GEESE - Generating and Evaluating Explanations for Semantic Entailment: A CALAMITA Challenge

Andrea Zaninello[1,2,*], Bernardo Magnini[1]

[1]Fondazione Bruno Kessler, Trento (Italy)
[2]Free University of Bozen-Bolzano (Italy)

Abstract
In the GEESE challenge, we present a pipeline to evaluate generated explanations for the task of Recognizing Textual Entailment (RTE) in Italian. The challenge focuses on evaluating the impact of generated explanations on the predictive performance of language models. Using a dataset enriched with human-written explanations, we employ two large language models (LLMs) to generate and utilize explanations for semantic relationships between sentence pairs. Our methodology assesses the quality of generated explanations by measuring changes in prediction accuracy when explanations are provided. Through reproducible experimentation, we establish benchmarks against various baseline approaches, demonstrating the potential of explanation injection to enhance model interpretability and performance.

Keywords
CALAMITA, CLiC-it, Explanation generation, Explainability, RTE, Recognizing Textual Entailment, Inference, Italian

1. Introduction and Motivation

The ability of a machine to justify its predictions and provide human-understandable explanations has been a key research objective of Machine Learning (ML) and Artificial Intelligence (AI) since their early stages [1, 2, 3]. In the past few years, the field of AI has experienced an unprecedented acceleration in most areas, such as computer vision [4], audio [5], video [6], and programming languages [7], and especially in Natural Language Processing (NLP), with the popularization of generative Large Language Models (LLMs) such as OpenAI's ChatGPT [8], Google's Gemini [9], or Meta's Llama [10].

These models are currently able to produce natural-sounding and coherent language, often indistinguishable from natural language [11, 12]. While these results open up new avenues for future applications and research, they also raise ethical issues considering the ubiquitous role of machines in our lives, and in sensitive fields like education, health, justice, and private life. In fact, the scarce transparency of neural architectures makes it hard to interpret their functioning (the so-called "black-box" problem). In addition, many of the currently available LLMs are not fully open-source, so the data they were trained on is not known to either researchers or the general public. Finally, these models have achieved such sizes that their results are difficult to replicate, making them a kind of "black box in a black box".

As a consequence, the need to develop methods to understand their reasoning is becoming central. Many recent efforts have been devoted to explaining such models [13], and the importance of interpretability and explainability in AI has become ever more urgent [14, 15, 16].

The role of explanations in NLP has been explored by a consistent body of research. Cambria et al. [17], for instance, provides a comprehensive survey of approaches for generating natural language explanations; Hartmann and Sonntag [18] examines the benefits of explanations for NLP models; Paranjape et al. [19] focuses on template-based explanations, Lampinen et al. [20] and Ye and Durrett [21] demonstrate the benefits of in-context explanations for large models in challenging reasoning tasks.

Explanation generation quality has traditionally been evaluated through automated *ovelap* metrics like BLEU [22], ROUGE [23], or BERT-Score [24] against a gold reference explanation written by humans. This usually implies costly human-explanation collection campaigns; additionally, these measures may neither fully capture the informativity or the effectiveness of an explanation, nor faithfully reflect human judgments.

Recently, human *simulatability* scores have been proposed as an alternative method to understand the quality of explanations from the perspective of the "utility to an end-user" [25]. Rather than focusing on the overlap between explanations and ground-truth data, this approach assesses how *explanations enhance predictive performance on a downstream task* compared to the input alone. While humans have traditionally been the predictors [26], recent research has demonstrated that trained models can automate this process, showing moderate to strong correlations with human judgments [27]. Pruthi et al. [28], for instance, measures explanation quality

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
azaninello@fbk.eu (A. Zaninello); magnini@fbk.eu (B. Magnini)
https://github.com/andreazaninello (A. Zaninello)
0000-0001-9998-1942 (A. Zaninello)

based on downstream performance: their methodology involves training a student model on explanations generated by a teacher, using automatic explanation generation techniques and training the student for the end task.

However, current LLMs may also benefit from explanation injection even if they are not explicitly trained to do so, and some works suggest using the explanation to augment the input to condition predictions of future data points on both the input *and* the explanation [29, 27]. In fact, LLMs are capable of understanding supplementary input content and including explanations in the input during inference without requiring additional supervision, which can indirectly demonstrate the role of explanations in the inference process.

These observations underline two crucial aspects:

- providing LLMs with quality explanations that allow them to *infer relevant latent information*, i.e. to provide additional background knowledge, improves performance compared to only using the input or to using spurious explanations;
- the quality of a (human or machine-generated) explanation can be measured based on its *helpfulness* (or impairment) to the (model's or human's) performance on a downstream task.

To contribute to this line of research, we propose **GEESE: Generating and Evaluating Explanations for Semantic Entailment** at CALAMITA [30], a pipeline to indirectly assess the effectiveness of explanations through the evaluation of their impact on the task of Recognizing Textual Entailment (RTE) in Italian[1].

2. Task Description and GEESE Explanatory Pipeline

Consider a pair of sentences $< s_1, s_2 >$, like the ones in the following example:

(1) *Il cielo è grigio oggi.*

(2) *Faresti bene a prendere l'ombrello.*[2]

Consider a semantic relation r holding between s_1 and s_2 (e.g., s_1 entails s_2, s_1 does not entail s_2, s_1 contradicts s_2). Let E be the set of possible explanations for r. GEESE's explanatory task consists in:

- generating an explanation $e_r \in E$ for the semantic relationship r for each $< s_1, s_2 >$ in the dataset;
- predict the relation with and without the generated explanation e_r;
- assess the quality of the generated explanations E_{gen} by taking the delta between prediction accuracy with and without explanation as a proxy of explanations' quality.

Step 1: Generate Explanation: A first LLM (M_1) is prompted to produce explanations $E_{gen} = \{e_1, e_2, ... e_n\}$ for a specific semantic relation r_c holding between a given sentence pair, denoted as $< s_1, s_2 >$. In the task, we focus on the entailment relationship, which can take three values: "YES" (sentence 1 is entailed by sentence 2), "NO" (sentence 1 is contradicted by sentence 2), "UNKNOWN" (sentence 1 is neither entailed nor contradicted by sentence 2). In our baselines, we focus on one explanation type (why-explanation), but other kinds of explanations or reasoning strategies (like counterfactual or example-based ones) are possible. In our baselines, we use llama-3-3B-instruct [31] as M_1.

Step 2: Use Explanation on Relation Prediction: A second LLM (M_2) is then provided with the generated explanations E_{gen} to evaluate if the generated explanations improve the task of predicting the correct relations. In practice, this is achieved by appending the explanation as a "hint" to the prompt, and asking the model to make a prediction thereof. This process aims to discover how effectively M_2 leverages the explanations from M_1 to perform the target task. We use llama-3-8B as M_2, but other combinations of M_1 and M_2 are possible.

Step 3: Evaluate Explanation Effectiveness Explanation effectiveness is evaluated by analyzing how providing different explanations generated in Step 1 affects the model M_2 prediction in Step 2. In practice, this is done by calculating the accuracy of the predictions of M_2 given the explanations and comparing them to the selected baselines (see Section 4).

3. Data description

3.1. Origin of data

The Recognizing Textual Entailment (RTE) task emerged in 2005 [32] as the problem of determining if two sentences stand in an *entailment* or *not-entailment* relationship. A common definition of "semantic entailment" (also referred to as *presupposition* in some studies) is that "A sentence S presupposes a proposition p if p must be true in order for S to have a truth-value (to be true or false)" [33]. A text t is said to entail another text (*hypothesis*, h) if h is true in every circumstance (possible world) in which t is true. RTE, however, suggests a more empirical definition, allowing for cases in which the truth of

[1] Code and data are made available at github.com/andreazaninello/calamita-geese

[2] (1) The sky is grey today. (2) You better take your umbrella with you.

the hypothesis is *highly plausible, for most practical purposes*, rather than certain. According to [34], this "shallow" definition better accounts for the types of uncertain inferences that are typically expected from text-based applications.

Recognizing Textual Entailment was formalized through a series of successful challenges and workshops that began in 2005 [32] and lasted until 2012. Starting from the RTE-3 edition, the task was extended from two labels to a three-label classification, splitting the not-entailment label into two classes, *contradiction* and *neutrality*. Given the interest in the task, an Italian version of the RTE-3 dataset was developed to explore language comprehension and textual entailment [35].

The dataset used in the challenge is the e-RTE-3-it dataset [36], which is an emended version enriched with human-written explanations of the RTE-3-it dataset [35].

3.2. Detailed data statistics

The dataset contains 1600 text-hypothesis sentence pairs in Italian (`text_t` and `text_h` in the dataset) divided into an 800-example validation and an 800-example test split. Each example is annotated with an entailment label (`label`): `"YES"` (entailed), `"NO"` (contradicted), or `"UNKNOWN"` (neutral).

3.3. Annotation details

The *e-RTE-3-it* dataset presents human explanations written in Italian by native speakers. For each text-hypothesis pair, annotators provided a *natural language explanation* justifying the given label (`explanation`) for the entailment relation ("why does S_1 stand in an r relation with S_2?")[3].

All annotations underwent quality control, involving two expert linguists who manually checked the explanations for grammaticality, fluency, and logical validity. This process ensured high quality of the final e-RTE-3-it explanations, informativeness, as well as minimal label leakage (see *infra*).

Label leakage [37] refers to the fact that the explanation may be directly suggesting the label without genuinely being informative. While the manual check of all original human explanations ensured minimal label leakage, to prevent this we automatically replace direct references to the label and to the task with placeholders in the human-written and generated explanations. In our implementation, this is done through regular expressions by substituting ("anonimize") the label strings (`"YES"`, `"NO"`, `"UNKNOWN"`) and all words starting with `entail.*`, `contradict.*`, `neutr.*`, `impl*`, `contradd.*` (verbs and nouns directly stating the kind of relationship) with "XXX".

We therefore also provide the following "anonymized" additional explanations for each example, which we use in our prompts:

- `anon_whyexp`: the anonimized explanation generated by llama3 as M_1;
- `anon_human`: the anonimized human-written explanation (from e-RTE-3-it).

3.4. Data format

The dataset is freely distributed in HuggingFace's Dataset format[4]. A snippet of the data is displayed in Table 1.

4. Metrics and baselines

We conduct baseline experiments using Llama-3.1-8B-Instruct as M_1 with a custom implementation in HuggingFace, and Llama-3-8B as M_2, using the LLM-Evaluation-Harness library [38] in a zero-shot setting[5].

We provide baselines for the following settings:

1. **no-exp**: No explanations provided (baseline);
2. **dummy**: The hypothesis itself (`text_t`) provided as a "non-informative" explanation, controlling for input length and providing a second baseline.
3. **human**: Human-written explanations (from e-RTE-3-it) anonimized (`anon_human`) provided as additional input;
4. **llama-3**: The explanation generated using LLama-3-8B-Instruct as M_1 (`anon_whyexp`).

4.1. Example of prompts for zero shots

All experiments have been carried out in a zero-shot setting using the following prompts[6].

(**M1 - Generation**):
```
Your task is to clarify the entailment relationship between a pair of sentences by explaining why a classifier predicted
```

[3] Additionally, the annotator provided a *confidence score* (1-5) reflecting their certainty about the provided explanation (which we don't use in the task), an optional *alternative label*, if they felt the initial label was inaccurate, along with *explanations* and *confidence scores*. We don't consider these annotations in the task, and only use the original label as our gold relationship and the human-written explanation for the original label as a strong baseline.

[4] https://huggingface.co/datasets/azaninello/explained-full-llama-3

[5] Generation parameters are: stop sequences: "</s>", "<|eot_id|>", max. gen. tokens: 128; do sample: false; temperature: 0

[6] All prompts are provided in English to ensure comparability across tasks and language models. In fact, using English as a "metalanguage" allows us to establish a comparable set-up to study LLMs' capability regardless of their specific training language.

FIELD	EXAMPLE
id	393
text_t	Gli Eschimesi, i cacciatori di foche sparsi nell'Artico, prevedono di sollecitare un giudizio della Commissione Interamericana dei Diritti dell'Uomo poiché gli Stati Uniti, contribuendo in modo sostanziale al riscaldamento globale, stanno minacciando la loro esistenza.
text_h	La caccia alle foche mette a rischio d'estinzione le specie.
label	UNKNOWN
text_label	neutrality
explanation	Nella frase non vengono citate le conseguenze della caccia sulle foche perciò non si può evincere il rischio per loro della caccia.
anon_whyexp	The XXX relationship between the two sentences is XXX because the relationship is not clearly established. Sentence 1 mentions the Inuit people, the Commission on Human Rights, and the United States, which are not directly related to the topic of the second sentence, which is about the threat of extinction of species due to hunting. The connection between the two sentences is not explicitly stated, and it is not clear how the information presented in Sentence 1 supports or XXX the information in Sentence 2. Therefore, the XXX label is XXX because the relationship between the two sentences is not clear. For example, if the XXX label was XXX, it would mean that Sentence 1 logically implies Sentence 2, but the information in Sentence 1 does not provide enough evidence to support this claim. Similarly, if the XXX label was XXX, it would mean that Sentence 1 logically XXX Sentence 2, but again, the information in Sentence 1 does not provide enough evidence to support this claim. Therefore, the XXX label is XXX because the relationship between the two sentences is not clear.
anon_human	Nella frase non vengono citate le conseguenze della caccia sulle foche perciò non si può evincere il rischio per loro della caccia.

Table 1
Example from the processed e-RTE-3-it dataset.

a specific entailment label. Sentence 1: text_t Sentence 2: text_h Entailment label: label. exp_type[7]

(**M2 - Prediction**): Your task is to predict the entailment label between two sentences, selecting one label among YES (entailment), NO (contradiction), or UNKNOWN (neutrality). Sentence 1: text_t Sentence 2: text_h Hint: anon_explanation. Entailment label:[8]

5. Baseline Results and Discussion

Baseline results, reported in Table 2, demonstrate the impact of incorporating explanations on the performance of language models in the Recognizing Textual Entailment tasks. The accuracy scores indicate that models utilizing explanations generated by Llama-3 achieve the highest accuracy at 78.12%. In comparison, using human-written explanations shows slightly lower accuracy compared to machine-generated, but higher scores compared to baselines, suggesting that explanations do enhance the models' understanding of semantic relationships.

Generated explanations, proving more effective than human-crafted ones, suggest that the quality and type of explanations provided can influence predictive performance, but also highlight the need for further research into optimizing explanation generation methods for improved outcomes in NLP tasks. In fact, note that generated explanations may be positively influenced by factors other than informativeness alone, such as the lengths of the explanations themselves, or may still be indirectly suggesting the right relationship despite the anonymization process described in 3.3.

For example, as reported by one of the anonymous reviewers, see "anon_whyexp" explanation in Table 1: "In other words, Sentence 2 **provides enough information to infer the truth** of Sentence 1". The generated explanation clearly (but not directly) hints at an "entail" label, potentially compromising the intended anonymity. The fairness of the comparison between human- and machine-generated explanation is an aspect that deserves further investigation.

[7]Variables are indicated in color. In our experiments exp_type = "Explain how the two sentences are connected." and the variables are read from each example.
[8]Variables are indicated in color. In our experiments, anon_explanation can take the following values: "Not given." (**no-exp**), text_h (**dummy**), anon_human (**human**), anon_whyexp (**llama-3**).

Tasks	n-shot	Metric	Value	Stderr
geese_dummy	0	acc	0.5850	0.0174
geese_noexp	0	acc	0.5437	0.0176
geese_llama3	0	acc	**0.7812**	0.0146
geese_human	0	acc	0.7575	0.0152

Table 2
Results for the 0-shot baseline experiments on the full test set.

6. Conclusion

The findings from the GEESE challenge underscore the significance of effective explanation generation in enhancing the capabilities of language models in RTE tasks. Preliminary results show that models provided with explanations, whether human-written or generated by LLMs, exhibit improved predictive accuracy compared to those lacking such inputs. This supports the hypothesis that explanations can facilitate a deeper understanding of semantic relationships, thus aiding model inference.

The GEESE challenge establishes a framework for generating and evaluating explanations in the domain of semantic entailment. By demonstrating the utility of explanation injection, we contribute to the ongoing discourse on interpretability in AI, advocating for a balanced approach that enhances model transparency while maintaining robustness. Our findings encourage further exploration into the interplay between explanations and model performance, paving the way for more interpretable and user-friendly AI systems. As language models continue to evolve, integrating effective explanation mechanisms will be crucial for ensuring their responsible deployment in sensitive applications.

7. Limitations

The study also highlights limitations, including potential biases in the generated explanations and the challenge of ensuring that explanations remain informative without directly revealing the answer. Future research could explore diverse explanation types and their varying impacts across different contexts and languages.

8. Ethical issues

We would like to draw the readers' attention on the following. Firstly, the potential for bias in both the training data and the generated explanations can perpetuate stereotypes or misinformation, leading to harmful consequences, particularly in sensitive domains such as healthcare or legal applications. There is also the risk that users may place undue trust in machine-generated explanations, mistakenly believing them to be infallible. Finally, the collection and use of data for training these models must adhere to strict privacy standards to ensure that individuals' rights are respected. Addressing these ethical challenges is essential to foster trust and ensure that AI technologies are developed and used responsibly.

9. Data license and copyright issues

We release our original content under the MIT License. Please refer to the original dataset's copyright and license regulations for information on the derived data.

Acknowledgments

This work has been partially funded by PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and the ANTIDOTE project (CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002)

References

[1] S. Lowry, G. Macpherson, A blot on the profession, 296 brit, MED. J 657 (1988) 657.

[2] L. M. Fagan, E. H. Shortliffe, B. G. Buchanan, Computer-based medical decision making: from mycin to vm, Automedica 3 (1980) 97–108.

[3] R. Bareiss, Exemplar-based knowledge acquisition: A unified approach to concept representation, classification, and learning, volume 2, Academic Press, 2014.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2021. arXiv:2112.10752.

[5] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg, B. Ramabhadran, Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning, CoRR abs/1907.04448 (2019). URL: http://arxiv.org/abs/1907.04448. arXiv:1907.04448.

[6] Y. Mirsky, W. Lee, The creation and detection of deepfakes: A survey, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3425780. doi:10.1145/3425780.

[7] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such,

D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, W. Zaremba, Evaluating large language models trained on code, CoRR abs/2107.03374 (2021). URL: https://arxiv.org/abs/2107.03374. arXiv:2107.03374.

[8] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[9] G. Team, Gemini: A family of highly capable multimodal models, 2024. URL: https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] T. Labruna, S. Brenna, A. Zaninello, B. Magnini, Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations, 2023. arXiv:2305.14556.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[14] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5. doi:10.1017/S0269888921000011.

[15] A. D. Selbst, J. Powles, Meaningful information and the right to explanation, International Data Privacy Law 7 (2017) 233–242. URL: https://doi.org/10.1093/idpl/ipx022. doi:10.1093/idpl/ipx022. arXiv:https://academic.oup.com/idpl/article-pdf/7/4/233/22923065/ipx022.pdf.

[16] L. Edwards, M. Veale, Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, Duke L. & Tech. Rev. 16 (2017) 18.

[17] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on xai and natural language explanations, Information Processing Management 60 (2023) 103111. URL: https://www.sciencedirect.com/science/article/pii/S0306457322002126. doi:https://doi.org/10.1016/j.ipm.2022.103111.

[18] M. Hartmann, D. Sonntag, A survey on improving NLP models with human explanations, in: Proceedings of the First Workshop on Learning with Natural Language Supervision, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 40–47. URL: https://aclanthology.org/2022.lnls-1.5. doi:10.18653/v1/2022.lnls-1.5.

[19] B. Paranjape, J. Michael, M. Ghazvininejad, H. Hajishirzi, L. Zettlemoyer, Prompting contrastive explanations for commonsense reasoning tasks, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4179–4192. URL: https://aclanthology.org/2021.findings-acl.366. doi:10.18653/v1/2021.findings-acl.366.

[20] A. Lampinen, I. Dasgupta, S. Chan, K. Mathewson, M. Tessler, A. Creswell, J. McClelland, J. Wang, F. Hill, Can language models learn from explanations in context?, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 537–563. URL: https://aclanthology.org/2022.findings-emnlp.38. doi:10.18653/v1/2022.findings-emnlp.38.

[21] X. Ye, G. Durrett, The unreliability of explanations in few-shot prompting for textual reasoning, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 30378–30392. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c402501846f9fe03e2cac015b3f0e6b1-Paper-Conference.pdf.

[22] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[23] L. C. ROUGE, A package for automatic evaluation

of summaries, in: Proceedings of Workshop on Text Summarization of ACL, Spain, 2004.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[25] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Advances in Neural Information Processing Systems, volume 29, 2016.

[26] S. Wiegreffe, A. Marasović, N. A. Smith, Measuring association between labels and free-text rationales, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10266–10284. URL: https://aclanthology.org/2021.emnlp-main.804. doi:10.18653/v1/2021.emnlp-main.804.

[27] P. Hase, S. Zhang, H. Xie, M. Bansal, Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4351–4367. URL: https://aclanthology.org/2020.findings-emnlp.390. doi:10.18653/v1/2020.findings-emnlp.390.

[28] D. Pruthi, R. Bansal, B. Dhingra, L. B. Soares, M. Collins, Z. C. Lipton, G. Neubig, W. W. Cohen, Evaluating explanations: How much do explanations from the teacher aid students?, Transactions of the Association for Computational Linguistics 10 (2022) 359–375. URL: https://aclanthology.org/2022.tacl-1.21. doi:10.1162/tacl_a_00465.

[29] P. Hase, M. Bansal, When can models learn from explanations? a formal framework for understanding the roles of explanation data, arXiv preprint arXiv:2102.02201 (2021).

[30] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[31] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. E. Tan, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Grattafiori, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Vaughan, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Franco, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer,

D. Civin, D. Beaty, D. Kreymer, D. Li, D. Wyatt, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Ozgenel, F. Caggioni, F. Guzmán, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Thattai, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, I. Damlaj, I. Molybog, I. Tufanov, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Prasad, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Huang, K. Chawla, K. Lakhotia, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Tsimpoukelli, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. P. Laptev, N. Dong, N. Zhang, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Li, R. Hogan, R. Battey, R. Wang, R. Maheswari, R. Howes, R. Rinott, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Kohler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Albiero, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wang, X. Wu, X. Wang, X. Xia, X. Wu, X. Gao, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Hao, Y. Qian, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[32] I. Dagan, O. Glickman, B. Magnini, The pascal recognising textual entailment challenge, in: Machine learning challenges workshop, Springer, 2005, pp. 177–190.

[33] G. Chierchia, S. Mcconnell-Ginet, Meaning and grammar: An introduction to semantics, 1990. URL: https://api.semanticscholar.org/CorpusID:62731986.

[34] I. Dagan, B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment: Rational, evaluation and approaches–erratum, Natural Language Engineering 16 (2010) 105–105.

[35] B. Magnini, A. Lavelli, S. Magnolini, Comparing machine learning and deep learning approaches on NLP tasks for the Italian language, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2110–2119. URL: https://aclanthology.org/2020.lrec-1.259.

[36] A. Zaninello, S. Brenna, B. Magnini, Textual entailment with natural language explanations: The italian e-rte-3 dataset, in: CLiC-it, 2023. URL: https://ceur-ws.org/Vol-3596/short21.pdf.

[37] P. Hase, S. Zhang, H. Xie, M. Bansal, Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4351–4367. URL: https://aclanthology.org/2020.findings-emnlp.390. doi:10.18653/v1/2020.findings-emnlp.390.

[38] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2024. URL: https://zenodo.org/records/12608602. doi:10.5281/zenodo.12608602.