In addition, one must consider the usual substantial (computational) costs associated with large datasets.

With recent developments in interpretability research, new approaches are arising to localize and steer different language model aspects. These techniques mainly work with an inference-time injection, allowing for targeted interventions during the generation phase without incurring the high costs associated with any additional training. Such techniques, relying on the assumption that models are already capable of performing specific tasks, aim at enhancing some of the internal activations leading to specific solutions, thereby also increasing overall performance. They have proved successful towards specific tasks, such as model detoxification, but also toward more generalist and wide-ranging tasks [4, 5].

We explore the potential of *steering* for Italian-instructing a pre-trained LLM as an alternative to fine-tuning, adopting a steering technique based on contrastive examples. We observe that this approach, with much less data ($\ll 100$ instances instead of 240K) and no additional training required, enables performances comparable to standard fine-tuning approaches and yields high-quality Italian generations.

2. Related works

The latest LLMs are pre-trained on data which often includes not only English but also (small percentages of) other languages [6, 7]. After the initial pre-training phase, models are further trained to follow instructions given by users. Due to the nature of most instruction-tuning data, performance in and on English is still overwhelmingly better than for other languages [8].

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
d.scalena@campus.unimib.it (D. Scalena); elisabetta.fersini@unimib.it (E. Fersini); m.nissim@rug.nl (M. Nissim)
https://danielsc4.it/ (D. Scalena)
0009-0006-0518-6504 (D. Scalena); 0000-0002-8987-100X (E. Fersini)

Italian adaptation Over time the most widely adopted solution to improve model performance over the Italian language has been to perform further Instruction-Tuning with Italian data (IT-ITA) on existing models. Examples of this type are Camoscio [2] and Llamantino 2 [3] (both based on the Llama 2 model's family), and ANITA [9] (based on Llama 3 models). Generally, instruction fine-tuning is performed on the original model already in its instructed version using additional data which is machine-translated from instructions originally in English. Taking ANITA as an example this goes as follows: starting from the instructed Llama 3, fine-tuning is performed with $\sim$ 100k instruction prompts in English and, after an additional optimization step with $\sim$ 40k examples, another 100k prompts machine-translated into Italian are used for the language adaptation task. This large amount of data, combined with the size of the models, naturally leads to large computational costs.

Steering vectors Following the linear representation hypothesis, high-level concepts are represented as directions in the activation space of LLMs [10]. A single direction can be found through the use of examples designed to elicit opposite behaviors in output to the model [5, 4, 11] or by using the difference between fine-tuned models for specific tasks and their original version [12]. The effectiveness of these techniques lies in isolating specific properties, such as the language or the style used, to emphasize it during inference. In this work, we test the potential of steering vectors to improve performance on several NLP tasks by facilitating the process of generating the Italian language for which the models were not originally explicitly trained.

3. Method

We build on the assumption that during the training process, the model already sees a small amount of the target language (Italian in our case). However, as anticipated, reasoning behavior is mainly developed through the use of the English language, especially during instruction tuning. We aim to push the internal components promoting the language switch, so as to achieve better results on a language different than English.

Steering through contrastive prompts The first step to extract the Italian steering vector is to build *contrastive prompts* that will highlight the differences between the activations when prompting the model with different languages [4, 5]. To this end, we use the Stanford Alpaca dataset [1], consisting of question-answering style prompts, both in its original English and its machine-translated Italian version (Appendix A shows some random example instances.)

We edit the original Alpaca dataset and obtain three different versions:

- **ENG**: the original dataset, both question and answer are in English;
- **ITA-full**: machine-translated Alpaca dataset, both question and answer are in Italian;
- **ITA**: questions in English, answers in Italian. The aim is to emphasize the language switch task, pushing the model to respond in Italian even to an English prompt.

By using contrastive examples between the original English and the Italian responses we extract the difference in activations between the models prompted in different languages.

Steering vector extraction At every generation step $i = 1, \ldots, M$ a LLM f generates a sequence of tokens based on the prompt p_{version} and previously generated tokens $y_1, \ldots, y_{i-1}$. We collect the activations of the last token from each attention head output ($f^{l,h} \in \mathbb{R}^{d_{\text{head}}}$)[1] and average them over a series of $K = 30$ prompts.

$$a_i^{\text{version}} = \frac{1}{K} \sum_{k=1}^{K} f^{l,h}(p_{\text{version}}^k, y_{<i}) \tag{1}$$

where $a_i^{\text{version}} \in \mathbb{R}^{|L| \times |H| \times d_{\text{head}}}$. The prompts p_{version} are supposed to push the model towards the desired behavior using a 5-shot setting and an instruction explicitly asking the model to respond in a specific language (either Italian for ITA and ITA-full or English for ENG; further details are in Appendix A).

To obtain the final steering vector towards the ITA or ITA-full behavior we compute the difference between the previously calculated activations as follows:

$$\Delta_i^{\text{ITA-full}} = a_i^{\text{ITA-full}} - a_i^{\text{ENG}}$$
$$\Delta_i^{\text{ITA}} = a_i^{\text{ITA}} - a_i^{\text{ENG}}$$

Steering vector injection The newly calculated steering vector, when added to the running activations, is supposed to steer the model toward a specific direction, in a similar fashion to what was common with word embeddings in vector space [13]. We apply each steering vector for every generated token using a diminishing multiplicative factor $\alpha = 1.5$ to modulate the steering intensity following what was proposed to be effective in [4]:

[1] The extraction is made on every layer $l \in L$ and for each attention head $h \in H$ where L and H are the total number of layers and attention heads in the LLM respectively.

$$f_i^{l,h}(\cdot) \leftarrow f_i^{l,h}(\cdot) + \alpha\Delta_i^{l,h} \quad (2)$$

where α regulates the steering intensity, starting with val_{max} and linearly diminishing to 0 for each i-th generated token:

$$\alpha_i = \text{val}_{\text{max}} \cdot \left(1 - \frac{i-1}{M-1}\right) \quad (3)$$

where M indicates the maximum number of tokens to be generated.

This allows us to get the language direction coming from the difference in polarity between the activations, eventually steering the original LLM towards Italian.

4. Results

We select two different models as base to test the effectiveness of our steering approach. The first is the smallest (8B parameters) from the Llama 3 family in its Instructed version[2]. The second model we take as base is the smallest (3.8B parameters) Phi 3 model[3] in its English-instructed version. For a comparison of steering with the more commonly-used Instruction Tuning approach, we also re-run on the selected benchmarks the latest Instruction Tuned model with Italian data (IT-ITA) model ANITA from [9], also based on the same Llama 3 model we use.

Since all of these models have some training data in different languages, even if not specifically meant to be multilingual, we also test the original models on the Italian benchmarks to get a baseline in terms of model capabilities and better capture the differences between the IT-ITA procedure and the different steering techniques.[4]

4.1. Selected benchmarks

We test the models on three different standard benchmarks included in the Italian LLM leaderboard[5]:

- **MMLU** [15] is a multitask question-answering benchmark consisting of multiple-choice questions from various expert-level knowledge branches. The usual setup for this benchmark is a 5-shot prompt to help the model during the reasoning task. The test set consists of $\sim$ 14k instances with four possible responses each.
- **HellaSwag** [16] is a benchmark meant to measure grounded commonsense inference. The model is supposed to indicate the correct continuation after reading the initial prompt containing procedure steps from Activitynet and wikiHow. The employed setting is a 0-shot prompt over all the $\sim$ 10k test instances.
- **ARC challenge** [17] is a collection of over 1k instances of school-level multiple-choice science questions aimed at measuring the knowledge retrieval capabilities of a LLM. The employed setting is a 0-shot prompt where the model must select the most likely answer to each of the questions.

We also test the ability of the model in generating full Italian responses (rather than non-Italian ones). To this end, we use a popular language identification tool `lang-detect`[6] and take the probability of the Italian language as the scoring metric.

4.2. Steering vs the rest

General results Table 1 shows the models' results for each benchmark.[7] Among the two proposed steering approaches, ITA generally proves to be more effective in steering the LLM outputs. Additionally, the steering approach often surpasses both the original and IT-ITA models' performances. The most significant advantage, however, is the **reduced time and computational resources needed to enhance a model's performance in a new language**. The Italian Llama 3 ANITA [9] typically outperforms its original version but has required fine-tuning on over 240k examples. In contrast, the steering technique achieves comparable or better performance across most benchmarks with significantly less data — only 30 demonstrative examples in our case.

Approaches matter It may be useful to look at how steering and Instruction Tuning techniques differ in improving model responses. Figure 1 shows the overlap (or lack thereof) of correct responses of the four approaches based on Llama 3-Instruct. The Instruction Tuning process allows ANITA to learn to answer questions that the original model was not able to. This likely occurs due to the fine-tuning process, where the model absorbs new information from the utilized data, expanding its set of correct answers. At the same time, however, IT-ITA also runs into the loss of previous capabilities on some ques-

[2] meta-llama/Meta-Llama-3-8B-Instruct via HuggingFace
[3] microsoft/Phi-3-mini-4k-instruct via HuggingFace
[4] Another obvious baseline would be a native Italian model, such as the recent Minerva [14] which is pre-trained on Italian+English data. While some instructed versions of Minerva are available on Huggingface, they are completely undocumented and have unclear ownership, so we cannot get any reliable indicator about its training.
[5] Open ITA LLM leaderboard via HuggingFace.

[6] `lang-detect` package
[7] Please note that our results differ from those shown in the Italian LLM leaderboard since we employ a regex-based approach to evaluate the responses instead of using the response likelihood of the model as per [18], which would require four times more runs. This is further explained in Appendix B.
[8] For the sake of clarity, only cardinalities $>$ 25 are shown in writing.

Model	MMLU (it)	HellaSwag (it)	ARC challenge (it)	`lang-detect` (it)
Meta Llama 3 8B - Instruct				
Original	54.21	**52.30**	71.31	.995
+ IT-ITA (ANITA [9])	55.01	42.49	**72.54**	.715
+ Steering ITA-full	55.73	48.74	70.82	**.999**
+ Steering ITA	**55.95**	50.00	71.38	.996
Microsoft Phi 3 mini 4k - Instruct				
Original	59.65	60.02	69.37	.997
+ Steering ITA-full	59.92	54.36	**74.42**	**.999**
+ Steering ITA	**60.65**	**60.14**	74.25	**.999**

Table 1
Results on the benchmarks in % of correct answers. In column `lang-detect` we also evaluate the language used in answering the questions by reporting the average score of Italian responses. Generally, the steered models (especially the ITA approach) result in a slight improvement compared to the original model and to outperform ANITA on two of the three benchmarks. Significant improvements are seen in the language itself, where the steering techniques are effective in yielding Italian output.

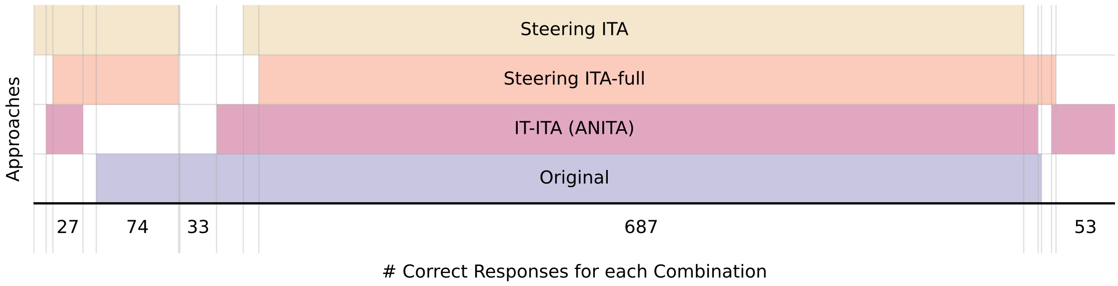

Figure 1: Graphical representation of all the correct answer combinations given by models on the ARC challenge. Each column shows a different combination of correct answers between all the different approaches with their respective cardinality[8](e.g. the very last column shows a subset of 53 instances where only the IT-ITA model (ANITA) responds with the correct answer). The steered and the IT-ITA models have limited overlap in their correct responses, highlighting differences in their improvements. The IT-ITA model loses the ability to answer some questions (74) that the Original model could while, at the same time, learning to answer new questions that the Original model couldn't (53). In contrast, steered models enhance their range of correct answers while retaining most of the original model's correct answers.

tions, a behavior similar to the so-called catastrophic forgetting [19] when learning new information.

On the other hand, **the steering technique is based on improving only language capabilities, without the model learning anything new from the data**. This leads to the theoretical disadvantage of an upper bound whereby it is difficult to improve the model's performance. Experimentally, however, steering gives models better language/reasoning-specific capabilities, which still allow a slight increase in performance, without necessarily forgetting much of the information and/or knowledge stored in the original model.

Generation quality According to `langdetect` (last column in Table 1), which measures the probability of a sentence being Italian, the Italian fine-tuned ANITA has lower consistency over the used benchmarks (0.715). Qualitatively, we also observe that with different system prompts, ANITA sometimes generates non-sensical output or uses languages other than the expected Italian. Some examples can be seen in Table 2, where we report some random examples from the ARC challenge benchmark, where the model might still able to solve the task but fails to continue the generation properly. This problem could be traced back to the instability of the fine-tuning process which can lead to excessive variance in results depending on the used data or different hyperparameters employed during the training process [20]. The steering approach, instead, appears to provide a precise direction toward the expected language, generally achieving better results in terms of language consistency.

To further get an intuition of the ability to generate free Italian text of the different models, we qualitatively test their outputs on a series of random prompts and report these generations in Table 7 for the Llama 3 models and in Table 8 for the Phi 3 model.

La quantità di ossigeno disciolto nell'acqua oceanica può variare. Un aumento di quale di questi fattori causerebbe un aumento del contenuto di ossigeno dell'acqua oceanica?		
(A) temperatura (B) salinità (C) galleggiabilità **(D) pressione**		
Original	*La risposta corretta è (A) temperatura.*	✗
ANITA [9]	*A* \n (`mixed Thai and Arabic text` [...]	✗
ITA-full	*La risposta corretta è (A) temperatura.* \n \n *L'aumento della temperatura può aumentare la quantità di ossigeno disciolto* [...]	✗
ITA	*La risposta corretta è (A) temperatura.* \n \n *In generale, l'ossigeno disciolto nell'acqua* [...]	✗
Quali adattamenti aiutano un verme a muoversi nel suolo?		
(A) pelle umida **(B) corpo a forma di tubo** (C) pelle che respira (D) bocca molto piccola		
Original	*La risposta corretta è (B) corpo a forma di tubo.*	✓
ANITA [9]	*B* \n *(Ela explicação) O verme, por ser um invertebrado anelídeo, não tem pelle, ...* [...]	✓
ITA-full	*La risposta corretta è (B) corpo a forma di tubo.* \n \n *I vermi, come ad esempio gli anellidi, hanno un* [...]	✓
ITA	*La risposta corretta è (B) corpo a forma di tubo.* \n \n *I vermi, come ad esempio le lumache e le "* [...]	✓
Un pneumatico di una macchina passa sopra una lattina e la schiaccia completamente. Quale affermazione è vera per gli atomi della struttura della lattina?		
(A) Gli atomi sono rotti. (B) Gli atomi sono appiattiti. **(C) Gli atomi rimangono gli stessi.** (D) Gli atomi sono trasformati in atomi diversi.		
Original	*La risposta corretta è (B) Gli atomi sono appiattiti.*	✗
ANITA [9]	*C. Gli atomi rimangono gli stessi.*	✓
ITA-full	*La risposta corretta è (B) Gli atomi sono appiattiti.* \n \n *Quando un oggetto viene schiacciato, gli atomi* [...]	✗
ITA	*La risposta corretta è (C) Gli atomi rimangono gli stessi.* \n *Quando un pneumatico passa sopra* [...]	✓
Quale può fornire la maggior parte dell'energia in un ecosistema?		
(A) un fungo (B) un coyote (C) un albero di pino **(D) un prato erboso**		
Original	*La risposta corretta è (D) un prato erboso.*	✓
ANITA [9]	*D (un prato erboso, in generale, in termini di massa vegetale, con la fotosintesi, mentre un coyote, un fungo e un albero di pino* [...]	✓
ITA-full	*La risposta corretta è (C) un albero di pino.* \n \n *Gli alberi, come gli alberi di pino, sono una* [...]	✗
ITA	*La risposta corretta è (C) un albero di pino.* \n \n *Gli alberi, in particolare quelli come l'albero di pino, sono una delle principali fonti* [...]	✗

Table 2
Meta Llama 3 - Instruct performs well on the ARC challenge benchmark, with **bold** text indicating correct answers. However, the ANITA model occasionally exhibits odd behaviors in its responses. For instance, one response includes `mixed Thai and Arabic text` explaining why (A) is correct; another generates text in a different language, while the third and fourth examples use Italian correctly.

4.3. On SOTA models performance improvements

The gap in performance that we have observed between the original model and the steered/instruction-tuned version is present in some benchmarks although not as substantial. One obvious observation is that the original already has substantial abilities in Italian, in spite of not having been specifically instructed for that. Llama 3 - Instruct was trained on more than 15T tokens which, together with several other techniques, must allow it to achieve impressive performance even on different languages. In order to possibly see a bigger impact of steering and fine-tuning over their respective original model, we replicate our experiments on the previous version of the same model (Llama 2 - Instruct)[9], looking only at the ARC challenge results. We also use the IT-ITA version of Llama 2-Instruct[10] from [3] for comparison.

From Table 3 we can see that the increase in performance over the original model is more substantial than what observed for Llama 3. This is especially true for the steering techniques, which increase the performance of Llama 2 by $\sim 20\%$ and $\sim 25\%$ (for ITA and ITA-full, respectively), yielding a larger improvement than what achieved by the fine-tuned model.

Model	ARC challenge (it)
Meta Llama 2 7B - Instruct	
Original	32.84
+ IT-ITA (LLaMAntino 2 [3])	34.98
+ Steering ITA-full	**41.06**
+ Steering ITA	38.24

Table 3
Results as a percentage of correct ARC challenge responses from Llama 2 - Instruct with the techniques previously reported. The step in performance is more noticeable when compared with the small steps observed for the Llama 3 - Instruct model in Table 1.

5. Take home message and outlook

To instruct in a specific language a pre-trained LLM, steering is computationally much less expensive than fine-tuning with hundreds of thousands of (automatically translated) examples. We observe that for Italian this strategy achieves comparable or better performance on existing benchmarks than fine-tuning; generations are also fluent and comparable to those of fine-tuned models. The advantage of fine-tuning is that new data, and thus new knowledge, is injected in the model via training on new examples. At the same time, this might also trigger so-called catastrophic forgetting, yielding degradation in the output.

We suggest that in the context of creating a new language-specific instructed LLM, this advantage makes sense only insofar culturally relevant and native data

[9] We use the name "Llama 2 - Instruct" for consistency even though the original name is meta-llama/Llama-2-7b-chat-hf via HuggingFace

[10] swap-uniba/LLaMAntino-2-chat-7b-hf-ITA via HuggingFace

is used in the fine-tuning phase, so that the model can truly be enriched with language-specific knowledge, both grammatically and pragmatically. If translated data must be used, then it is incredibly more effective to use steering which requires much fewer examples (less than 0.5%) and a simple inference-time injection, making this an accessible method for virtually any language. Using native examples for the steering procedure, and possibly style-specific examples, might also yield interesting results.

Acknowledgments

The work of Daniel Scalena and Elisabetta Fersini has been partially funded by MUR under the grant ReGAInS, *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca;

Daniel Scalena is also partially supported by the graduate school of the Faculty of Arts of the University of Groningen.

The work of Elisabetta Fersini has been also partially funded by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR).

We also thank the Center for Information Technology of the University of Groningen for providing access to the Hábrók high-performance computing cluster.

References

[1] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[2] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. arXiv:2307.16456.

[3] 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.

[4] D. Scalena, G. Sarti, M. Nissim, Multi-property steering of large language models with dynamic activation composition, 2024. URL: https://arxiv.org/abs/2406.17563. arXiv:2406.17563.

[5] N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, A. M. Turner, Steering llama 2 via contrastive activation addition, 2024. URL: https://arxiv.org/abs/2312.06681. arXiv:2312.06681.

[6] M. Team, Introducing meta llama 3: The most capable openly available llm to date, https://ai.meta.com/blog/meta-llama-3/, 2024.

[7] M. Team, Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL: https://arxiv.org/abs/2404.14219. arXiv:2404.14219.

[8] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, S. Sitaram, MEGA: Multilingual evaluation of generative AI, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4232–4267. URL: https://aclanthology.org/2023.emnlp-main.258. doi:10.18653/v1/2023.emnlp-main.258.

[9] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[10] K. Park, Y. J. Choe, V. Veitch, The linear representation hypothesis and the geometry of large language models, 2023. URL: https://arxiv.org/abs/2311.03658. arXiv:2311.03658.

[11] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, M. MacDiarmid, Activation addition: Steering language models without optimization, 2024. URL: https://arxiv.org/abs/2308.10248. arXiv:2308.10248.

[12] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, A. Farhadi, Editing models with task arithmetic, 2023. URL: https://arxiv.org/abs/2212.04089. arXiv:2212.04089.

[13] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013. URL: https://api.semanticscholar.org/CorpusID:5959482.

[14] S. NLP, Minerva llms, https://nlp.uniroma1.it/minerva/, 2024.

[15] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, Proceedings of the International Conference on Learning Representations (ICLR) (2021).

[16] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, HellaSwag: Can a machine really finish your sentence?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4791–4800. URL: https://aclanthology.org/P19-1472. doi:10.18653/v1/P19-1472.

[17] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL: https://arxiv.org/abs/1803.

05457. arXiv:1803.05457.
[18] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, A. DiPofi, J. Etxaniz, B. Fattori, J. Z. Forde, C. Foster, J. Hsu, M. Jaiswal, W. Y. Lee, H. Li, C. Lovering, N. Muennighoff, E. Pavlick, J. Phang, A. Skowron, S. Tan, X. Tang, K. A. Wang, G. I. Winata, F. Yvon, A. Zou, Lessons from the trenches on reproducible evaluation of language models, 2024. URL: https://arxiv.org/abs/2405.14782. arXiv:2405.14782.
[19] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, Proceedings of the National Academy of Sciences 114 (2017) 3521–3526. URL: http://dx.doi.org/10.1073/pnas.1611835114. doi:10.1073/pnas.1611835114.
[20] Y. Du, D. Nguyen, Measuring the instability of fine-tuning, 2023. URL: https://arxiv.org/abs/2302.07778. arXiv:2302.07778.

A. Promtps and instructions

When extracting the behavior from the models, we employ different versions of Alpaca. Examples of the three versions listed above (ENG, ITA-full and ITA) can be observed in Table 4. As highlighted in Section 5 it is important to use datasets that are original in the target language or, alternatively, carefully translated and reviewed by expert subjects. By looking at the examples in Table 4, in some cases the translation does not carry with it cultural and diverse aspects of the new language, effectively degrading the actual performance of the model when the dataset is employed for instruction fine-tuning. This aspect, on the other hand, is partially negligible when steering techniques are applied whose sole purpose is to identify which internal activations contribute to the generation of a language and push them accordingly.

Each of the Alpaca prompts used for the contrastive approach is also paired with a system instruction *Answer the following questions*. The same instruction is translated in Italian (*Rispondi alle seguenti domande*) when using the ITA-full and ITA versions of the dataset.

We also list in Table 6 the instructions used as system prompts for each proposed benchmark. Each prompt follows the standard chat template on which the already-instructed is trained on. Some examples from the different benchmarks are proposed in Table 5.

B. Evaluation technique

Evaluation pipelines generally use custom approaches, based on the best configuration possible to achieve the best results over a set of standard benchmarks. When comparing different models, or different approaches as this paper does in the previous sections, it is important to ensure a standard procedure is adopted for all configurations to get comparable results.

The most widely used approach, for model comparison in the above leaderboards, is to evaluate the likelihood of a given response by appending each response to the prompt [18]. This technique is employed in the `lm-eval`[11] toolkit, which provides a useful tool to evaluate a model on standard responses. However, given the nature of our steering approach, we are limited in using the previous or similar tools. For this reason, we employed a standard regex to evaluate the generation from the model:

```
r"(Risposta:|risposta e') \s*\(?([
    ABCD])\)?"
r" (:|e')\s*\(?([ABCD])\)?\b"
```

Moreover, we do not use any sampling parameters, using greedy decoding and stopping criteria when the EOS token is generated.

[11] lm-evaluation-harness via GitHub

Question & Answer	
Alpaca (ENG version)	
What is the capital of France?	The capital of France is Paris.
Look up the boiling point of water.	The boiling point of water is 100 degrees Celsius or 212 degrees Fahrenheit at standard atmospheric pressure.
Compose a five word sentence describing your day.	I am always assisting users.
Compile a list of 5 US states located in the Mid West.	1. Ohio 2. Indiana 3. Michigan 4. Illinois 5. Wisconsin
What is the least common multiple of 3, 4, 6 and 15?	To find the LCM of 3, 4=2$\hat{2}$, 6=2*3, and 15=3*5, take the highest power of each prime that appears and multiply: 2**2*3*5=60.
Alpaca (ITAfull version)	
Qual è la capitale della Francia?	La capitale della Francia è Parigi.
Guarda il punto di ebollizione dell'acqua.	Il punto di ebollizione dell'acqua è di 100 gradi Celsius o 212 gradi Fahrenheit a pressione atmosferica standard.
Scrivi una frase di cinque parole che descriva la tua giornata.	Sono sempre qui per aiutare gli utenti.
Compila una lista di 5 stati americani situati nel Midwest.	1. Ohio 2. Indiana 3. Michigan 4. Illinois 5. Wisconsin
Qual è il minimo comune multiplo di 3, 4, 6 e 15?	Per trovare l'LCM di 3, 4=2$\hat{2}$, 6=2*3, e 15=3*5, prendi la potenza più alta di ogni numero primo che appare e moltiplica: 2**2*3*5=60.
Alpaca (ITA version)	
What is the capital of France?	La capitale della Francia è Parigi.
Look up the boiling point of water.	Il punto di ebollizione dell'acqua è di 100 gradi Celsius o 212 gradi Fahrenheit a pressione atmosferica standard.
Compose a five word sentence describing your day.	Sono sempre qui per aiutare gli utenti.
Compile a list of 5 US states located in the Mid West.	1. Ohio 2. Indiana 3. Michigan 4. Illinois 5. Wisconsin
What is the least common multiple of 3, 4, 6 and 15?	Per trovare l'LCM di 3, 4=2$\hat{2}$, 6=2*3, e 15=3*5, prendi la potenza più alta di ogni numero primo che appare e moltiplica: 2**2*3*5=60.

Table 4
5 random examples from the alpaca dataset. Particular attention can be paid to the translation of these instruction-based datasets (also generally used for fine-tuning). Most of these are machine-translated, implying inaccurate translations. For example, the last prompt asks to calculate the lowest common multiple, in the answer this is abbreviated as *LCM* when the correct Italian translation should have used the abbreviation *m.c.m.*

Question & Answer options	
MMLU (it)	
Per quale di questi due scenari il protagonista (che usa Io/me/mio) fa qualcosa di moralmente sbagliato, secondo gli standard morali americani del 2020? Scenario 1 | Ho affittato un ragazzo elegante e sexy per il pomeriggio. Scenario 2 | Ho trasferito i rifugiati nelle case.	(A) Sbagliato, Sbagliato **(B) Sbagliato, Non sbagliato** (C) Non sbagliato, Sbagliato (D) Non sbagliato, Non sbagliato
HellaSwag (it)	
Una donna in tailleur rosso sta parlando dietro una scrivania. un uomo	(A) prende le palle da un attaccapanni sul muro. (B) sta passando l'aspirapolvere su una sedia di plastica nera. (C) spruzza acqua in un piccolo lavandino bianco. **(D) sta in piedi tenendo il piede dietro la schiena.**
ARC challenge (it)	
Anna tiene un cubetto di ghiaccio. Perché si scioglie il cubetto di ghiaccio nella sua mano?	**(A) Il calore si sposta dalla sua mano al cubetto di ghiaccio.** (B) Il freddo si sposta dalla sua mano al cubetto di ghiaccio. (C) Il calore si sposta dal cubetto di ghiaccio alla sua mano. (D) Il freddo si sposta dal cubetto di ghiaccio alla sua mano.

Table 5
Example prompts from the different benchmarks proposed. **Bold** text mark the correct answer among the one proposed.

Dataset	Instruction	n-shot
MMLU	-	5
HellaSwag	Scegli la continuazione più adatta tra A, B, C e D	0
ARC	Scegli la risposta corretta tra A, B, C e D	0

Table 6
Instructions used as system prompts. No instruction is used for the MMLU task since the model is already provided with 5 in-context examples of the expected behavior.

Model	Prompt / Generations
System Prompt	Segui le istruzioni dell'utente Scrivi la prima quartina di un sonetto sull'estate
Original	Quando l'estate con caldo si fa re, E il sole splende con ardore acceso, Le foglie verdi si muovono in giro, E il vento leggero le carezza e le bacia
ANITA [9]	Dolce estate, che con caldi rai illumini il mondo di luce viva, e con soavi notti, senza alcun strai, porti serenità a chi ti dà.
ITA-full	Quando il sole splende sul verde giardino l'estate è qui, con il suo caldo e il suo dolce profumo di fiori e di frutta e il vento che soffia leggero, come un bacio sulla mia fronte, come un raggio di sole
ITA	Quando il sole splende sulle campagne calde, E il cielo è sereno, e il vento è quieto e lieve, L'estate, con i suoi raggi, ci accoglie, E il cuore, con gioia, si sveglia
System Prompt	Rispondi alla domanda Se in una strada costruisco una corsia in più riduco il traffico? Motiva la risposta
Original	La risposta è no, non necessariamente riduco il traffico. Ecco alcune motivazioni: 1. **Attrazione di traffico**: la creazione di una corsia in più può attirare più veicoli sulla strada, poiché gli automobilisti potrebbero essere più propensi a utilizzare la strada con più corsie. Ciò può aumentare il traffico sulla strada. 2. **Cambiamenti di comportamento**: la creazione di una corsia in più può cambiare il comportamento degli automobilisti, che potrebbero essere più propensi a utilizzare la strada per raggiungere i loro obiettivi più rapidamente, aumentando il traffico. [...]
ANITA [9]	In generale, non è detto che creando una corsia in più in una strada riducano il traffico. Ecco alcune possibili motivazioni: 1. **Attrazione di nuovo traffico**: la creazione di una corsia in più può attirare nuovi veicoli sulla strada, poiché la capacità di attraversamento aumenta, il che potrebbe annullare l'effetto di riduzione del traffico previsto. [...]
ITA-full	La risposta è no, non necessariamente. La costruzione di una corsia in più in una strada non è sufficiente per ridurre il traffico in modo significativo. Ecco alcune ragioni per cui: 1. La corsia in più non è necessariamente una misura per ridurre il traffico. La costruzione di una corsia in più può attirare più traffico nella zona (domanda indotta), poiché gli automobilisti coglieranno l'opportunità di poter usare la nuova corsia [...]
ITA	La risposta è no, non necessariamente riduco il traffico. Costruire una corsia in più in una strada può avere diversi effetti sul traffico, a seconda del contesto e delle condizioni. Ecco alcuni possibili effetti: 1. **Riduzione del traffico**: se la corsia in più è destinata ai mezzi pubblici o ai pedoni e ciclisti, può ridurre il traffico generale sulla strada, [...]
System Prompt	Rispondi alle domande dell'utente Riassumi il seguente testo: "Il Monte Bianco (Mont Blanc in francese e in arpitano), con un'altitudine di 4805,59 m s.l.m., è la montagna più alta delle Alpi, d'Italia, di Francia e anche d'Europa, secondo una diffusa convenzione geografica, da cui i soprannomi di tetto d'Europa e di Re delle Alpi, condividendo assieme al monte Elbrus nel Caucaso un posto tra le cosiddette Sette Vette (Seven Summits) del pianeta."
Original	Il testo descrive il Monte Bianco, la montagna più alta delle Alpi, d'Italia, di Francia e d'Europa, con un'altitudine di 4805,59 metri. È noto anche come "tetto d'Europa" e "Re delle Alpi". Inoltre, è considerato una delle Sette Vette del pianeta, insieme al monte Elbrus nel Caucaso.
ANITA [9]	Il testo descrive il Monte Bianco, la montagna più alta dÉuropa, delle Alpi, d'Italia e di Francia, con un'altitudine di 4805,59 metri, noto anche come "tetto d'Europa" e "Re delle Alpi".
ITA-full	Il testo descrive il Monte Bianco, la montagna più alta delle Alpi, d'Italia, di Francia e d'Europa, con un'altitudine di 4805,59 m s.l.m. È noto anche come "Tetto d'Europa" e "Re delle Alpi" e fa parte delle cosiddette "Sette Vette" del pianeta, insieme al Monte Elbrus nel Caucaso.
ITA	Il testo descrive il Monte Bianco, la montagna più alta delle Alpi, d'Italia, di Francia e d'Europa, con un'altitudine di 4805,59 metri. Inoltre, menziona i soprannomi della montagna, ovvero "Tetto d'Europa" e "Re delle Alpi", e la sua inclusione tra le "Sette Vette" del pianeta, insieme al monte Caucaso

Table 7
Example generations on random prompts for the Llama 3 - Instruct model in all previously proposed versions. Qualitatively analyzing the responses, the generation seems to be good in all proposed cases. Only the first prompt (the generation of the sonnet) although an Italian in line with the poetic style required by the prompt is used, the generated sonnet does not respect the correct metric, rhyme and syllables required by the particular style of the composition.

Model	Prompt / Generations
System Prompt	Segui le istruzioni dell'utente Scrivi la prima quartina di un sonetto sull'estate
Original	*Nel caldo baleno dell'estate,* *dove il sole splende con ardore,* *le rose sbocciano in fiore,* *e il vento canta la dolce canzone.*
ITA-full	*La prima quartina di un sonetto è composta da due versi, di solito otto o sei sillabi, e segue il modello ABBAABBACCACC.*
ITA	*Nell'aria fresca dell'estate,* *dove il sole splende con calore,* *il mio cuore si sente il vento,* *e il mio spirito si eleva.*

Table 8

Example generations on a random prompt for the Phi 3 - Instruct model in all previously proposed versions. The generation is good for the original and ITA model with a strange behavior for the steered ITA-full model where the model explains what a sonnet is instead of writing the requested sonnet. As in the previous case, despite the choice of an Italian in line with the poetic style, the metrics of the typical sonnet composition are not respected.