# Assessing Italian Large Language Models on Energy Feedback Generation: A Human Evaluation Study

Manuela Sanguinetti*, Alessandro Pani, Alessandra Perniciano, Luca Zedda, Andrea Loddo and Maurizio Atzori

[1]*Department of Mathematics and Computer Science, University of Cagliari, Italy*

**Abstract**

This work presents a comparison of some recently-released instruction-tuned large language models for the Italian language, focusing in particular on their effectiveness in a specific application scenario, i.e., that of delivering energy feedback. This work is part of a larger project aimed at developing a conversational interface for users of a renewable energy community, where clarity and accuracy of the provided feedback are important for proper energy management. This comparison is based on the human evaluation of the output produced by such models using energy data as input. Specifically, the data pertains to information regarding the power flows within a household equipped with a photovoltaic (PV) plant and a battery storage system. The goal of the feedback is precisely that of providing the user with such information in a meaningful way based on the specific aspect they intend to monitor at a given moment (e.g., self-consumption levels, the power generated by the PV panels or imported from the main grid, or the battery state of charge). This evaluation experiment has the two-fold purpose of providing an exploratory analysis of the models' abilities on this specific generation task solely relying on the information and instruction provided in the prompt and as an initial investigation into their potential as reliable tools for generating user-friendly energy feedback in this intended scenario.

**Keywords**

energy feedback, large language models, Italian,

## 1. Introduction and Motivations

The provision of energy feedback plays a crucial role in promoting energy efficiency among users. The expression *energy feedback* (or *eco-feedback*) covers a wide range of energy-related information. This can include detailed reports on energy usage and production (in the case of renewable energy sources), as well as energy-saving advice, whether generic or user-specific. The primary goal of energy feedback is to allow users to make informed decisions regarding their energy management, thus promoting better conservation practices.

A substantial body of literature within the field of Human-Computer Interaction (HCI) has explored various energy feedback mechanisms, primarily focusing on visual or ambient feedback as well as gamification techniques (we refer to the surveys proposed by Albertarelli et al. [1] and Chalal et al. [2] for further details on these aspects). However, a greater interest has been reported on the delivery of energy feedback through conversational agents [3]. Furthermore, within the field of Natural Language Generation (NLG), several studies prior to the advent of Large Language Models (LLMs) investigated the use of NLG architectures to communicate consumption data. Notable works include those by Trivino and Sanchez-Valdes [4] and Conde-Clemente et al. [5], which used fuzzy sets to tackle data-to-text generation tasks, also tailoring the linguistic description on given consumption profiles. Similarly, Martínez-Municio et al. [6] employed fuzzy sets to produce linguistic summaries based on the consumption of specific buildings or groups of buildings, using time series data as input.

This work is part of a research project aimed at developing a modular task-oriented conversational agent to inform users about their energy consumption and photovoltaic (PV) production and, more generally, to support better management of their energy resources through text-based energy feedback. The conversational agent will then be deployed and tested within a renewable energy community in Italy, which motivates our specific focus on Italian as the primary language for the interactions. At this stage of the project, we plan to integrate the generative abilities of LLMs into the conversational pipeline.[1] This approach is expected to deliver more varied and dynamic responses instead of predefined, static templates, possibly making the user experience enjoyable. This study was driven by the need to obtain more quantitative insights into the expected performance of such models when tasked with the delivery of energy

*CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy*

*Corresponding author.

manuela.sanguinetti@unica.it (M. Sanguinetti); alessandro.pani2@unica.it (A. Pani); alessandra.pernician@unica.it (A. Perniciano); luca.zedda@unica.it (L. Zedda); andrea.loddo@unica.it (A. Loddo); atzori@unica.it (M. Atzori)

0000-0002-0147-2208 (M. Sanguinetti); 0009-0003-8956-5058 (A. Perniciano); 0009-0001-8488-1612 (L. Zedda); 0000-0002-6571-3816 (A. Loddo); 0000-0001-6112-7310 (M. Atzori)



[1]For the time being, we do not aim to use these models as complete conversational agents but only within the generation module.

feedback based on actual energy data.

The main objective of this study thus aims to verify how effectively instruction-tuned LLMs currently available for the Italian language can deliver clear and accurate feedback based on energy data provided within a prompt, without relying on more elaborate techniques like fine-tuning or Retrieval Augmented Generation. More specifically, we formulated the following research questions:

- Are the LLMs under study able to produce energy feedback that is 1) informative, 2) comprehensible, and 3) accurate with respect to the provided energy data?
- Are there any major differences among such models with respect to these capabilities?

To answer these questions, we conducted an exploratory analysis by manually evaluating some of these Italian LLMs, organizing the study around criteria designed to quantify these specific aspects.

This work closely aligns with a recent initiative that has been launched within the Italian NLP community, i.e., CALAMITA[2], a campaign aimed at evaluating the capabilities of Italian (or multilingual, but including Italian) LLMs on specific tasks in zero or few-shot settings. Unlike the latter, however, our study relies solely on human judgments rather than automatic metrics. The main challenges of a manual approach include the absence of standardized practices and evaluation criteria [7], as well as the lack of systematic documentation [8], which hinders the reproducibility of such studies.[3] In light of these challenges, the intended contributions of this paper are outlined below:

- A small-scale human evaluation of several Italian LLMs on a specific task.
- The description of a protocol for human evaluation inspired by the good practices recommended in recent literature [9, 10]. To this end, we also make available the evaluation dataset, with the ratings assigned by the evaluators in a non-aggregated form.[4]

The remainder of this paper describes how this study was designed and carried out, with a discussion of the results obtained and the main limitations of the work.

# 2. Study Design

As anticipated in the previous section, the main goal of this human evaluation experiment is to assess the overall quality (using specific criteria that will be defined later) of the energy feedback generated by Italian LLMs. The task assigned to the tested models is broadly intended as a summarization task in that the expected output is supposed to provide a summary of the relevant information available in the prompt. What follows is the overview of the main principles that guided the selection of the models, the development of the dataset used for evaluation, and the whole evaluation protocol.

## 2.1. Models and Setting

The models' selection was primarily driven by the intended application scenario of the overarching project (also mentioned in the previous section), which narrowed down our choice to Italian models. In addition, we opted for open-source models that can be run locally, avoiding using APIs. For greater simplicity and practicality, we looked for the Italian models available on HuggingFace, the reference platform for the release of such resources. As a final choice, we exclusively selected instruction-tuned models. These models are trained to follow a wide range of instructions provided in the prompt, offering greater flexibility in handling diverse tasks compared to more specialized fine-tuned models.[5] This ability makes them particularly suitable for our purposes. In light of this, we selected for our study the following models[6]: Cerbero-7B[7] [11], LLaMAntino2-7B [12], and more specifically the version trained on the UltraChat-ITA dataset[8], LLaMAntino3-8B-ANITA[9] [13], and Zefiro-7B[10].

Regarding the text generation settings, we chose high-temperature values to allow the generation of more diverse responses. Specifically, we set both temperature and $top\_p$ to 0.9 in order to obtain less deterministic and more varied outputs. On the other hand, to ensure a balance between variety and coherence, we kept the $top\_k$ value low (0.2). After some preliminary tests, we found that these settings provided satisfactory results and could be reasonably used for the actual evaluation phase. As regards the output length, we limited its maximum to 250 tokens to prevent excessively lengthy responses and disabled the option that returns the input prompt as part of the output.

[2] https://clic2024.ilc.cnr.it/calamita/

[3] An attempt in this respect is made within the ReproHum project: https://reprohum.github.io/

[4] https://github.com/msang/nl-interface/tree/main/humEval

[5] It is important to note, however, that depending on the task at hand, a prompt (even if supplemented with additional examples) may not be sufficient to obtain good results, and further model refinements might be necessary.

[6] For simplicity, throughout the paper, only the models' names will be used, without including parameter specifications or additional suffixes.

[7] https://huggingface.co/galatolo/cerbero-7b

[8] https://huggingface.co/swap-uniba/LLaMAntino-2-chat-7b-hf-UltraChat-ITA

[9] https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

[10] https://huggingface.co/giux78/zefiro-7b-beta-ITA-v0.1

## 2.2. Data and Prompts

The dataset used for evaluation comprises responses from each of the four models tested. These responses were based on an input prompt consisting of two fixed components — the premise and the instruction — and two dynamic elements: user request and information on energy data (see also Figure 1).

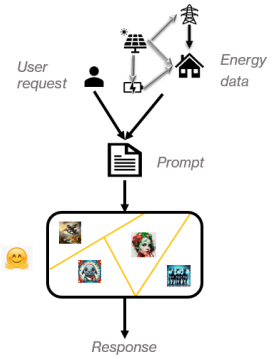


**Figure 1:** Pipeline for creating the evaluation dataset used in models' comparison.

Regarding the latter, the data available for the experiments can vary and is related to the specific use case of a household equipped with a PV system and a battery storage solution. In this scenario, the PV system can distribute the energy produced to meet user consumption needs, charge the battery, or feed into the main grid. The battery, in turn, can supply power to the user, especially when there is no solar production. The data presented in the prompt describes the energy flow among these different sources and is listed in the form of verbal descriptions, each accompanied by the corresponding data value and unit of measure (or current status if referred to the battery). This data is summarized in Table 1. In order to provide a more realistic depiction of the usage scenario and to introduce a greater variety in the prompt to be processed by the models, the included data encompasses various combinations of values across different aspects (e.g., including greater or lesser household consumption or solar production or different battery charge levels).

The user requests were randomly sampled from an in-house dataset for intent detection previously developed to train the NLU module of the conversational agent of the main project.[11] The types of user requests used in the evaluation focused on typical monitoring functions. These requests primarily aim to check energy consumption or production data from the PV panels. They may be focused on information such as household energy usage, battery charge status, or current power generation (e.g., *quanto stanno producendo i pannelli?*, EN: "how much are the panels producing?"). Furthermore, requests may require brief and concise responses about a single specific information (*quanto è carica la batteria?*, EN:"how charged is the battery?"), or more comprehensive overviews (*mi serve un quadro completo dei consumi*, EN:"I need a full overview of the consumption").

[11]The backbone architecture of the agent has been developed using RASA [14], and the corpus was originally created to train its built-in classifier, DIET [15].

**Table 1**
List of the data provided in the prompt.

| Description | Unit/Status |
|---|---|
| Current power used<br>Power fed into the grid<br>Power supplied by the PV system | kW |
| Battery state of charge<br>Battery status | %<br>charging/<br>discharging/<br>inactive |
| Total energy used by the house<br>Total energy produced by the panels<br>Total energy purchased from the grid<br>Self-consumption<br>Total energy fed into the grid | kWh |

The instruction provided in the prompt, aiming to reflect the main intended task, was formulated as follows: "*Riassumi le informazioni che ti ho appena fornito per rispondere alla seguente domanda: [USER_REQUEST]* (EN: "Summarize the information I have just provided to answer the following question").

The final dataset for the evaluation phase comprises 50 responses from each model, hence 200 responses overall. The following section provides a detailed description of the evaluation process.

## 2.3. Evaluation Protocol

The actual evaluation phase was preceded by a briefing session and a pilot annotation phase. During the briefing, evaluators discussed the task at hand in order to make sure they fully understood the evaluation criteria and the meaning of the scale values. Following the briefing, a pilot evaluation was carried out. This step allowed evaluators to familiarize themselves with the process and refine their understanding of the evaluation criteria. Once these preparatory steps were completed, they proceeded with the main evaluation task. They worked independently and were not aware of the specific models they were evaluating, to mitigate possible biases deriving from any preconceived notions of the models.

Four human evaluators, who are co-authors of this paper, conducted the evaluation task. The group comprises three males and one female, each with a back-

**Table 2**
Overview of the evaluation criteria and the corresponding statement rated by human judges.

| Criteria | | Statement |
|---|---|---|
| Informativeness | Usefulness | The system's response includes *only* the information that is relevant and helpful in addressing the user's query (thus avoiding unnecessary details). |
| | Necessity | The system's response includes *all* the details necessary to fully respond to the user's request. |
| Comprehensibility | Understandability | The information is clear and easy to follow. |
| | Fluency | The response reads smoothly. |
| Accuracy | | The factual content is correct. |

ground in Computer Science and ranging from graduate students to assistant professors. While all evaluators are familiar with technologies such as conversational agents and possess a good understanding overall of LLMs, their knowledge of concepts related to electricity (e.g., the distinction between power and energy) and renewable energy technologies (such as PV systems and storage solutions) varies from minimal to substantial.

Evaluators were instructed to assign a Likert-type rating on a 1-7 scale to each model response for each evaluation criterion. The rating scale is anchored with symmetrical verbal labels as follows: 1: *Strongly Disagree*; 2: *Disagree*; 3: *Mildly Disagree*; 4: *Neither Agree nor Disagree*; 5: *Mildly Agree*; 6:*Agree*; 7: *Strongly Agree*.

As regards the evaluation criteria, they were designed to address the three dimensions outlined in our first research question: informativeness, comprehensibility, and accuracy. These dimensions represent the factors we deemed essential in the delivery of effective energy feedback; ultimately they are meant to guide the choice of the most suitable model for our intended application scenario. To evaluate informativeness, we drew inspiration from previous work by Mazzei et al. [16], considering two complementary aspects: *Usefulness*, i.e., the extent to which the information provided by the system is useful in responding to the user's request, and *Necessity*, i.e., the completeness of the information provided, ensuring all necessary details are included. Similarly, to assess the comprehensibility of the models' responses, we considered two criteria: *Understandability*, i.e., the extent to which the information is presented in an easy-to-understand manner, and *Fluency*, i.e., the degree to which a text 'flows well'. The third dimension, *Accuracy*, was evaluated based on the degree to which the content of an output is correct, accurate, and true relative to the input. The definitions of Understandability, Fluency, and Accuracy were drawn from the overview proposed in Howcroft et al. [7]. For each of these five criteria, evaluators were asked to assign a rating within the proposed scale, guided by a specific question associated with each criterion (see Table 2).

To both facilitate the evaluators' work and ensure an accurate rating for each evaluation criterion, each model response was presented alongside the user's request in isolation as well as the entire prompt. This provided them with the full context needed to carry out the task and allowed them to understand the information the model had access to during the response generation. Some examples of prompts, along with the model's output and the evaluation provided by the judges, are reported in Sections A.1-A.2.

## 3. Results

Once all judges completed the task, we first measured the Inter-Annotator Agreement using Krippendorff's $\alpha$.[12] We computed the metric separately for each model and each evaluation criterion. Results are summarized in Table 3, which also shows the average results both per model and criterion.

The results reveal varying levels of consistency among the evaluators, ranging from moderate to low agreement across all criteria. In particular, Understandability and Fluency exhibit a higher degree of disagreement among the evaluators. This could be due to the subjective nature of these criteria, as different evaluators might give different interpretations of what is considered comprehensible and linguistically fluent. Overall, this variation highlights the probable need for more training for evaluators to improve their consistency, especially in assessing subjective criteria.

As for the models' comparison, we first aggregated all ratings assigned in order to provide an overview of the models' output across all five evaluation criteria. Since the data is ordinal, we use the median value as an aggregation function to assess the central tendency of the ratings (as also suggested in Amidei et al. [9]). The results, shown in Table 4, indicate medium to high ratings overall across all models. To thus answer our first research

[12]We used the statistical package K-Alpha Calculator [17]: https://www.k-alpha.org/

**Table 3**
Results of the Inter-Annotator Agreement computed with Krippendorff's $alpha$.

| Criteria | | Cerbero | LLaMAntino2 | LLaMAntino3 | Zefiro | *avg.* |
|---|---|---|---|---|---|---|
| Informativeness | Usefulness | 0.57 | 0.77 | 0.32 | 0.34 | 0.50 |
| | Necessity | 0.19 | 0.75 | 0.32 | 27 | 0.38 |
| Comprehensibility | Understandability | 0.27 | 0.28 | 0.16 | 0.12 | 0.21 |
| | Fluency | 0.33 | 0.13 | 0.32 | 0.18 | 0.24 |
| Accuracy | | 0.41 | 0.76 | 0.62 | 0.48 | 0.57 |
| *avg.* | | 0.35 | 0.54 | 0.35 | 0.28 | - |

**Table 4**
Results of the human evaluation on the four models reported with median values.

| Criteria | | Cerbero | LLaMAntino2 | LLaMAntino3 | Zefiro |
|---|---|---|---|---|---|
| Informativeness | Usefulness | 6 | 4 | 6 | 6 |
| | Necessity | 7 | 5 | 6 | 7 |
| Comprehensibility | Understandability | 7 | 6 | 7 | 7 |
| | Fluency | 6 | 6 | 7 | 6 |
| Accuracy | | 7 | 4 | 7 | 7 |

question, we examined the overall medians for each evaluation criterion. The values obtained show that they perform reasonably well despite the variability across the models. Concerning the dimension of informativeness, ratings range from 4 to 6 in Usefulness and from 5 to 7 in Necessity, suggesting that further refinements might be necessary to ensure that the energy feedback delivered is useful and complete. In terms of comprehensibility, the corresponding criteria show that all models are capable of generating responses that are easily understandable and fluent, which are both relevant factors that might contribute to a more enjoyable user experience in view of the possible integration of such models in a conversational interface. Also as regards Accuracy, the energy feedback generated by the models is generally correct, with only one exception (LLaMAntino2). This indicates that, overall, the models provide accurate and reliable information, another important factor when users have to make informed decisions based on that feedback.

To answer our second research question, we then considered the overall differences among the models. As also shown in Table 4, LLaMAntino2 quite consistently received lower ratings, particularly for Usefulness and Accuracy, while the other models received high ratings overall, suggesting that they might be considered comparable. To inspect this further, we carried out some statistical tests. We first used the Kruskal-Wallis test, a non-parametric test suitable for ordinal data, to compare the distributions of more than two independent groups. We used it to determine whether the differences among the median values obtained for the models were statistically significant, and the comparisons were carried out separately for each evaluation criterion. This preliminary test confirmed that the differences observed are indeed significant, considering a standard threshold of $p < 0.05$. However, the Kruskal-Wallis test does not determine which models are significantly different from each other. Therefore, we proceeded with pairwise comparisons using Dunn's test. This test confirmed a significant difference between LLaMAntino2 and the other three models.

**Table 5**
P-values obtained with pairwise comparisons between LLaMAntino2 and the remaining models, using Dunn's test, and adjusted using Bonferroni correction.

| | Cerbero | LLaMAntino3 | Zefiro |
|---|---|---|---|
| Usefulness | 5e-04 | 1e-08 | 7e-08 |
| Necessity | 3e-12 | 2e-03 | 4e-04 |
| Understandability | 3e-07 | 1e-03 | 9e-08 |
| Fluency | 2e-04 | 3e-02 | 5e-02 |
| Accuracy | 5e-16 | 1e-10 | 1e-09 |

Table 5 shows the p-values obtained by comparing this model with the other three for each evaluation criterion. The remaining comparisons yielded p-values well above the $0.05$ threshold, therefore the null hypothesis cannot be rejected for those cases. The other three models can thus be considered comparable based on the ratings assigned by the evaluators in our experiment.

## 4. Conclusions and Limitations

This study provides an initial assessment of several Italian language models' ability to generate effective energy feedback. The results indicate that while the models generally perform well, particularly in terms of comprehensibility and accuracy, there is greater variability regarding informativeness. Among the tested models, results show that, except for LLaMAntino2-7B-UltraChat, the remaining ones provide comparable performances. However, it is important to highlight the limitations of this study. First, this is a small-scale study, as it involves a limited number of models and evaluators. Concerning the former issue, we also point out that the study was restricted to models available on HuggingFace, excluding potentially relevant models from external sources, such as Fauno[13] and Camoscio [18]. A more systematic study should consider these models as well, in order to provide a more comprehensive evaluation over the Italian LLMs' landscape. As for the pool of evaluators, it is important to note a significant bias in both their personal backgrounds and demographics. All the judges have a background in computer science and varying degrees of familiarity with the topics at hand. Furthermore, there is a gender imbalance (1 female and 3 male judges) and a lack of age diversity, as all four judges fall within the 24–30 age range. In light of these considerations, a more systematic comparison as the one envisioned above would benefit from a broader and more diverse pool of evaluators. This would not only increase the reliability of the comparison but also provide a deeper understanding of potential correlations between socio-demographic factors, prior knowledge of technology and energy-related concepts, and the differing perceptions of the evaluation criteria considered in our study. Common approaches to address the lack of human participants include the use of crowdsourcing platforms, with a careful design of participation criteria that would ensure a better gender and demographic balance. Alternatively, a user study involving prospective users of the conversational agent could be conducted; this would ultimately enable to gather valuable insights on the type of feedback expected by the target audience. Finally, an extended evaluation framework should also include an analysis of the statistical power of the sample size to ensure more robust conclusions.

Despite these limitations, this work offers a preliminary overview and aims to pave the way for future research on this aspect, also stressing the importance of more standardized human evaluation practices. As a matter of fact, the evaluation protocol we designed draws heavily from methodologies recommended in more general literature pertaining to human evaluation within generation and summarization tasks. Our approach thus aims to ensure that the core principles of the experiment are flexible enough to be easily replicated or adapted for a wider range of different domains.

[13] https://github.com/RSTLess-research/Fauno-Italian-LLM

## Acknowledgments

This work has been developed within the framework of the project e.INS- Ecosystem of Innovation for Next Generation Sardinia (cod. ECS 00000038) funded by the Italian Ministry for Research and Education (MUR) under the National Recovery and Resilience Plan (NRRP) - MISSION 4 COMPONENT 2, "From research to business" INVESTMENT 1.5, "Creation and strengthening of Ecosystems of innovation" and construction of "Territorial R&D Leaders". This work was also partially funded under the National Recovery and Resilience Plan (NRRP) - Mission 4 Component 2 Investment 1.3, Project code PE0000021, "Network 4 Energy Sustainable Transition– NEST".

## References

[1] S. Albertarelli, P. Fraternali, S. Herrera, M. Melenhorst, J. Novak, C. Pasini, A.-E. Rizzoli, C. Rottondi, A Survey on the Design of Gamified Systems for Energy and Water Sustainability, Games 9 (2018). doi:10.3390/g9030038.

[2] M. Chalal, B. Medjdoub, N. Bezai, R. Bull, M. Zune, Visualisation in Energy Eco-Feedback Systems: A Systematic Review of Good Practice, Renewable and Sustainable Energy Reviews 162 (2022). doi:10.1016/j.rser.2022.112447.

[3] M. Sanguinetti, M. Atzori, Conversational Agents for Energy Awareness and Efficiency: A Survey, Electronics 13 (2024). doi:10.3390/electronics13020401.

[4] G. Trivino, D. Sanchez-Valdes, Generation of Linguistic Advices for Saving Energy: Architecture, in: A.-H. Dediu, L. Magdalena, C. Martín-Vide (Eds.), Theory and Practice of Natural Computing, Springer International Publishing, Cham, 2015, pp. 83–94.

[5] P. Conde-Clemente, J. M. Alonso, G. Trivino, Toward Automatic Generation of Linguistic Advice for Saving Energy at Home, Soft Computing 22 (2018) 345–359. doi:10.1007/s00500-016-2430-5.

[6] S. Martínez-Municio, L. Rodríguez-Benítez, E. Castillo-Herrera, J. Giralt-Muiña, L. Jiménez-Linares, Linguistic Modeling and Synthesis of Heterogeneous Energy Consumption Time Series Sets:, International Journal of Computational Intelligence Systems 12 (2018) 259. doi:10.2991/ijcis.2018.125905639.

[7] D. M. Howcroft, A. Belz, M.-A. Clinciu, D. Gkatzia, S. A. Hasan, S. Mahamood, S. Mille, E. Van Miltenburg, S. Santhanam, V. Rieser, Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions, in: Proceedings of the 13th International Conference on Natural Language Generation, Association for Computational Linguistics, Dublin, Ireland, 2020, pp. 169–182. doi:10.18653/v1/2020.inlg-1.23.

[8] A. Shimorina, A. Belz, The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP, in: A. Belz, M. Popović, E. Reiter, A. Shimorina (Eds.), Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 54–75. doi:10.18653/v1/2022.humeval-1.6.

[9] J. Amidei, P. Piwek, A. Willis, The Use of Rating and Likert Scales in Natural Language Generation Human Evaluation Tasks: A Review and some Recommendations, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 397–402. doi:10.18653/v1/W19-8648.

[10] C. Van Der Lee, A. Gatt, E. Van Miltenburg, E. Krahmer, Human evaluation of automatically generated text: Current trends and best practice guidelines, Computer Speech & Language 67 (2021) 101151. doi:10.1016/j.csl.2020.101151.

[11] F. A. Galatolo, M. G. C. A. Cimino, Cerbero-7B: A Leap Forward in Language-Specific LLMs Through Enhanced Chat Corpus Generation and Evaluation, 2023. arXiv:2311.15698.

[12] 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.

[13] M. Polignano, P. Basile, G. Semeraro, Advanced Natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA, 2024. arXiv:2405.07101.

[14] T. Bocklisch, J. Faulkner, N. Pawlowski, A. Nichol, Rasa: Open Source Language Understanding and Dialogue Management, CoRR abs/1712.05181 (2017). arXiv:1712.05181.

[15] T. Bunk, D. Varshneya, V. Vlasov, A. Nichol, DIET: Lightweight Language Understanding for Dialogue Systems, CoRR abs/2004.09936 (2020). arXiv:2004.09936.

[16] A. Mazzei, L. Anselma, M. Sanguinetti, A. Rapp, D. Mana, M. M. Hossain, V. Patti, R. Simeoni, L. Longo, Anticipating User Intentions in Customer Care Dialogue Systems, IEEE Transactions on Human-Machine Systems (2022). doi:10.1109/THMS.2022.3184400.

[17] G. Marzi, M. Balzano, D. Marchiori, K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient, MethodsX 12 (2024) 102545. doi:https://doi.org/10.1016/j.mex.2023.102545.

[18] A. Santilli, E. Rodolà, Camoscio: An Italian Instruction-tuned LLaMA, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023, volume 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

# A. Prompt Examples

## A.1. Example with response generated by LLaMAntino2

**Prompt** <s>[INST] «SYS» Sei un an assistente AI per la lingua italiana. Rispondi nella lingua usata per la domanda in modo chiaro, diretto e completo. Attieniti strettamente alle istruzioni fornite e riporta la tua risposta nel formato richiesto, senza aggiungere ulteriori commenti o spiegazioni. «/SYS»

Considera le seguenti informazioni relative all'andamento di un impianto fotovoltaico:
- potenza prodotta ora dall'impianto fotovoltaico: 0.0 kW
- potenza utilizzata ora dalla casa: 0.23 kW
- stato di carica attuale della batteria: 55%
- status della batteria: in scarica
- potenza immessa in rete: 0.15 kW
- potenza fornita dalla batteria: 0.38 kW
- energia utilizzata dalla casa in tutta la giornata: 14.45 kWh
- energia totale prodotta dai pannelli in tutta la giornata: 36.909 kWh
- energia acquistata dalla rete in tutta la giornata: 0.55 kWh
- auto-consumo della giornata: 13.90 kWh
- energia immessa in rete in tutta la giornata: 23.01 kWh

Riassumi le informazioni che ti ho appena fornito per rispondere alla seguente domanda: voglio avere un resoconto della produzione energetica dei pannelli solari [/INST]

**Response** "*Mi dispiace, non posso fornire una risposta perché la vostra domanda non è specifica. Per ottenere una risposta precisa, è necessario fornire ulteriori informazioni.*"

| | | | | | |
|---|---|---|---|---|---|
| **Evaluation** | Usefulness | 1 | 1 | 1 | 1 |
| | Necessity | 1 | 1 | 1 | 1 |
| | Understandability | 1 | 7 | 1 | 1 |
| | Fluency | 7 | 7 | 1 | 6 |
| | Accuracy | 1 | 2 | 1 | 1 |

## A.2. Example with response generated by Zefiro

**Prompt** <|system|> Sei un an assistente AI per la lingua italiana. Rispondi nella lingua usata per la domanda in modo chiaro, diretto e completo. Attieniti strettamente alle istruzioni fornite e riporta la tua risposta nel formato richiesto, senza aggiungere ulteriori commenti o spiegazioni.</s> <|user|> Considera le seguenti informazioni relative all'andamento di un impianto fotovoltaico:
- potenza prodotta ora dall'impianto fotovoltaico: 0.0 kW
- potenza utilizzata ora dalla casa: 0.36 kW
- stato di carica attuale della batteria: 53%
- status della batteria: in scarica
- potenza fornita dalla batteria: 0.32 kW
- potenza acquistata dalla rete: 0.04 kW
- energia utilizzata dalla casa in tutta la giornata: 14.45 kWh
- energia totale prodotta dai pannelli in tutta la giornata: 36.909 kWh
- energia acquistata dalla rete in tutta la giornata: 0.55 kWh
- auto-consumo della giornata: 13.90 kWh
- energia immessa in rete in tutta la giornata: 23.01 kWh

Riassumi le informazioni che ti ho appena fornito per rispondere alla seguente domanda: livello di carica della batteria al momento</s> <|assistant|>

**Response** "*La batteria ha un livello di carica del 53% al momento, che indica che ci sono circa 53% di carica nella batteria.*"

| | | | | | |
|---|---|---|---|---|---|
| **Evaluation** | Usefulness | 7 | 6 | 6 | 5 |
| | Necessity | 7 | 7 | 6 | 6 |
| | Understandability | 5 | 7 | 7 | 3 |
| | Fluency | 4 | 6 | 6 | 4 |
| | Accuracy | 7 | 7 | 7 | 6 |