Claim Check-Worthiness Detection:
How Well do LLMs Grasp Annotation Guidelines?

Anonymous ACL submission

Abstract

The increasing threat of disinformation calls for automating parts of the fact-checking pipeline. Identifying text segments requiring fact-checking is known as *claim detection* (CD) and *claim check-worthiness detection* (CW), the latter incorporating complex domain-specific criteria of worthiness and often framed as a ranking task. Zero- and few-shot LLM prompting is an attractive option for both tasks, as it bypasses the need for labeled datasets and allows verbalized claim and worthiness criteria to be directly used for prompting. We evaluate the LLMs' predictive and calibration accuracy on five CD/CW datasets from diverse domains, each utilizing a different worthiness criterion. We investigate two key aspects: (1) how best to distill factuality and worthiness criteria into a prompt and (2) what amount of context to provide for each claim. To this end, we experiment with varying the level of prompt verbosity and the amount of contextual information provided to the model. Our results show that optimal prompt verbosity is domain-dependent, adding context does not improve performance, and confidence scores can be directly used to produce reliable check-worthiness rankings.

1 Introduction

Automating fact-checking is becoming crucial in response to rising amounts of data and disinformation. Fact-checking is typically done on claims – to warrant fact-checking, a claim must be both *factual* (i.e., related to purported facts) and *check-worthy* (i.e., of interest to society). The NLP tasks of identifying factual and check-worthy claims are known as *claim detection* (CD) and *claim check-worthiness detection* (CW), respectively. While both tasks are typically defined as classification tasks, CW can also be framed as a ranking task, mimicking the prioritization process employed by fact-checking organizations (FullFact, 2020).

Both CD and CW are challenging for several reasons. Firstly, the underlying concepts of factual claims and worthiness resist straightforward definitions. To grasp factuality, Konstantinovskiy et al. (2021) presented a thorough categorization of factual claims, while Ni et al. (2024) provided a definition distinguishing opinions. Defining check-worthiness is made more challenging by its subjective, context-dependent nature and temporal variability. Assessing it usually requires choosing more specific criteria, such as relevance to the general public (Hassan et al., 2017a) or policymakers, potential harm (Nakov et al., 2022), or alignment with a particular topic (Stammbach et al., 2023; Gangi Reddy et al., 2022)). Another challenge is identifying the situational context (including previous discourse and speaker information) required to determine claim factuality and check-worthiness.

CD and CW have been addressed using traditional supervised ML and fine-tuning pre-trained language models. However, acquiring labeled datasets can be challenging – they must align with a specific language, domain, and genre and match the desired factuality and worthiness criteria. Moreover, dataset annotation is costly and requires redoing if criteria change. LLMs present a viable alternative to supervised methods owing to their strong zero- and few-shot performance (Kojima et al., 2022; Brown et al., 2020). Over time, fact-checking organizations have refined principles for claim prioritization, and few-shot prompting offers a seamless way to transfer this knowledge to the model. Thus, an effective strategy might entail zero- and few-shot prompting with check-worthiness criteria from annotation guidelines.

In this paper, we study the predictive and calibration accuracy of zero- and few-shot LLM prompting for CD and CW. We experiment with five datasets, each with a different factuality or worthiness criterion outlined in the accompanying annotation guidelines. We investigate two key aspects: (1) how to best distill factuality and worthiness criteria from the annotation guidelines into the prompt and (2) what amount of context to provide for each claim. For (1), we experiment with varying the level of prompt verbosity, starting from brief zero-shot prompts to more detailed few-shot prompts that include examples. For (2), we experiment with expanding the prompt with co-text and other components of the claim's situational context. Finally, we consider CW as a ranking task using LLM's confidence scores as a proxy for priority. We show that *gpt-4-turbo* with worthiness criteria adopted from annotation guidelines can yield accuracy and ranking scores comparable to or outperforming existing CD/CW methods. Our results demonstrate the potential of using LLMs for claim check-worthiness detection with minimal prompt engineering.

2 Related Work

The CD and CW tasks constitute the first part of the fact-checking pipeline. Typically framed as classification tasks, they are handled using traditional supervised machine learning (Hassan et al., 2017b; Wright and Augenstein, 2020; Hassan et al., 2017a; Gencheva et al., 2017), or fine-tuning PLMs (Stammbach et al., 2023; Sheikhi et al., 2023).

Recently, the use of LLMs for CD and CW is starting to take on. Sawinski et al. (2023) and Hyben et al. (2023) compare the performance of fine-tuned PLMs with LLMs using zero- and few-shot learning as well as fine-tuning. Although zero- and few-shot approaches for LLMs underperform, the authors note their reliance on internal definitions of worthiness and limited prompt testing. As part of the fully automated fact-checking system relying only on LLMs, Li et al. (2023) implement a CD module using a verbose few-shot prompt, yet they do not report performance metrics. Finally, Ni et al. (2024) tackle CD by proposing a three-step prompting approach to examine model consistency. However, neither Li et al. (2023) nor Ni et al. (2024) address the CW task. To our knowledge, there is no work on CW focused on describing specific worthiness criteria using verbose prompts.

3 Datasets

Our experiments use five datasets in English covering diverse topics and genres:

ClaimBuster (CB) (Hassan et al., 2017a) is a widely used dataset of claims from USA presidential debates. It uses ternary labels (*non-factual*, *unimportant factual*, *check-worthy factual*) and thus allows for the distinction between check-worthy and unimportant factual claims, therefore covering both the CD and CW tasks;

CLEF CheckThat!Lab 2022 (CLEF) (Alam et al., 2021) is a dataset of tweets relating to COVID-19. It comprises two sets: a set of tweets containing claims and a subset of those containing check-worthy claims, thus covering both the CD and CW tasks. Check-worthiness is defined as the need for professional fact-checking, excluding claims that are "too trivial to check";

EnvironmentalClaims (ENV) (Stammbach et al., 2023) focuses on environmental articles and reports. It defines specific criteria for an environmental claim that extend beyond the topic itself (e.g., the claim should be explicit, focus on environmental impact);

NewsClaims (NEWS) (Gangi Reddy et al., 2022) is a dataset of sentences from news articles on COVID-19, with metadata available for positives (speaker, object, claim span). The annotators were asked to judge whether a claim falls into one of the four topic-specific categories – essentially forming the worthiness criteria;

PoliClaim (POLI) (Ni et al., 2024) covers the same topic as ClaimBuster (politics, speeches of governors) but labels only factual claims, leaving out check-worthiness. The binary labels are obtained by aggregating responses to two questions.

We selected these datasets because they disclose annotation guidelines to some degree of detail and cover different topics, genres, and worthiness criteria. Table 1 summarizes the datasets' characteristics (cf. Appendix A for more details). The CB and CLEF datasets cover both tasks, where CB uses ternary labels annotated at once, while CLEF uses binary labels and separate annotation questions for CD and CW. The datasets were either originally annotated using a binary scheme (ENV), Likert scale (CLEF-CW), multi-class based on topic (NEWS), or a follow-up prompt for uncertain instances (POLI). All the datasets provided aggregated binary labels, except CB, where aggregation from ternary into binary CD and CW labels is straightforward. The reported inter-annotator agreement for POLI and CLEF is substantial (Landis and Koch, 1977), while the agreement for ENV and NEWS datasets is moderate, confirming the complexity of the domain-dependant CW task.

	CB	CLEF	ENV	NEWS	POLI
Task	CD+CWD	CD+CWD	CWD	CWD	CD
Labels	ternary	binary*	binary	binary	binary*
# instances	23,533	3,040	2,647	7,848	52 speeches
# instances used	1,032	251	570	6,129	816
Genre	debates	tweets	news articles	reports	speech transcripts
Topic	politics	healthcare	environment	healthcare	political
Co-text	4 previous, on request	-	not available	inconclusive	1 previous, 1 following
Agreement	-*	0.75/0.7	0.47	0.405	0.69
Agreement metric	-	Fleiss-κ	Krippendorff-α	Krippendorff-κ	Cohen-κ

Table 1: Characteristics of the CD and CW datasets used in our experiments. *CB reported no agreement evaluation, but the test set used is agreed upon by experts.

4 Experiments and Results

In our experiments, we use OpenAI models *gpt-turbo-3.5* and *gpt-4-turbo* (cf. Appendix C for a description of parameters). We also experimented with smaller, open-source models, but their responses often did not match the target labels.

4.1 Prompt verbosity

We first investigate how prompt verbosity affects LLMs' predictive accuracy. We hypothesize that the optimal verbosity level depends on the dataset, reflecting the factuality and worthiness criteria differences between the domains. While a brief prompt might lack essential details, a comprehensive prompt featuring extensive definitions and examples may make the task more ambiguous for the model. Based on the content and style of annotation guidelines, we define the following four levels of verbosity (cf. Appendix E for full prompts):

Level V0 serves as the baseline. We use a naive zero-shot prompt relying on the models' internal definitions of *factual* and *check-worthy* for the CD and CW tasks, respectively. As this prompt does not include the specific worthiness criteria from the guidelines, it serves as a domain-agnostic baseline;

Level V1 uses prompts that include the task definition and the set of possible labels but omit detailed explanations of the labels or principles;

Level V2 expands on V1 by adding a more detailed explanation of the labels or general annotation principles (or both, in the case of PoliClaim);

Level V3 builds on V2 with examples provided to annotators. This level is closest to annotation guidelines, encompassing all or most information that the datasets' authors provided in the papers accompanying the datasets.

	CB		CLEF		ENV	NEWS	POLI
	CD	CW	CD	CW	CD	CW	CW
V0	.833	.805	.797	.467	.416	**.583**	**.844**
V1	.883	.885	.799	.552	**.773**	.572	.679
V2	.908	.889	**.806**	**.583**	.69	.48	.541
V3	**.919**	**.927**	.781	.556	.596	.523	.563

Table 2: F1 scores of *gpt-4-turbo* for CD and CW tasks across datasets and prompt verbosity levels (V1–V3). Level V0 corresponds to the naive-prompting baseline.

Table 2 shows the F1 scores by verbosity level for *gpt-4-turbo* (*gpt-3.5-turbo* generally performed worse; cf. Appendix A.2). Predictive accuracy for CD is generally higher than for CW, proving that capturing worthiness is more difficult than identifying factual claims. The optimal verbosity level is not consistent across datasets: the performance increases with verbosity levels for CB, but the trend is reversed for ENV. We observe no consistent trend for CLEF, POLI, and NEWS datasets. The naive baseline prompt outperforms our prompts on POLI and NEWS datasets. On the other hand, our prompts outperform some of the previously reported results (cf. Appendix A.2), proving the potential of prompting with annotation guidelines.

4.2 Incorporating context

Claims are never made in isolation; their context matters not only for verifying their veracity but also for gauging their factuality and worthiness even before fact-checking. We investigate how LLMs' predictive accuracy depends on the amount of situational context provided to the model. To this end, we leverage the context information available in the CB and POLI datasets and expand the prompts by adding the co-text of the claim (Level C1), speaker information (Level C2), or both (Level C3) (NEWS includes additional information, cf. Appendix A.3).

	CB		POLI
	CD	CW	CD
C0	**.919**	**.927**	**.844**
C1	.877	.874	.790
C2	.906	.913	.707
C3	.846	.843	.692

Table 3: F1 scores of *gpt-4-turbo* by level context information added to the prompt (C1–C3). Level C0 corresponds to the prompt of optimal verbosity level with no context information (V3 for CB and V1 for POLI).

The amount of co-text included in the prompt is the same as what was originally shown to the annotators (cf. Table 1). Speaker information pertains to the speaker's identity and political party. The context information was appended to the prompts (cf. Appendix A for a detailed description).

We only extend the prompts with optimal verbosity level (V3 for CB and V1 for POLI). Table 3 shows the F1 scores by context levels for *gpt-4-turbo*. Surprisingly, expanding the prompt with co-text and speaker information did not improve the model's accuracy on either dataset. This could be because context information is not required for these datasets, or because our specific prompt structure might be suboptimal and confuse the model. We leave further investigation for future work.

4.3 Rank-based evaluation

In light of resource constraints, fact-checking organizations have devised principles to prioritize claims based on their check-worthiness. This invites the question of whether zero- and few-shot LLM prompting could be used for that purpose. To investigate this, we frame CW as a ranking task and rank the claims based on the LLM's confidence for the positive class. We used token likelihood of the positive class as a measure of confidence. The quality of the so-obtained ranking will depend on how well the LLM is calibrated. Thus, we first evaluate the LLMs' calibration accuracy using the expected calibration error (ECE). Figure 1 shows the predictive accuracy (F1 score) against calibration accuracy ($1-$ECE) across datasets and prompt verbosity levels (we only use prompts at context level C0, i.e., we add no context information). For each dataset, we select the prompt that scores high on both predictive and calibration accuracy. The prompts with the highest F1 scores are usually also the best-calibrated ones, except for NEWS, where we select level V1 as Pareto-optimal.

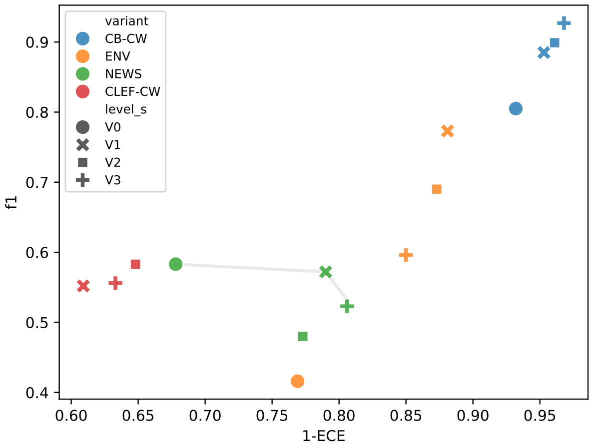

Figure 1: F1 scores and calibration accuracy for the CW task, across datasets and levels of prompt verbosity

	CB	CLEF	ENV	NEWS
AP	.951	.469	.767	.67
P@10	1	.5	.9	1
P@R	.924	.487	.761	.533

Table 4: Rank-based performance scores for the CW task: average precision (*AP*), precision-at-10 (*P@10*), and precision-at-R (*P@R*)

Table 4 shows the rank-based performance scores for the selected prompts on *gpt-4-turbo*: average precision (AP), precision-at-10 (P@10), and precision-at-R, where R equals the total number of positives in the dataset. The rank-based performance scores mirror the classification accuracy scores: they are high for datasets with high predictive accuracy (CB and ENV) and lower for datasets with lower predictive accuracy (NEWS and CLEF). Our results suggest that LLM models with high predictive accuracy also produce well-calibrated scores using ECE and may be readily used as check-worthiness rankers.

5 Conclusion

We addressed the claim detection and check-worthiness tasks using zero- and few-shot LLM prompting based on existing annotation guidelines. The optimal level of prompt verbosity, ranging from minimal prompts to detailed prompts that include criteria and examples, depends on the domain and style of guidelines. Adding claim context (co-text and speaker information) does not improve the performance. For models with high predictive accuracy, confidence scores can be directly used to produce reliable check-worthiness rankings.

Limitations

Datasets. In our experiments, we do not use datasets created by fact-checking organizations. While the datasets were created specifically for the tasks of CD and CW, and most were annotated by experts, the datasets were constructed for research purposes. To most accurately evaluate the potential of using our approach in fact-checking organizations, a dataset annotated according to official factuality or check-worthiness criteria with appropriate annotation guidelines should be used.

Models. Due to hardware constraints, no open-source LLMs of a larger size were used in our experiments. We acknowledge the importance of relying on open-source models in the research community, and the lack of insight that results from disregarding larger open-source models. Using closed-source models has the additional caveat of possible leakage of the dataset, which is a growing concern in the community (Balloccu et al., 2024). We also note that the outstanding results on the ClaimBuster dataset (CB) could be due to data leakage, considering the dataset was published several years ago and has a wide reach in the research of automatic fact-checking.

Languages. In this work, we only do experiments on datasets in English. This is for two reasons: (1) the necessity to understand the annotation guidelines to draft prompts using them and (2) the lack of datasets in other languages. However, we acknowledge that disinformation is a global problem and that tackling it requires working with multiple languages.

CLEF&NEWS. The results of weak performance on the CLEF and NEWS datasets could lie in the worthiness criteria used and the way the criteria are articulated. In CLEF, a Likert scale was used for the annotation. However, the levels in the scale do not completely correspond to gradation, as is usually the case. The negative labels include both tweets that do not need fact-checking (the label "No, no need to check") and those worth fact-checking but not requiring experts' attention (the label "No, too trivial to check"). This distinction probably creates ambiguity for the model, as demonstrated by poor predictive accuracy and calibration. On the NEWS dataset, the CW task essentially amounts to topic classification. It is unclear how the model should handle sentences unrelated to the provided topics. The authors in the corresponding research paper report performance scores with F1 not exceeding .7, even for annotators.

Risks

Although we intend to combat the spread of disinformation with this work, there is still a potential for misuse. The prompts and insights reported in this work could potentially be used to create disinformative claims adapted to make their detection more difficult. A big challenge of disinformation detection is the growing use of generative models for creating disinformative claims. The prompts provided in this work could be reverted for generative purposes, achieving the exact opposite effect than what our work aims to achieve.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

FullFact. 2020. The challenges of online fact checking.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017b. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Martin Hyben, Sebastian Kula, Ivan Srba, Robert Moro, and Jakub Simko. 2023. Is it indeed bigger better? the comprehensive study of claim detection lms applied for disinformation tackling.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2).

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *CLEF 2022: Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*, pages 368–392. CEUR Workshop Proceedings (CEUR-WS.org).

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators.

Marcin Sawinski, Krzysztof Wecel, Ewelina Ksiezniak, Milena Stróżyna, Wlodzimierz Lewoniewski, Piotr Stolarski, and Witold Abramowicz. 2023. Openfact at checkthat!-2023: Head-to-head GPT vs. BERT - A comparative study of transformers language models for the detection of check-worthy claims. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 453–472. CEUR-WS.org.

Ghazaal Sheikhi, Samia Touileb, and Sohail Khan. 2023. Automated claim detection for fact-checking: A case study using Norwegian pre-trained language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–9, Tórshavn, Faroe Islands. University of Tartu Library.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

A Dataset Information

In this section, we provide details on the datasets used in our experiments. In

A.1 Test set selection

Here, we provide details on the test set selection for each dataset. Furthermore, we state which set the authors used for evaluation and whether the results can be comparable.

ClaimBuster. The dataset does not have an explicit test set. The authors instead used 4-fold cross-validation on different-sized subsets during their experiments (4,000, 8,000 ... 20,000). However, a high-quality *groundtruth* set is available in the dataset. It contains 1,032 samples that experts agreed on and was used for screening during annotation. Also, all the test sets the authors used contain the screening sentences. For the quality of labels and to have somewhat comparable results to the authors, we selected the *groundtruth* set for experiments.

CLEF. The dataset consists of both a *dev* and a *test* set. Since the *test* set was used to evaluate teams participating in the CLEF CheckThat! the challenge, we opted to do our experiments on this set to compare to the metrics of the best-submitted solution.

EnvironmentalClaims. The dataset contains both a *dev* and *test* set of equal size, whereas the original work publishes metrics on both sets separately. We selected the *test* set for our experiments.

NewsClaims. The dataset provides both a *dev* and a *test* set; however, the disclosed sets contain only positive instances. The complete dataset consists of around 10% of positive instances, with a high number of low-quality negative instances created by errors in sentencizing and filtering – instances containing only names, dates, links. The dataset also contains duplicate instances, also in the set of positives. To create a viable subset and avoid high costs during inference, we sampled the negative instances from a normal distribution with the parameters fitted to the length of the instances. We chose to sample the same number of instances as there are positives without duplicates, creating a higher baseline.

PoliClaim. The dataset provides an explicit *test* set consisting of both gold labels and labels resulting from inference on 4 political speeches. To be able to compare results, we opted to use the complete *test* set.

A.2 Original metrics

ClaimBuster. As previously mentioned, the authors used 4-fold cross-validation on different-sized subsets during their experiments (4,000, 8,000 ... 20,000). They evaluate using $f_w avg$ with the highest score of .818. Our highest scores on $f_w avg$ are .933 on *gpt-4-turbo* and .906 on *gpt-3.5-turbo*, which is a significant improvement. The authors also evaluate ranking, where our results improve on P@k.

CLEF. The best results of the Task 1 on CLEF CheckThat!2022 were accuracy of .761 for claim detection, and the F1 of .698 on check-worthiness detection. While our approach underperforms for check-worthiness detection (F1 of .583), it achieves higher accuracy for claim detection (.776 on Level V2).

EnvironmentalClaims. The authors report F1 on the test set, the highest achieved is .849. Our approach achieves .773 for Level V1.

NewsClaims. The authors report F1 on the whole dataset. However, it is inconclusive whether it is evaluated on the binary or multiclass labels. They report the highest F1 of .309, which our approach surpasses on the subset we selected, with F1 of .583.

PoliClaim. The authors evaluate on accuracy. On the *test* set, they achieve an accuracy of .764 on *gpt-3.5* and .862 on *gpt-4*. Our approach achieves the same maximum accuracy using *gpt-3.5-turbo* with prompt Level V3, but lower accuracy for *gpt-4*.

A.3 Context information

ClaimBuster. During the annotation of the ClaimBuster dataset, 4 preceding statements could be viewed with an extra button, which was used in 14% of all cases. Since the dataset covers presidential debates with multiple speakers, including the moderator and audience questioners, it is not completely clear how the speakers were differentiated in the provided preceding sentences. Therefore, we selected the method of differentiating the speakers arbitrarily – 'A' was used for the speaker of the statement that is meant to be annotated, and 'B' for the opposing speaker.

EnvironmentalClaims. No additional contextual or co-textual information was provided in the dataset. The annotators were not shown any context during annotation due to budget. The authors considered annotating whole paragraphs instead of sentence-level annotation but decided against it due to time and budget constraints.

PoliClaim. The annotators were provided with the preceding and following sentences of the one they are annotating. Since there is only one speaker (as opposed to ClaimBuster, which covers debates), there is no need for denoting the speaker, minimizing confusion in prompts. In annotation guidelines, context was explicitly mentioned, as well as clarified in examples. In our experiments, we used two versions of the prompts – one mentioning context for experiments with co-text expansion and one without the mention of context used when only one sentence from the speech is provided. The two alternatives are shown in E.

	CB		CLEF		ENV	NEWS	POLI
	CD	CW	CD	CW	CD	CW	CW
V0	.853	.718	.656	.496	.484	**.531**	.707
V1	.57	.739	.745	.438	.71	.371	.751
V2	.774	.800	.719	.468	**.701**	.348	.657
V3	**.872**	**.862**	**.757**	.445	.65	.206	**.803**

Table 5: F1 scores of *gpt-3.5-turbo* for CD and CW tasks across datasets broken down by prompt verbosity level (V1–V3). Level V0 corresponds to the naive-prompting baseline.

CLEF. The dataset consists of tweets covering COVID-19 topics. For the check-worthiness task, annotators were shown metadata such as time, account, number of likes and reposts. However, this information is not readily available in the dataset and requires crawling the tweets to obtain it. It was also not available in the dataset of the CLEF2022 CheckThat! Challenge, which was derived from the original dataset. Since we wanted to make our effort comparable to alternative methods used in the competition, we did not opt for crawling the tweets to acquire metadata.

NewsClaims. The research paper introducing the dataset has inconsistencies regarding the co-text provided to annotators. While it is stated in the paper that whole articles are provided for co-text, in the screenshot of the annotation platform, only three preceding and following sentences were provided. Regarding context, the work emphasizes the importance of metadata such as claim object, speaker and span, and provides that data for positive instances (sentences containing claims related to 4 specified COVID-19 subtopics). The effort of annotating the claims with metadata is worthwhile, however we decided against using it in inference since no such data is available for negative instances.

B Alternative Results

In Table 5 we present the results of prompt expansion for *gpt-3.5-turbo-0125*. In general, *gpt-4-turbo* outperforms this model, except for POLI, where Level V3 on *gpt-3.5-turbo* achieves the highest F1 score of .803.

C Model Information

For OpenAI models, we use *gpt-3.5-turbo-0125* and *gpt-4-0125-preview*. We use a temperature of 0 for all experiments. To get confidence, we use

	CB		CLEF		POLI	ENV	NEWS
	CD	CW	CD	CW	CD	CW	CW
V0	.094	.068	.259	.601	.231	.322	.142
V1	.050	.047	.196	.391	.119	.210	.271
V2	.043	.039	.194	.352	.127	.277	.373
V3	.039	.032	.222	.367	.150	.194	.348

Table 6: ECE score by prompt level per dataset for *gpt-4-turbo*. 'CD' and 'CW' mark claim detection and claim check-worthiness detection, respectively, while 'V0' marks the score for the naive baseline

	CB		CLEF		ENV	NEWS	POLI
	CD	CW	CD	CW	CD	CW	CW
V0	.033	.068	.212	.359	.189	.246	.257
V1	.323	.085	.386	.609	.088	.260	.229
V2	.103	.071	.279	.560	.097	.280	.327
V3	.061	.050	.285	.646	.100	.379	.196

Table 7: ECE score by prompt level per dataset for *gpt-3.5-turbo*. 'CD' and 'CW' mark claim detection and claim check-worthiness detection, respectively, while 'V0' marks the score for the naive baseline

logprobs and *n_probs=5*, to account for the target labels ending up as less probable tokens. We use a random seed of 42 in all experiments, to avoid stochastic answers as much as possible. The run was executed once per model and prompt variant. Inference was done through the OpenAI API. GPU hours are hard to estimate.

The initial experiments on open-source models were done on *neural-chat:7b-v3.3-q5_K_M* and *mistral:7b-instruct-v0.2-q5_K_M*. A total of 5 GPU hours was used.

D Calibration

In this section, the ECE per prompt verbosity level is shown for *gpt-3.5-turbo* (Table 7), and *gpt-4-turbo* (Table 6). The ECE is calculated with the parameters $n_bins = 10$ and $norm = l1$.

E Complete prompts

This section provides the complete prompts used in our experiments. The instructions were given in system prompts, while the instances were in user prompts. The added context information is also appended to user prompts.

For each dataset, the three prompt levels are shown, with the content expanded in relation to the previous level highlighted. To visually separate the levels, Level V2 is highlighted in yellow, while

Level V3 is highlighted in pink.

For CLEF, two alternative prompts are given, since for CD and CW different annotation guidelines were used. For POLI, parts of the Level V2 and Level V3 prompts regarding surrounding sentences are either provided or not, based on whether context expansion is used (surrounding sentences are given in prompts C1 and C3). Those parts are highlighted in blue.

Baseline prompts. Two naive zero-shot prompts serve as a baseline. For the claim detection task, the baseline is:

Does the following <sentence> contain a factual claim? Answer only with Yes or No.

For the claim check-worthiness detection task, the baseline is:

Does the following <sentence> contain a check-worthy claim? Answer only with Yes or No.

User prompts. The user prompts were based on how the instance was reffered to in the corresponding annotation gudelines. The instances are surrounded with HTML tags. The same is done for context expansion on CB and POLI.

Level	Prompt
V1	Categorize the <sentence> spoken in the presidential debates into one of three categories: Non-Factual Sentence (NFS), Unimportant Factual Sentence (UFS) or Check-worthy Factual Sentence (CFS). Use only one of the three labels (NFS, UFS or CFS), do not provide any additional explanation.
V2	Categorize the <sentence> spoken in the presidential debates into three categories: Non-Factual Sentence (NFS): Subjective sentences (opinions, beliefs, declarations) and many questions fall under this category. These sentences do not contain any fac tual claim. Unimportant Factual Sentence (UFS): These are factual claims but not check-worthy. The general public will not be interested in knowing whether these sentences are true or false. Fact-checkers do not find these sentences as important for checking. Check-worthy Factual Sentence (CFS): They contain factual claims and the general pub lic will be interested in knowing whether the claims are true. Journalists look for these type of claims for fact-checking. Use only one of the three labels (NFS, UFS and CFS), do not provide any additional explanation.
V3	Categorize the <sentence> spoken in the presidential debates into three categories: Non-Factual Sentence (NFS): Subjective sentences (opinions, beliefs, declarations) and many questions fall under this category. These sentences do not contain any factual claim. Here are two such examples. ”But I think it’s time to talk about the future.“ “You remember the last time you said that?” Unimportant Factual Sentence (UFS): These are factual claims but not check-worthy. The general public will not be interested in knowing whether these sentences are true or false. Fact-checkers do not find these sentences as important for checking. Some examples are as fol lows. “Next Tuesday is Election day.” “Two days ago we ate lunch at a restaurant.” Check-worthy Factual Sentence (CFS): They contain factual claims and the general public will be interested in knowing whether the claims are true. Journalists look for these type of claims for fact-checking. Some examples are: “He voted against the first Gulf War.” “Over a million and a quarter Americans are HIV-positive.” Use only one of the three labels (NFS, UFS and CFS), do not provide any additional explanation.

Table 8: System prompts used for inference on the ClaimBuster dataset.

Level	Prompt
V1	Your task is to label the <sentence>. The information I need is whether it is an environmental claim. A broad definition for such a claim is given by the European Commission: Environmental claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services. Answer only with Yes or No.
V2	Your task is to label the <sentence>. The information I need is whether it is an environmental claim. A broad definition for such a claim is given by the European Commission: Environmental claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services. General principles: You will be presented with a <sentence> and have to decide whether the <sentence> contains an explicit environmental claim. Do not rely on implicit assumptions when you decide on the label. Base your decision on the information that is available within the sentence. However, if a sentence contains an abbreviation, you could consider the meaning of the abbreviation before assigning the label. In case a sentence is too technical/complicated and thus not easily understandable, it usually does not suggest to the average consumer that a product or a service is environmentally friendly and thus can be rejected. Likewise, if a sentence is not specific about having an environmental impact for a product or service, it can be rejected. Answer only with Yes or No.
V3	Your task is to label the <sentence>. The information I need is whether it is an environmental claim. A broad definition for such a claim is given by the European Commission: Environmental claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services. General principles: You will be presented with a sentence and have to decide whether the sentence contains an explicit environmental claim. Do not rely on implicit assumptions when you decide on the label. Base your decision on the information that is available within the sentence. However, if a sentence contains an abbreviation, you could consider the meaning of the abbreviation before assigning the label. In case a sentence is too technical/complicated and thus not easily understandable, it usually does not suggest to the average consumer that a product or a service is environmentally friendly and thus can be rejected. Likewise, if a sentence is not specific about having an environmental impact for a product or service, it can be rejected. Examples: <sentence>: Farmers who operate under this scheme are required to dedicate 10% of their land to wildlife preservation. Label: Yes Explanation: Environmental scheme with details on implementation. <sentence>: UPM Biofuels is developing a new feedstock concept by growing Brassica Carinata as a sequential crop in South America. Label: No Explanation: Sentence context would be required to understand whether it is a claim. Answer only with Yes or No, don't provide any additional explanation.

Table 9: System prompts used for inference on the EnvironmentalClaims dataset.

Level	Prompt
V1	A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Does the <tweet> contain a verifiable factual claim? Answer only with Yes or No, don't provide any additional explanation.
V2	A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Factual claims include the following: Stating a definition; Mentioning quantity in the present or the past; Making a verifiable prediction about the future; Reference to laws, procedures, and rules of operation; References to images or videos (e.g., "This is a video showing a hospital in Spain."); Statements about correlations or causations. Such correlation and causation needs to be explicit, i.e., sentences like "This is why the beaches haven't closed in Florida." is not a claim because it does not say why explicitly, thus it is not verifiable. Tweets containing personal opinions and preferences are not factual claims. Note: if a tweet is composed of multiple sentences or clauses, at least one full sentence or clause needs to be a claim in order for the tweet to contain a factual claim. If a claim exist in a sub-sentence or sub-clause then tweet is not considered to have a factual claim. For example, "My new favorite thing is Italian mayors and regional presidents LOSING IT at people violating quarantine" is not a claim, how ever, it is an opinion. Moreover, if we consider "Italian mayors and regional pres idents LOSING IT at people violating quarantine" it would be a claim. In addition, when answering this question, annotator should not open the tweet URL. Does the <tweet> contain a verifiable factual claim? Answer only with Yes or No.
V3	A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Factual claims include the following: Stating a definition; Mentioning quantity in the present or the past; Making a verifiable prediction about the future; Reference to laws, procedures, and rules of operation; References to images or videos (e.g., "This is a video showing a hospital in Spain."); Statements about correlations or causations. Such correlation and causation needs to be explicit, i.e., sentences like "This is why the beaches haven't closed in Florida." is not a claim because it does not say why explicitly, thus it is not verifiable. Tweets containing personal opinions and preferences are not factual claims. Note: if a tweet is composed of multiple sentences or clauses, at least one full sentence or clause needs to be a claim in order for the tweet to contain a factual claim. If a claim exist in a sub-sentence or sub-clause then tweet is not considered to have a factual claim. For example, "My new favorite thing is Italian mayors and regional presidents LOSING IT at people violating quarantine" is not a claim, however, it is an opinion. Moreover, if we consider "Italian mayors and regional presidents LOSING IT at people violating quarantine" it would be a claim. In addition, when answering this question, annotator should not open the tweet URL. Does the <tweet> contain a verifiable factual claim? Answer only with Yes or No. Examples: Tweet: Please don't take hydroxychloroquine (Plaquenil) plus Azithromycin for COVID19 UNLESS your doctor prescribes it. Both drugs affect the QT interval of your heart and can lead to arrhythmias and sudden death, especially if you are tak ing other meds or have a heart condition. Label: Yes Explanation: There is a claim in the text. Tweet: Saw this on Facebook today and it's a must read for all those idiots clearing the shelves coronavirus toiletpapercrisis auspol Label: No Explana tion: There is no claim in the text. Answer only with Yes or No, don't provide any additional explanation.

Table 10: System prompts used for inference on the CLEF dataset for claim detection.

Level	Prompt
V1	It is important that a verifiable factual check-worthy claim be verified by a professional fact-checker, as the claim may cause harm to society, specific person(s), company(s), product(s), or some government entities. However, not all factual claims are important or worth fact-checking by a professional fact-checker, as this very time-consuming. Do you think that a professional fact-checker should verify the claim in the <tweet>? Labels: No, no need to check; No, too trivial to check; Yes, not urgent; Yes, very urgent. Decide on one label. Then, answer only with Yes or No.
V2	It is important that a verifiable factual check-worthy claim be verified by a professional fact-checker, as the claim may cause harm to society, specific person(s), company(s), product(s), or some government entities. However, not all factual claims are important or worth fact-checking by a professional fact-checker, as this very time-consuming. Do you think that a professional fact-checker should verify the claim in the <tweet>? Labels: No, no need to check: the tweet does not need to be fact-checked, e.g., be- cause it is not interesting, a joke, or does not contain any claim. No, too trivial to check: the tweet is worth fact-checking, how ever, this does not require a professional fact-checker, i.e., a non-expert might be able to fact-check the claim. For example, one can verify the information using reli able sources such as the official website of the WHO, etc. An example of a claim is as follows: "The GDP of the USA grew by 50% last year." Yes, not urgent: the tweet should be fact-checked by a professional fact-checker, however, this is not urgent or critical; Yes, very urgent: the tweet can cause immediate harm to a large number of people; therefore, it should be verified as soon as possible by a professional fact-checker; Decide on one label. Then, answer only with Yes or No.
V3	It is important to verify a factual claim by a professional fact-checker, which can cause harm to the society, specific person(s), company(s), product(s) or government entities. However, not all factual claims are important or worthwhile to be fact-checked by a professional fact-checker as it is a time-consuming procedure. Do you think that a professional fact-checker should verify the claim in the <tweet>? Labels: No, no need to check: the tweet does not need to be fact-checked, e.g., be- cause it is not interesting, a joke, or does not contain any claim. No, too trivial to check: the tweet is worth fact-checking, however, this does not require a professional fact-checker, i.e., a non-expert might be able to fact-check the claim. For example, one can verify the information using reliable sources such as the official website of the WHO, etc. An example of a claim is as follows: "The GDP of the USA grew by 50Yes, not urgent: the tweet should be fact-checked by a professional fact-checker, however, this is not urgent or critical; Yes, very urgent: the tweet can cause immediate harm to a large number of people; therefore, it should be verified as soon as possible by a professional fact-checker; Examples: Tweet: Wash your hands like you've been chopping jalapeños and need to change a contact lens" says BC Public Health Officer Dr. Bonnie Henry re. ways to protect against #coronavirus

Table 11: System prompts used for inference on the CLEF dataset for claim check-worthiness detection.

Level	Prompt
V1	The task is to select verifiable statements from political speeches for fact-checking. Given a <statement> from a political speech, answer the question. Does the <statement> explicitly present any verifiable factual information? Answer with A, B or C only. A - Yes, B - Maybe, C - No.
V2	The task is to select verifiable statements from political speeches for fact-checking. Given a <statement> from a political speech, answer the question following the guidelines. Definitions and guidelines: Fact: A fact is a statement or assertion that can be objectively verified as true or false based on empirical evidence or reality. Opinion: An opinion is a judgment based on facts, an attempt to draw a reasonable conclusion from factual evidence. While the underlying facts can be verified, the derived opinion remains subjective and is not universally verifiable. Context: Make sure to consider a small context of the target statement (the previous and next sentence) when annotating. Some statements require context to understand the meaning. Factual claim: A factual claim is a statement that explicitly presents some verifiable facts. Statements with subjective components like opinions can also be factual claims if they explicitly present objectively verifiable facts. Opinion with Facts: Opinions can also be based on factual information. When does an opinion explicitly present a fact: Many opinions are more or less based on some factual information. However, some facts are explicitly presented by the speakers, while others are not. What is verifiable: The verifiability of the factual information depends on how specific it is. If there is enough specific information to guide a general fact-checker in checking it, the factual information is verifiable. Otherwise, it is not verifiable. The question: Does the <statement> explicitly present any verifiable factual information? Answer with A, B or C only. A - Yes, the statement contains factual information with enough specific details that a fact-checker knows how to verify it. E.g., Birmingham is small in population compared to London. B - Maybe, the statement seems to contain some factual information. However, there are certain ambiguities (e.g., lack of specificity) making it hard to determine the verifiability. E.g., Birmingham is small compared to London. (lack of details about what standard Birmingham is small) C - No, the statement contains no verifiable factual information. Even if there is some, it is clearly unverifiable. E.g., Birmingham is small.

Table 12: System prompts of Level V1 and Level V2 used for inference on the PoliClaim dataset for claim check-worthiness detection. The blue highlight shows instructions for regarding context.

Level	Prompt
V3	The task is to select verifiable statements from political speeches for fact-checking. Given a statement from a political speech and its context, answer the question following the guidelines. Definitions and guidelines: Fact: A fact is a statement or assertion that can be objectively verified as true or false based on empirical evidence or reality. Opinion: An opinion is a judgment based on facts, an attempt to draw a reasonable conclusion from factual evidence. While the underlying facts can be verified, the derived opinion remains subjective and is not universally verifiable. Factual claim: A factual claim is a statement that explicitly presents some verifiable facts. Statements with subjective components like opinions can also be factual claims if they explicitly present objectively verifiable facts. Context: Make sure to consider a small context of the target statement (the previous and next sentence) when annotating. Some statements require context to understand the meaning. For example: E1. "... Just consider what we did last year for the middle class in California, sending 12 billion dollars back - the largest state tax rebate in American history. <statement> But we didn't stop there. <> We raised the minimum wage. We increased paid sick leave. Provided more paid family leave. Expanded child care to help working parents..." Without the context, the sentence marked with <statement> seems an incomplete sentence. With the context, we know the speaker is claiming a bunch of verifiable achievements of their administration. E2. "... When I first stood before this chamber three years ago, I declared war on criminals and asked for the Legislature to repeal and replace the catch-and-release policies in SB 91. <statement> With the help of many of you, we got it done. <> Policies do matter. We've seen our overall crime rate decline by 10 percent in 2019 and another 18.5 percent in 2020! ..." The part marked with <statement> claims that the policies against crimes have been "done", which is verifiable. It needs context to understand it. Opinion with Facts: Opinions can also be based on factual information. For example: E1. "I am proud to report that on top of the local improvements, the state has administered projects in almost all 67 counties already, and like I said, we've only just begun." The speaker's "proud of" is a subjective opinion. However, the content of pride (administered projects) is factual information. E2. "I first want to thank my wife of 34 years, First Lady Rose Dunleavy." The speaker expresses their thankfulness to their wife. However, there is factual information about the first lady's name and the length of their marriage. When does an opinion explicitly present a fact: Many opinions are more or less based on some factual information. However, some facts are explicitly presented by the speakers, while others are not. Explicit presentation means the fact is directly entailed by the opinion without extrapolation: E1. "The pizza is delicious." This opinion seems to be based on the fact that "pizza is a kind of food". However, this fact is not explicitly presented. E2. "I first want to thank my wife of 34 years, First Lady Rose Dunleavy." The name of the speaker's wife and their year of marriage are explicitly presented. What is verifiable: The verifiability of the factual information depends on how specific it is. If there is enough specific information to guide a general fact-checker in checking it, the factual information is verifiable. Otherwise, it is not verifiable. E1. "Birmingham is small." is not verifiable because it lacks any specific information for determining veracity. It leans more toward subjective opinion. E2. "Birmingham is small, compared to London" is more verifiable than E1. A fact-checker can retrieve the city size, population size ...etc., of London and Birmingham to compare them. However, what to compare to prove Birmingham's "small" is not specific enough. E3. "Birmingham is small in population size, compared to London" is more verifiable than E1 and E2. A fact-checker now knows it is exactly the population size to be compared. The question: Does the <statement> explicitly present any verifiable factual information? Answer with A, B or C only. A - Yes, the statement contains factual information with enough specific details that a fact-checker knows how to verify it. E.g., Birmingham is small in population compared to London. B - Maybe, Maybe, the statement seems to contain some factual information. However, there are certain ambiguities (e.g., lack of specificity) making it hard to determine the verifiability. E.g., Birmingham is small compared to London. (lack of details about what standard Birmingham is small) C - No, the statement contains no verifiable factual information. Even if there is some, it is clearly unverifiable. E.g., Birmingham is small.

Table 13: System prompts of Level V3 used for inference on the PoliClaim dataset for claim check-worthiness detection. The blue highlight shows instructions for regarding context.