VeryfIT - Benchmark of Fact-Checked Claims for Italian: A CALAMITA Challenge

Jacopo Gili[1], Viviana Patti[1,†], Lucia Passaro[2,†] and Tommaso Caselli[3,†]

[1]Department of Computer Science, University of Turin, Italy
[2]Department of Computer Science, University of Pisa, Italy
[3]CLCG, University of Groningen, The Netherlands

Abstract

Achieving factual accuracy is a known pending issue for language models. Their design centered around the interactive component of user interaction and the extensive use of "spontaneous" training data, has made them highly adept at conversational tasks but not fully reliable in terms of factual correctness. VeryfIT addresses this issue by evaluating the in-memory factual knowledge of language models on data written by professional fact-checkers, posing it as a true or false question. Topics of the statements vary but most are in specific domains related to the Italian government, policies, and social issues. The task presents several challenges: extracting statements from segments of speeches, determining appropriate contextual relevance both temporally and factually, and ultimately verifying the accuracy of the statements.

Keywords

fact checking, benchmark, factual knowledge, Italian, fake news, CALAMITA, CheckIT!

1. Challenge: Introduction and Motivation

The pollution of the information ecosystem by means of misleading or false information has reached unprecedented levels at a global scale. This has been possible thanks to a combination of multiple factors, among which the collapse of (local and national) journalism; an increasing sense of distrust in science and evidence-based facts; and the presence of computational amplification tools such as bots [1, 2]. In this sense the rise of Large Language Models (LLMs) with the constant increase of their performances has introduced both opportunities and challenges in the fight against misinformation: while LLMs possess the capability to generate coherent and contextually relevant text, they also pose risks by potentially producing deceptive misinformation at scale [3, 4].

Testing factual and common sense knowledge in LLMs has been a common although not easy task involving mostly multi-choice question answering, a method easy to automate and not prone to ambiguity, and spanning across wide ranges of academic and professional domains like mathematics, medicine, history, law, general knowledge and many others [5, 6, 7, 8, 9, 10, 11, 12].

Developing benchmarks to test the ability of LLMs to accurately evaluate factual knowledge is more relevant than ever considering the ease of access of these tools to non-experts for any purpose (entertainment, education, professional settings) and the increasing integration of these technologies in every day activities.

Notably, most of these tasks and corresponding benchmarks are in English with other languages being represented through machine-translated data or no data at all. This is true for Italian too. For instance, SQUAD-IT [13] is a machine-translated version of the SQUAD dataset [14] and it is the reference for evaluating models on QA-tasks.

While machine-translation has been constantly improving, it can indeed easily introduce artefacts in the output text impairing naturalness and correctness, moreover translated data can be subjected to the loss of nuance and context as translations may not capture cultural nuances or contextual meanings, leading to misunderstandings or misinterpretations in the target language: certain phrases or idioms may not have direct equivalents in other languages, and the presence of linguistic constructions typical of the source language may be encouraged excessively [15].

By using data from a professional fact-checking agency[1] we can test knowledge memorization of LMs and to what extend intra-memory conflicts, resulting in "hallucinations", arise. Furthermore, doing so using Italian data centered around the Italian and European contexts ensures testing LM's functionalities directly in Italian.

This task is based on CheckIT! [16], a resource of expert fact-checked claims designed to fill a gap for the development of AI- assisted fact-checking pipelines for

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

† These authors contributed equally.

jacopo.gili584@edu.unito.it (J. Gili); viviana.patti@unito.it (V. Patti); lucia.passaro@unipi.it (L. Passaro); t.caselli@rug.nl (T. Caselli)
0009-0007-1343-3760 (J. Gili); 0000-0001-5991-370X (V. Patti); 0000-0003-4934-5344 (L. Passaro); 0000-0003-2936-0256 (T. Caselli)

[1]Data have been obtained from Pagella Politica

Italian.

2. Challenge: Description

The challenge is a binary classification task in a zero-shot setting: for each atomic statement, any LM is asked to determine its factuality *with respect to the time it was uttered* by answering only with one of the two labels, "Vero" (true) or "Falso" (false). A third label for half true statements could have been easily kept as it was already part of the dataset from which the data is sourced, but in this first stage we opted for the binary setting as to limit task complexity.

Some cases in the dataset exhibit complexities due to the combination of multiple pieces of information within a single claim, which can affect the final determination of veracity. For instance, consider the following scenario:

Original claim	Translation
«Se è vero che oltre l'82% dei morti da Covid hanno più di 70 anni, non si capisce perché meno della metà degli over 80 sia stato vaccinato finora»	«If it is true that over 82% of Covid deaths are over 70 years old, it is not clear why less than half of those over 80 have been vaccinated so far»

Table 1
Example of a claim

The informations concerning this statement are:

1. Out of all the deceased due to the Covid19 pandemic, 82% are people over 70 years old.
2. Less than half of the citizens over 80 years old had administered at least one dose of vaccine against Covid19.

This example also highlights the importance of incorporating the appropriate temporal context in the verification process. Factual information, especially involving statistics or reports about the state of the world, evolves over time and failing to account for this can invalidate the conclusions drawn by experts. Although more complex statements require a broader knowledge base, by now language models have shown understanding abilities well over this level and should not be subjugated by it.

3. Data description

The **VeryfIT** dataset consists of **2,021** claims taken from CheckIT! [16]. Not all claims were included due to the binary format of the task as VeryfIT classifies claims as either "Vero" [True] or "Falso" [False], whereas CheckIT! recognizes an intermediate "Ni" [Half true] label. As a result, all claims with the "half-true" verdict were discarded.

Furthermore, we considered pertaining to the task to provide also a smaller subset of claims, "**VeryfIT_small**", balanced on the political orientation of the politician speaking, as misinformation can occur on all topics but when referring to political misinformation each side of the political spectrum has some more widespread topics and recurrent formulations.

Additionally, an annotation task was carried out on the VeryfIT_small subset aimed at the clarification of statements presenting a level of ambiguity that would have proven detrimental to the task: around 12% of the statements have available an alternative version "enriched" of informations vital to the task. We will refer to them as "enriched statements" (subsection 3.2).

In conclusion, 2 versions of the dataset are available: VeryfIT (2,021 claims) and VeryfIT_small (352 claims of which 43 with an enriched version).

3.1. Creation of VeryfIT_small

The first step to achieve this goal was to exclude around 400 out of the 2,021 claims of VeryfIT for which information about the political orientation of the speaker was not available.

We then mapped, using Wikipedia as a source, the political orientation of the parties (and thus of the authors of the claims at the moment of remark) into eight fine-grained, commonly recognized political categories: far-left, left, center-left, center, center-right, right, far-right. An illustration on the list of all the parties and their corresponding political orientation is reported in Table 2. An additional label 'transverse' was added to indicate a non precise placement in the political spectrum. This label includes one party ("Movimento 5 Stelle"), members of the Italian institutions above political parties (e.g. the President of the Republic), and experts not affiliated to any political party or political coalition like members of a *technical government* [2].

At first glance, the Italian political spectrum may appear only slightly unbalanced. Despite the absence of a far-left representation, the distribution of parties across the spectrum is relatively symmetrical. Out of the 23 political parties in the data, six are from the left, two from the center-left, six from the center, three from the center-right, two from the right, and three from the far-right. However, the distribution of claims is not as well balanced, with a larger number of claims from the rights and far-right parties than the rest as reported in table 3.

To ensure the balance of our benchmark we decided to reduce the label granularity from eight to four, by col-

[2] https://en.wikipedia.org/wiki/Technocratic_government_(Italy)

Political party	Orientation label
Alleanza Verdi e Sinistra	left
Alternativa Popolare	center-right
Articolo Uno	center-left
Azione	center
Coraggio Italia	center-right
Europa Verde	left
Forza Italia	right
Fratelli d'Italia	far-right
Impegno Civico	center
Indipendente	transverse
Italexit	far-right
Italia Viva	center
Lega Nord	far-right
Liberi e uguali	left
Movimento 5 Stelle	transverse
Nuovo Centro Destra	center-right
Partito Democratico	center-left
Più Europa	center
Popolo della Libertà	right
Possibile	left
Radicali Italiani	center
Scelta Civica	center
Sinistra Ecologia Libertà	left
Sinistra italiana	left
Tecnico	transverse

Table 2
VeryfIT data: Italian political parties and their orientation.

	Claims		
Political side	**True**	**False**	**Total**
Left	44	28	72
Center-left	323	110	433
Center	105	82	187
Center-right	8	2	10
Right	79	84	163
Far-right	156	241	397
Transverse	209	146	355
total	924	693	1,617

Table 3
VeryfIT data after exclusion of claims where information about political orientation of the speaker was not available: Distribution of verdict labels in the political spectrum.

lapsing labels far-left, left and center-left into '*left*' [SX], and far-right, right and center-right into '*right*' [DX]. Labels *center* [C] and *trasversal* [T] remained untouched. The re-aggregated coarse-grained labels are reported in Table 4.

Although the distribution is still unbalanced between the two end point (SX and DX), this setting, with the lowest cardinality being 187 (for C) easily allows us generate a perfectly balanced dataset along the political orientations. For the first version of **VeryfIT_small**, each block contributes with 88 claims resulting in a total of **352 entries**, with future works planned to expand it.

	Claims		
Political side	**True**	**False**	**Total**
Left [SX]	367	138	505
Center [C]	105	82	187
Right [DX]	243	327	570
Transverse [T]	209	146	355

Table 4
VeryfIT data after exclusion of claims where information about political orientation of the speaker was not available: Distribution of verdict labels in the political spectrum after label collapse.

	Claims		
Political side	**True**	**False**	**Total**
Left [SX]	64 [13]	24 [2]	88 [15]
Center [C]	46 [4]	42 [7]	88 [11]
Right [DX]	40 [4]	48 [5]	88 [9]
Transverse [T]	50 [2]	38 [6]	88 [8]
total	200 [23]	152 [20]	352 [43]

Table 5
VeryfIT_small: Final distribution of verdict labels in the political spectrum. Highlighted in green the number of labels of enriched statements (explained in subsection 3.2).

3.2. Enriched statements

Given the specificity of the statements, many of which require detailed knowledge of topics related to Italian institutions and policies, and the occasional ambiguity arising from their oral nature, the task has been further divided into two sub-tasks with slight data modifications, aimed at adding vital context to statements that were excessively reliant on information external to the statements themselves. The altered statements account for **around 12%** of the **VeryfIT_small** dataset, as excessive human intervention would undermine the core principle of testing on natural data, aligned with what language models might be asked to handle in real-life scenarios. In most cases, minimal adjustments were made, such as retaining the original claim but adding the name of the politician speaking or clarifying specific references.

The goal of partially or entirely removing the initial layer of complexity, by simplifying the extraction of the relevant information from the statement for verification, is to highlight a stronger correlation between the benchmark results and the language model's actual factual knowledge: when working with natural data, the model's responses may stem from its difficulty in comprehending the specific information it is being asked to verify. However, with altered data, its responses are more directly influenced by gaps in its knowledge.

Examples of enriched statements are reported in Table 6:

Original statement	Enriched statement
Abbiamo 490 grandi elettori	Gli elettori dell'area di centrosinistra che voteranno per l'elezione del Presidente della Repubblica saranno 490.
Oggi in Italia sono 796 quelli che pagano più di 1 milione di euro	Oggi in Italia sono 796 quelli che dichiarano un reddito superiore ad 1 milione di euro.
[Alle europee] io ho battuto Salvini in molti capoluoghi di provincia	[Alle europee] io [Carlo Calenda] ho battuto Salvini in molti capoluoghi di provincia.
In parlamento stiamo facendo un lavoro che risponde a una prerogativa costituzionale. Certamente si sarebbero tutti auspicati, me compresa, tempi più brevi ma non stiamo perdendo tempo. Stiamo svolgendo un ruolo che ci compete e che la Costituzione da' al parlamento.	L'elezione dei membri della Corte Costituzionale e del Consiglio Superiore della Magistratura (Csm) è un dovere che la costituzione italiana dà al parlamento.

Table 6
Comparison of Original and Enriched Statements

The reasons for enriching the statements in table 6 all revolve around the lack of pivotal information to determine factuality: The first statement is completely missing the context and presents an unclear term "grandi elettori" [big voters], relatively known in the political context, but that could be mistaken for a physical feature or for a consideration regarding the age of voters; the second statement has an unclear formulation as "pagare" [to pay] does not refer univocally to taxes; the third statement is missing the subject; the fourth and last statement is missing part of its context as "stiamo facendo *un lavoro*" [we are doing a job] "stiamo svolgendo *un ruolo*" [we are playing a role] both refer to a very specific duty of the parliament that does not get mentioned directly.

Preliminary results obtained through the chat function of Claude 3.5 Sonnet[3] and GPT-4o [4] show that respectively two out of the four statements (Claude) and one out of the four statements (GPT) reported in Table 6 get wrongly classified when presented in the original version, while providing the models with the enriched versions brings up the correctly classifications to four out of four for both models. These results however can only partially prove the effectiveness of enriched statements as different models when presented a partial context could provide different verdicts, even guessing the right one.

3.3. Annotation details

During the making of the VeryfIT datasets, it was noticed that not all the statements were actual claims: in articles with multiple claims to check, the 'statement' field was filled with a short title resuming them all, often in the format "[name of the politician] on [topic]". Regular expressions were used to highlight statements not starting with '"' or '«', the two symbols used to denote a dialogue or part of a speech, and a manual check brought to the exclusion of around 170 statements. Moreover around 30 statements with formats resembling "[name of the politician] is [right/wrong] on [topic]: [statement]" were reformulated as claims by removing hints about the factuality verdict and the author of the statement. A couple examples are brought up in table 7.

Original statement	Reworded statement
Giulia Grillo sbaglia: i medici e gli infermieri italiani non sono i meno pagati	i medici e gli infermieri italiani sono i meno pagati
Secondo Di Maio il governo investe nelle centrali a carbone, ma è il contrario	Il governo investe nelle centrali a carbone
No, per la Corte dei Conti non ci saranno 17 miliardi di nuove tasse	Per la Corte dei Conti ci saranno 17 miliardi di nuove tasse

Table 7
Examples of reworded statements

Another important annotation step has been producing the enriched statements. A human annotator[5] reviewed the VeryfIT_small dataset, identifying statements that could benefit from additional context, and produced enriched variations of those statements. In most cases, minimal adjustments were made, such as retaining the original claim but adding the name of the politician speaking or clarifying anaphoric references.

[3] https://claude.ai/chat
[4] https://chatgpt.com/
[5] All the annotations noted in the report was done by the first author of the paper, master student in Computer Science with a background in Natural Language Processing

The decision of applying this annotation step to the VeryfIT_small subset, instead of the full dataset, is related to the amount of manual work it would have required.

Additionally another annotation step involved completing the "macro_area" [topic] field for all the 352 entries of VeryfIT_small. Although this field was included in the original dataset, it was missing a value in approximately 15% of the entries. This was done manually, classifying statements into the pre-existing topic labels which are: 'questioni sociali' [social matters], 'economia' [economy], 'esteri' [foreign affairs], 'giustizia' [justice], 'istituzioni' [institutions], 'ambiente' [environment], 'altro' [others]. The new labels were chosen by comparing unlabelled statements with statements that already had a label and inspecting the contents of the articles from which they were extracted, sometimes only needing to look at the 'tags' field to find all the information needed. To avoid even the smallest imprecision that would have impaired the original label system made by journalist, non-certain labels were put in the 'altro' category.

Statistics about the distribution of these labels can be found in section 3.6.

3.4. Data format

Brief explanation of the data fields:

- **annotato**: If True, the statement has a revised version.
- **id**: ID of the corresponding article in CheckIT!.
- **statement_date**: Date of statements diffusion.
- **statement**: The statement.
- **verdict**: Factuality verdict.
- **orientamento**: Orientation of the political party of the politician author of the statement.
- **macro_area**: Topic of the statement.
- **tags**: List of tags.
- **statement_revised**: Revised version of the statement, if present.

Fields such as 'macro_area' and 'tags' serve as indicators of the topic, the former providing a general categorization and the latter offering more specific details. These informations were included with in mind future tasks that could reveal differences in factual knowledge across different subjects.


```
{
  "annotato": False,
  "id": 991,
  "statement_date": 2019-07-12,
  "statement": "[Il salario minimo n.d.r.]
       Manca solo a noi e ai Paesi dell'Est
       Europa",
  "verdict": "Falso",
  "orientamento": 'C',
  "macro_area": "questioni sociali",
  "tags": "['questioni sociali', 'panzana
       pazzesca', 'italia', 'eu', 'salario
       minimo']",
  "statement_revised": ""
},
{
  "annotato": True,
  "id": 123,
  "statement_date": 2023-02-14,
  "statement": "Il canone in bolletta fu una
       mia scelta. Costava 113 euro. Averlo
       fatto pagare a tutti ha portato a un
       abbassamento del costo da 113 a 90 euro
       ",
  "verdict": "Vero",
  "orientamento": 'C',
  "macro_area": "altro",
  "tags": "["canone", "rai", "bolletta", "
       costo"]",
  "statement_revised": "Il canone in bolletta
       fu una mia scelta [di Matteo Renzi].
       Costava 113 euro. Averlo fatto pagare a
        tutti ha portato a un abbassamento del
        costo da 113 a 90 euro"
}
```


Figure 1: Data format

3.5. Example of prompts used for zero shots

The models are expected to be evaluated on this task in a zero-shot setting, thereby also better resembling the conditions of a real use-case scenario. The prompt we suggest to use for the evaluation is basic, and urges the model to limit its answer to just the letter corresponding to the answer. The original prompt in Italian, together with its English translation, are reported in Box 1.

Prompt

Il seguente statement, nella data indicata, è vero o falso? Rispondi solo con "Vero" o "Falso".

The following statement, on the date indicated, is true or false? Answer only with "True" or "False".

Box 1: Zero-shot prompt

The prompt does not contain any information about the subject of the question or any other informative cues apart from the time reference needed to anchor the claim in a temporal context. In this way, our benchmark not only tests the model in question answering, but also indirectly tests the instruction-following abilities of the model in a language different than English.

3.6. Detailed data statistics

The full VeryfIT! dataset is composed of 2,021 entries in the italian language. Out of these claims, 352 form the VeryfIT_small dataset in which the entries are equally split across the three main sides of a semplification of the classical political spectrum (left, right, center) and a fourth label 'trasversal', used to address non precise placement in the political spectrum or complete absence of affiliation to any political party or political coalition.

Of the 352 claims in the VeryfIT_small dataset, 43 have available an enriched variation of the statement, providing additional context alongside the original statement.

The distribution of claims and factuality labels across topics is presented in Table 8, Table 9, Table 10, Table 11.

	Claims		
Macro_area	**True**	**False**	**Total**
questioni sociali	256	170	426
economia	264	155	419
istituzioni	243	77	320
esteri	105	53	158
giustizia	60	26	86
altro	46	32	78
ambiente	42	18	60
un-noted	180	294	474
total	1,196	825	2,021

Table 8
VeryfIT: Distribution of claims and factuality labels per topics ordered by total value.

Further statistics on the original CheckIT! dataset is available in Figure A and Table A in Appendix A.

4. Metrics

Accuracy serves as the evaluation metric of the task due to its intuitive interpretation and broad applicability. Accuracy provides a clear measure of a classifier's overall performance by calculating the proportion of correct predictions among total cases examined.

No other metrics were chosen for the task.

	Orientation label						
Macro_area	**SX**	**CSX**	**C**	**CDX**	**DX**	**E-DX**	**T**
questioni sociali	19	105	27	5	27	101	80
economia	11	119	43	1	52	54	52
istituzioni	10	81	10	3	38	33	71
esteri	4	32	17	0	11	41	33
giustizia	3	11	1	0	13	11	24
altro	1	17	8	1	6	8	14
ambiente	1	10	2	0	2	6	14
un-noted	23	59	79	0	14	145	68

Table 9
VeryfIT data after exclusion of claims where information about political orientation of the speaker was not available: Distribution of claims per topic and positioning in the political spectrum.

	Claims		
Macro_area	**True**	**False**	**Total**
questioni sociali	50 [2]	37 [4]	87 [6]
economia	53 [4]	37 [4]	90 [8]
istituzioni	46 [11]	17 [5]	63 [16]
esteri	26 [4]	19 [3]	45 [7]
ambiente	8 [1]	10	18 [1]
giustizia	7	8	15
altro	10 [1]	24 [4]	34 [5]
total	200 [23]	152 [20]	352 [43]

Table 10
VeryfIT_small: Distribution of claims and factuality labels per topics ordered by total value. Highlighted in green the number of labels of enriched statements.

	Orientation label			
Macro_area	**SX**	**C**	**DX**	**T**
questioni sociali	22 [2]	15 [1]	25 [2]	25 [1]
economia	28 [2]	30 [5]	19 [1]	13
istituzioni	19 [8]	8 [1]	15 [3]	21 [4]
esteri	9 [2]	13 [1]	12 [2]	11 [2]
ambiente	2	7	3 [1]	6
giustizia	2	4	3	6
altro	6 [1]	11 [3]	11	6 [1]
total	88 [15]	88 [11]	88 [9]	88 [8]

Table 11
VeryfIT_small: Distribution of claims per topic and positioning in the simplified political spectrum. Highlighted in green the number of labels of enriched statements.

5. Limitations

The totality of the data comes from an expert, reliable source. For this reason, the quality of the verdicts is assured to be high. One possible limitation is due to the time-relatedness of said verdicts: claims can be truth and false at times depending on the temporal context

in which they are evaluated. LMs could have an hard time discerning informations pertaining specific time intervals, given that they could also not have been trained on data related to them.

Another limitation could be the depth of the factual knowledge required to understand and consequently answer the questions of the dataset. As previously stated, VeryfIT data is about italian/european context and touches details of various fields that most probably not even the citizens would know about!

Remarkably, the risk of the data being present in training corpuses for LMs should be mitigated as the CheckIT! dataset is not publicly released.

Finally, fact-checking is a very complex task and statements could carry different degrees of truthness, more than a binary setting can express. We chose to limit for now the task to a binary classification challenge to not make it too complicated, but we do not exclude further development towards a multi-label setting to better capture the nuances of the fact-checking process.

6. Ethical issues

No ethical issue has arisen from the making of this task, all the data has been sourced through agreements with the original authors.

7. Data license and copyright issues

The data cannot be publicly released due to a Data Sharing Agreement between University of Groningen and Pagella Politica. At the moment of writing of this contribution to obtain VeryfIT! contact dr. Tommaso Caselli.

References

[1] T. Economist, Disinformation is on the rise. how does it work?, 2024. URL: https://www.economist.com/science-and-technology/2024/05/01/disinformation-is-on-the-rise-how-does-it-work.

[2] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[3] OpenAI, Disrupting deceptive uses of ai by covert influence operations, 2024.

[4] C. Chen, K. Shu, Combating misinformation in the age of llms: Opportunities and challenges, AI Magazine (2024). URL: https://doi.org/10.1002/aaai.12188. doi:10.1002/aaai.12188.

[5] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2021. URL: https://arxiv.org/abs/2009.03300. arXiv:2009.03300.

[6] A. Srivastava, D. Kleyjo, Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, Transactions on Machine Learning Research (2023).

[7] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, et al., Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset, Advances in Neural Information Processing Systems 36 (2024).

[8] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring how models mimic human falsehoods, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3214–3252. URL: https://aclanthology.org/2022.acl-long.229. doi:10.18653/v1/2022.acl-long.229.

[9] P. Wang, A. Chan, F. Ilievski, M. Chen, X. Ren, Pinto: Faithful language reasoning using prompt-generated rationales, in: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022.

[10] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL: https://arxiv.org/abs/2009.13081. arXiv:2009.13081.

[11] A. Talmor, J. Herzig, N. Lourie, J. Berant, Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL: https://arxiv.org/abs/1811.00937. arXiv:1811.00937.

[12] L. C. Passaro, A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event, Information Sciences 615 (2022) 657–677. URL: https://www.sciencedirect.com/science/article/pii/S0020025522008167. doi:https://doi.org/10.1016/j.ins.2022.07.128.

[13] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: C. Ghidini, B. Magnini, A. Passerini, P. Traverso (Eds.), AI*IA 2018 – Advances in Artificial Intelligence, Springer International Publishing, Cham, 2018, pp. 389–402.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, 2016. URL: https://arxiv.org/abs/1606.05250. arXiv:1606.05250.

[15] I. Plaza, N. Melero, C. del Pozo, J. Conde, P. Reviriego, M. Mayor-Rocher, M. Grandury, Spanish and llm benchmarks: is mmlu lost in translation?,

arXiv preprint arXiv:2406.17789 (2024).

[16] J. Gili, L. Passaro, T. Caselli, Checkit!: A corpus of expert fact-checked claims for italian, in: F. Boschetti, G. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics, CEUR Workshop Proceedings, CEUR Workshop Proceedings (CEUR-WS.org), 2023. Publisher Copyright: © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).; 9th Italian Conference on Computational Linguistics, CLiC-it 2023 ; Conference date: 30-11-2023 Through 02-12-2023.

Appendix A

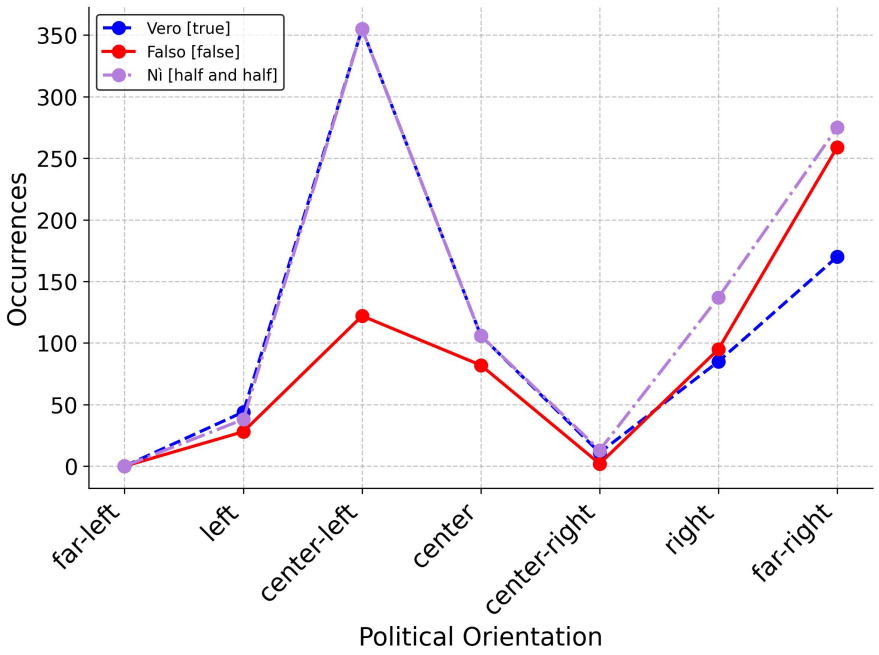

Figure A: Original data from subset d1 of CheckIT!: Claims distribution in the political spectrum in reference with factual veracity.

	Orientamento						
Macro_area	**SX**	**CSX**	**C**	**CDX**	**DX**	**E-DX**	**T**
economia	21	243	74	6	119	145	142
questioni sociali	30	215	62	12	50	203	174
istituzioni	11	150	24	6	81	54	144
esteri	7	75	25	0	19	99	80
ambiente	5	30	8	0	2	9	29
giustizia	3	23	4	0	23	23	35
altro	33	96	97	2	23	171	107
total	110	832	294	26	317	704	711

Table A
Original data from subset d1 of CheckIT!: Distribution of claims per topic and positioning in the full political spectrum. Far-left label is omitted as non-present in the dataset.