ABRICOT 🍑 - ABstRactness and Inclusiveness in COntexT: A CALAMITA Challenge

Giovanni Puccetti[1,*], Claudia Collacciani[2], Andrea Amelio Ravelli[3], Andrea Esuli[1] and Marianna Marcella Bolognesi[3]

[1]Istituto di Scienza e Tecnologia dell'Informazione "A. Faedo"
[2]Independent researcher
[3]ABSTRACTION Research Group – Università di Bologna

Abstract

The ABRICOT Task is designed to evaluate Italian language models on their ability to understand and assess the abstractness and inclusiveness of language, two nuanced features that humans naturally convey in everyday communication. Unlike binary categorizations such as abstract/concrete or inclusive/exclusive, these features exist on a continuous spectrum with varying degrees of intensity. The task is based on a manual collection of sentences that present the same noun phrase (NP) in different contexts, allowing its interpretation to vary between the extremes of abstractness and inclusiveness. This challenge aims to verify the how LLMs perceive subtle linguistic variations and their implications in natural language.

Keywords

Abstraction, Inclusiveness, Context, LLM evaluation, Italian Language Models

1. Challenge: Introduction and Motivation

The ability to convey both specific information (about individuals or events) and generalisations (about categories) with the same lexical item is one of the key feature of natural languages. Consider the examples in 1:

1. a) **the lion** escaped yesterday from the zoo.
   b) **the lion** is a predatory cat.

The noun phrase (NP) *the lion* can describe either a specific individual (1a) or the entire category of large African felines (1b), thus it expresses a variable degree of inclusiveness of the possible number of individuals to which the NP correctly applies in each sentence it occurs. This demonstrates how human language follows a principle of economy, enabling a one-to-many mapping between lexical labels and meanings.

The syntactic form of the NP (definite, indefinite, or plural) does not provide sufficient information to discriminate between the two meanings, and we need to enlarge our focus to take into account the whole context in which the NP occurs [1]. This phenomenon can be observed in all languages [2], affecting nearly all nouns that can be used in referring expressions. Indeed, natural languages do not have explicit markers for generic NPs [3]; the genericity/specificity of an NP is derived from the meaning of the entire sentence. In other words, we cannot interpret language one word at a time; we need to consider the whole sentence or utterance as context to disambiguate and decipher the meaning of each single word composing it, and thus to understand the message conveyed through language.

Generalizations about kinds and categories, as in 1b, are called *generics* and are fundamental to human cognition, because they allow us to conceptualize properties linked to categories, shaping how we perceive the world [4].

Moreover, distinguishing between generic and non-generic meanings for abstract entities is less straightforward than for concrete ones, and for this reason evaluate the inclusiveness of an abstract noun or a NP is even more challenging. Indeed, inclusiveness is not an exclusive feature of concrete-only entities. Consider the examples in 1:

2. a) Colorless green **ideas** sleep furiously.
   b) Be less curious about people and more curious about **ideas**.

The concept behind the word *idea* is always referring to an abstract entity, with slightly different grades of abstractness, but it shows a greater variation in terms of inclusiveness. The noun *ideas* in 2a includes only a restricted number of elements with respect to the universe

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
giovanni.puccetti@isti.cnr.it (G. Puccetti); claudia.collacciani2@unibo.it (C. Collacciani); andreamelio.ravelli@unibo.it (A. A. Ravelli); andrea.esuli@isti.cnr.it (A. Esuli); m.bolognesi@unibo.it (M. M. Bolognesi)
https://gpucce.github.io/ (G. Puccetti); https://github.com/claudiacollacciani (C. Collacciani); https://www.unibo.it/sitoweb/andreaamelio.ravelli (A. A. Ravelli); https://esuli.it/ (A. Esuli); https://www.unibo.it/sitoweb/m.bolognesi (M. M. Bolognesi)

Token: Margherita
Text: Le margherite di fronte alla mia casa saranno in piena fioritura.
Abstractness: 0.177
Inclusiveness: 0.187

(a) Example of sample for the *Margherita* token.

Token: Ambizione
Text: La sua ambizione lo rovinerà.
Abstractness: 0.478
Inclusiveness: 0.083

(b) Example of sample for the *Ambizione* token.

Token: Benzina
Text: La benzina è nella bottiglia del latte.
Abstractness: 0.064
Inclusiveness: 0.063

(c) Example of sample for a more concrete *Benzina* token.

Token: Benzina
Text: In Italia è disponibile la benzina a 95 ottani.
Abstractness: 0.573
Inclusiveness: 0.653

(d) Example of sample for a more abstract *Benzina* token.

Figure 1: Examples from the abricot dataset.

of the ideas (namely, only *colorless green* ones), while the reference in 2b shows a higher level of inclusiveness, not distinguishing among them on the basis of their color.

The ability to distinguish, interpret and use correctly the variability that natural language offers along these two graduated semantic features, abstractness and inclusiveness, is of paramount importance if we want to make *talking machines* which not only simulate language, but can also *reason* about natural language and the knowledge of the world it depicts.

The CALAMITA special event [5] offers the possibility to challenge Large Language Models on their ability to understand the abstractness and inclusiveness of the words, and compare with humans their behaviour in judging Italian sentences. With this report we present the ABRICOT 🍑 Task: ABstRactness and Inclusiveness in COntexT.

2. Challenge: Description

The ABRICOT 🍑 Task aims to challenge Italian language models on their understanding of abstractness and inclusiveness, features that we, as humans, naturally express in everyday language. These features are not discrete binary dichotomies like `abstract/concrete` or `inclusive/exclusive`; instead, they shade on a continuous spectrum, with the two extremes at opposite ends. The collection of sentences in this Task shows the same NP in a variety of different contexts, so that its meaning can oscillate between the extremes of both the axis of abstractness and inclusiveness.

We ask the participant models to express a judgment on a 5 point Likert scale for both the features of inclusiveness and abstractness of the target noun or NP in each sentence.

This task have some similarities with the CONcreTEXT Task[1] [6], which has been presented at the 2020 edition of EVALITA.[2] Both tasks focus on the abstractness/concreteness of target words in natural Italian sentences, asking judgments by means of Likert scales, but the ABRICOT 🍑 Task goes beyond by including also the inclusiveness feature of the targets. Moreover, for the construction of this dataset we considered exclusively nouns or NPs as targets, and in order to limit to the minimum the impact of the variability deriving from different semantic role or syntactic function, all the sentences have been selected with the target noun as subject of the main verb.

2.1. Tasks

We propose two separate tasks for this benchmark, Task 1: *abstractness* and Task 2: *inclusiveness* the two tasks are formally identical, we use the same metric and the same samples, however they measure two different scores, respectively *abstractness_mean* and *inclusiveness_mean*, the first meant to measure the abstractness of the word in context and the second its inclusiveness.

Since both these concepts are evident but fuzzy also for humans, we don´t expect language models to have a perfect understanding of them and we will limit our metrics to regression ones. Despite the tasks being very similar from a formal perspective, we show how models' performance on these two tasks varies and there is sensible difference between the results in the two tasks.

[1] lablita.github.io/CONcreTEXT
[2] www.evalita.it

3. Data description

3.1. Origin of data

The 20 target NPs of the dataset for the ABRICOT 🍑 Task are derived (and translated in Italian) from the set of target nouns in the Situation Entities Corpus (SitEnt [7]), a collection of English sentences in which specificity and genericity have been annotated with a binary labelling scheme (i.e., `GENERIC` vs. `NON-GENERIC`). Using those as seeds, representative Italian sentences have been manually harvested from OpenSubtitles[3] and WikiHow.[4] These are widely used sources, the first contains the openly available subtitles of an extensive collection of movies and TV series, while the second is a website gathering articles on *how-to* do a variety of different things.

More specifically, the sentences have been extracted from the Italian section of the multilingual The Human Instruction Dataset [8], a structured collection of WikiHow instructions pages, and from the Italian sub-corpus of the OpenSubtitles2018 corpus [9].

Our protocol proposes to the annotators groups of sentences (from a minimum of 4 to a maximum of 8), all containing the same noun, each to be evaluated using a continuous slider, from which values ranging from 0 to 1 will then be extracted.

After the annotation, the reliability of our data has been computed using the Intraclass Correlation Coefficient (ICC(k)). Human ratings have been then averaged, and the resulting figures will be used as gold standard.

An example of the samples present in the dataset can be seen in Figure 1 where examples with the NPs *margherita* (lilly), *ambizione* (ambition) and *benzina* (gasoline) are reported. In particular, Figure **??** and 1d show two examples containing the same token but in different contexts and report the effect of the context on the abstractness and inclusiveness of the token.

The data is stored on OSF [10].[5]

3.2. Data format

The data is proposed in a tabular format, with 12 columns:

- *ID*: a unique identifier for the sample;
- *target token*: the focus of the dataset, to be assinged an abstraction score in context;
- *target lemma*: the lemma of the target token;
- *text*: the sentence where the token appears;
- *begin*: the index of the first character of the token in the sentence;
- *end*: the index of the last character of the token in the sentence;
- *domain*: the source where the token come from;
- *inclusiveness mean*: the average inclusiveness score assigned by the annotators;
- *inclusiveness std*: the standard deviation of the inclusiveness scores;
- *abstractness mean*: the average abstractness score assigned by the annotators;
- *abstractness std*: the standard deviatio n of the abstractness scores;

[3] https://www.opensubtitles.org
[4] https://www.wikihow.com
[5] https://osf.io/ja89x/?view_only=91d683c7399c45f9aa63f2b34cfe6617

Abstractness Prompt:
Assegna un valore di astrazione da 1 a 5 alla parola parola nel contesto della frase seguente: {frase} Descrizione dei valori: 1 - La parola è estremamente concreta (e.g. un cane specifico) 2 - La parola è lievemente concreta (e.g. un cane di una certa razza) 3 - La parola è neutra (e.g. un cane tra tanti) 4 - La prola è lievemente astratta (e.g. un cane è un animale da compagnia) 5 - La parola è estremamente astratta (e.g. il cane è un mammifero).

(a) Prompt used for the Inclusiveness Task.

Inclusiveness Prompt:
Assegna un valore di inclusività da 1 a 5 alla parola parola nel contesto della frase seguente: {frase} Descrizione dei valori: 1 - La parola è estremamente specifica (e.g. un cane specifico) 2 - La prola è lievemente specifica (e.g. un cane di una certa razza) 3 - La parola è neutra (e.g. un cane tra tanti) 4 - La parola è lievemente inclusiva (e.g. un cane è un animale da compagnia) 5 - La parola è estremamente inclusiva (e.g. il cane è un mammifero)

(b) Prompt used for the Inclusiveness Task.

Figure 2: Prompts used for the evaluation.

3.3. Example of prompts used for zero or/and few shots

We use different prompts for the two tasks, they are shown in Figure 2, we ask the model to directly output a score from 1 to 5 specific to the task, we then propose an explanation for each point from 1 to 5, explaining the (approximate) meaning of assigning that score together with a very high-level example and on top of the explanation, we use 3-shot evaluation, we found 0-shot to be difficult

		ambizione	benzina	bicchiere	bici	bottiglia	cameriere	coscienza	effetto	farina	giardino
abstractness	mean	0.65	0.42	0.51	0.52	0.34	0.47	0.81	0.57	0.46	0.50
	std	0.18	0.26	0.19	0.27	0.26	0.22	0.06	0.24	0.26	0.29
inclusiveness	mean	0.41	0.48	0.52	0.58	0.35	0.42	0.53	0.43	0.48	0.54
	std	0.35	0.34	0.26	0.30	0.32	0.30	0.28	0.29	0.32	0.34
		ironia	margherita	mucca	orchestra	orologio	ospedale	patata	persona	saggezza	strategia
abstractness	mean	0.77	0.38	0.43	0.43	0.44	0.63	0.47	0.55	0.72	0.66
	std	0.14	0.22	0.25	0.29	0.27	0.22	0.27	0.27	0.13	0.12
inclusiveness	mean	0.38	0.36	0.45	0.32	0.47	0.71	0.56	0.41	0.49	0.51
	std	0.29	0.36	0.38	0.31	0.35	0.28	0.31	0.30	0.33	0.33

Table 1
Mean and standard deviation of the abstractness and inclusiveness for each token across all different possible contexts.

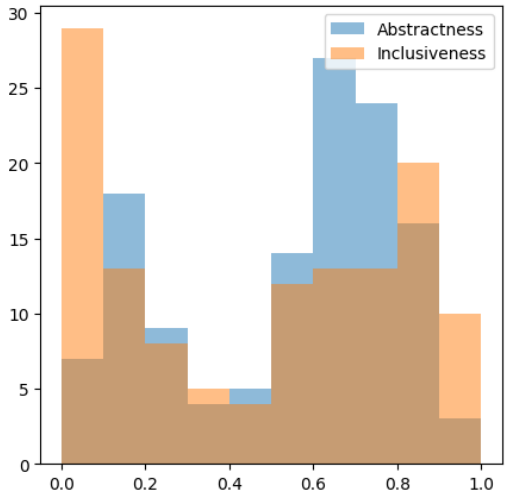

Figure 3: Distribution of the abstractness and inclusiveness scores in the dataset.

for this dataset as without some reference example, the scoring becomes too variable.

With a 3-shot approach and the prompts we used, all models we test appear to be able to understand the task and performance improves with these prompts when compared to less specific ones.

3.4. Detailed data statistics

The dataset contains 127 samples each sample focused on a token, the same token appears more than once in the dataset, on average 6.35 times, in different contexts.

While the dataset contains 127 samples (a limited amount), Figure 3 shows that both abstractness and inclusiveness are well spread across the dataset and there are samples for all values between 0 and 1. Interestingly, while the two concept under study are different, the two scores are similarly distributed across the dataset, but there is a higher number of samples with abstractness value around 0.8 while for inclusiveness the peak is around 0.1, showing a partial anti-pattern between the two scores, and the concept they are meant to distill.

	mistral 7b	llama-3.1-8b	llama-3.1-70b
abstractness	0.22	0.30	0.53
inclusiveness	0.00	0.30	0.41

Table 2
Pearson correlation between the model predicsions and the human annotations for abstractness and inclusiveness scores, measure for three different models, mistral 7b, llama-3.1-8b and llama-3.1-70b.

To investigate the relevance of the context in the assessment of abstraction and inclusiveness, Table 1 shows the mean and standard deviation of the abstractness and inclusiveness of a token when varying context, for all the tokens in the dataset. The standard deviation is often between 0.2 and 0.4 for a score bound between 0 and 1, this shows significant sensitivity to context and highlights how, even if tokens are repeated, each sample is valuable on its own and provides different insights about the token.

4. Metrics

We measure Pearson correlation between the abstractness and inclusiveness scores predicted by the model and the gold human annotation. More specifically, since it is challenging to have the models output a continuous value for the abstractness or inclusiveness of a token in context, we have them generate a discrete score from 1 to 5.

The evaluation is done following a likelihood based approach, after prompting the model to answer our question, we pick the highest likelihood token among 1, 2, 3, 4 and 5 and pick that as the model selection. After doing so for each sample, we compute the Pearson correlation between these values and a discretized version of the continuous scores (discretization does not affect the results)

assigned by humans to the same samples.

Table 2 shows our evaluation of three powerful, Emglish-first language models, mistral 7b [11], llama-3.1-8b and llama-3.1-70b [12], note that we use the instruct version of all three models, and we omit it from the names.

These initial results show that the models are able to capture both abstractness and inclusiveness, with the exception of mistral 7b that fails at understanding inclusiveness (Pearson correlation is 0). At the same time, a powerful LLM like llama-3.1-70b is not able to capture the full complexity of the task, with a Pearson correlation that is as low as 0.53 for abstractness and 0.41 for inclusiveness. This shows that while not alien to the concept of abstractness and inclusiveness, the models are still far from fully understanding it.

Assessing abstractness seems to be easier for LLMs, since every model performs better in this task than in the inclusiveness one. This is interesting although hard to interpret. One possible explanation is that abstractness is a feature that is already made explicit by the choice of the stimuli. Those words do show a variation between different contexts of use, and this is one of the objectives of such challenges with contextual information, but we can also organize these nouns, out of context, discretely along the axis of variation between abstract (e.g. *ambizione – ambition*) and concrete (e.g. *benzina – petrol*). On the contrary, inclusiveness cannot be resolved in any way without considering a proper context; a word form by itself does not convey any information about how much generic, thus inclusive, is the concept behind that lexical label. In light of this, we can hypothesize that when a model has to deal with abstractness/concreteness, it may not be able to rank two occurrences of the same word in slightly different contexts, but for sure it can judge as more concrete or more abstract all the occurrences of one target word with respect to those of another. But when it comes to inclusiveness, thus evaluate if one occurrence is more specific or generic than another, the model is probably struggling more.

Another possible interpretation of these unbalanced results between abstractness and inclusiveness may depend on the quantity of information about the two features: while on abstractness/concreteness there are many studies available online (on English and Italian, as well as on other languages), inclusiveness (and also genericity/specificity, which are the most used terms in literature to refer to this semantic feature) is an understudied topic. We can thus hypothesize that knowledge about abstractness is more formalised in training data, while inclusiveness is not.

Moreover, we confirm that also for this task larger models perform better, Llama 3.1-70b outperforms llama-3.1-8b by a large margin, and that training on more data provides stronger models also in this case, indeed, llama 3.1 outperforms mistral 7b also by a large margin.

Finally, we remark that we avoid testing models that have been tuned for Italian to let participants to the Challenge measure the performance improvements provided by Italian focused training.

5. Conclusions

We propose the ABRICOT benchmark, a dataset composed of 127 humanly annotated samples to measure the abstraction and concreteness of words. Each sample is annotated by 5 - 7 raters who ranked them with a continuous score from 0 to 1 from most concrete to most abstract and a second one measured in the same way from least to most inclusive.

We propose two Tasks, measuring abstractness and inclusiveness and we test three powerful language models on our benchmark, *mistral 7b*, *llama 3 8b* and *llama 3 70b*, we show that when correlating their generations with the humans scores, the highest result on abstractness is 0.53 achieved by the largest llama 3 while on inclusiveness the correlation is bound by 0.41, showing that inclusiveness is harder to understand than abstractness.

We hope that the ABRICOT benchmark will foster the development of new language models in Italian as well as new benchmarks investigating phenomena with a theoretical linguistic foundation such as abstractness and inclusiveness.

6. Limitations

The main limitation of the datasets is the low number of samples it contains, in particular since samples can repeat tokens and there are indeed only 20 unique ones. This can limit the validity of the models assessment, since the topics and vocabulary we cover is rather limited, although we have shown that in terms of both abstractness and inclusiveness, the dataset is well spread and provides a good coverage of both concepts.

Acknowledgments

This work was partially supported by the Project PRIN 2022EPTPJ9 (WEMB – "Word EMBeddings: From Cognitive Linguistics to Language Engineering, and Back"), funded by the Italian Ministry of University and Research (MUR), and the Project ERC-2021-STG-101039777 (ABSTRACTION), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] M. Krifka, F. J. Pelletier, G. Carlson, A. ter Meulen, G. Chierchia, G. Link, Genericity: An introduction, in: G. N. Carlson, F. J. Pelletier (Eds.), The Generic Book, University of Chicago Press, 1995, pp. 1–124.

[2] L. Behrens, Genericity from a cross-linguistic perspective, Linguistics (2005) 275–344.

[3] O. Dahl, The marking of the episodic/generic distinction in tense-aspect systems, in: G. N. Carlson, F. J. Pelletier (Eds.), The Generic Book, University of Chicago Press, 1995.

[4] D. L. Chatzigoga, Genericity, in: The Oxford Handbook of Experimental Semantics and Pragmatics, Oxford University Press, 2019, pp. 156–177.

[5] 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: Overview, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024.

[6] L. Gregori, M. Montefinese, D. P. Radicioni, A. A. Ravelli, R. Varvara, CONcreTEXT@EVALITA2020: The Concreteness in Context Task., in: EVALITA, 2020.

[7] A. Friedrich, A. Palmer, M. P. Sørensen, M. Pinkal, Annotating genericity: a survey, a scheme, and a corpus, in: Proceedings of the 9th Linguistic Annotation Workshop, 2015, pp. 21–30.

[8] P. Chocron, P. Pareti, Vocabulary alignment for collaborative agents: a study with real-world multilingual how-to instructions, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 159–165. URL: https://doi.org/10.24963/ijcai.2018/22. doi:10.24963/ijcai.2018/22.

[9] P. Lison, J. Tiedemann, M. Kouylekov, OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1275.

[10] A. A. Ravelli, G. Puccetti, M. Bolognesi, Abricot: Abstractness and inclusiveness in context, 2024. URL: osf.io/ja89x. doi:10.17605/OSF.IO/JA89X.

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

[12] 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. Al-

varado, 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.