Identifying Factual Inconsistencies in Summaries: Grounding Model Inference via Task Taxonomy

Anonymous ACL submission

Abstract

Factual inconsistencies pose a significant hurdle for the faithful summarization by generative models. While a major direction to enhance inconsistency detection is to derive stronger Natural Language Inference (NLI) models, we propose an orthogonal aspect that underscores the importance of incorporating task-specific taxonomy into the inference. To this end, we consolidate key error types of inconsistent facts in summaries, and incorporate them to facilitate both the zero-shot and supervised paradigms of LLMs. Extensive experiments on ten datasets of five distinct domains suggest that, zero-shot LLM inference could benefit from the explicit solution space depicted by the error type taxonomy, and achieves state-of-the-art performance overall, surpassing specialized non-LLM baselines, as well as recent LLM baselines. We further distill models that fuse the taxonomy into parameters through our designed prompt completions and supervised training strategies, efficiently substituting state-of-the-art zero-shot inference with much larger LLMs.

1 Introduction

As abstractive summarization has been advanced significantly via generative models such as BART (Lewis et al., 2020) and Large Language Models (LLMs), factual inconsistencies remain one of the key concerns for ensuring high-quality faithful summaries (Maynez et al., 2020a; Kryscinski et al., 2020; Goyal et al., 2023), where certain facts from the summary are not aligned with those presented in the original document. Previous works have studied extensively that employ various paradigms to reason inconsistencies, ranging from specialized BERT-variants (Devlin et al., 2019) such as DAE (Goyal and Durrett, 2020), QAFactEval (Fabbri et al., 2022), to recent LLMs equipped with general comprehension capabilities (Luo et al., 2023; Wang et al., 2023; Liu et al., 2023a).

In particular, one outstanding direction for factual inconsistency detection is to frame it as a Natural Language Inference (NLI) problem, assessing the entailment between the document and summary (Bowman et al., 2015). Intuitively, irrelevant or inconsistent facts in the summary should reflect a low level of entailment through NLI models. Prior to LLMs, BERT-based NLI models have been successfully practiced by approaches such as SummaC (Laban et al., 2022) to identify summary inconsistencies. In this new era of LLMs, several pioneering works have shown that zero-shot prompting of LLMs is already effective with NLI-style scoring, where LLMs directly classify the summary consistency or provide a consistency score (Luo et al., 2023; Wang et al., 2023; Liu et al., 2023a).

While it is a promising direction to keep enhancing factual inconsistency recognition by deriving stronger NLI models, such as FactCC (Kryscinski et al., 2020), DocNLI (Yin et al., 2021), FalseSum (Utama et al., 2022), AMRFact (Qiu et al., 2024), in this work, we propose approaches from an orthogonal aspect, which examines the incorporation of explicit solution space into the inference, such that either zero-shot prompting or a trained model reaches decisions according to explicit task-specific cues, i.e. an explicit error type taxonomy.

Our motivation stems from the distinct nature between summary inconsistencies and NLI: summaries are grounded by the original document, thus leaning towards *reiteration*, whereas NLI tackles a broader problem that involves *extrapolation*. Since we roughly deem the scope of summary inconsistency detection as smaller, one can consolidate and leverage its task-specific taxonomy to rationalize a more effective and explainable inference.

As there exist numerous annotation schemas adopted by previously introduced datasets, factual errors are firstly unified into a fine-grained taxonomy by AGGREFACT (Tang et al., 2023), which we consolidate upon and identify five common error types that are salient for recognizing summary inconsistencies, including *Predicate Error*, *Entity Error*, *Circumstantial Error*, *Coreference Error* and *Addition Error* (Section 3), covering a wide variety of datasets (Table 1). The identified error types are then utilized to anchor the inference of factual inconsistencies. Specifically, we examine their efficacy with LLMs in both zero-shot and supervised paradigms, and demonstrate the utility of task-specific taxonomy in complementary to the sole NLI-style classification.

For the zero-shot setting (Section 4), we craft the instruction tailored for each error type in the prompt, directing LLMs to reason specific error types according to the given guidance. To handle long summaries, we additionally propose a window-based prompting scheme, as an effective alternative to the vanilla prompting. For a comprehensive evaluation, our experiments are conducted on 10 datasets across five domains, including summarization on different news sources, daily or professional dialogues, official reports and narrative stories. Moreover, we employ models from OpenAI (ChatGPT, GPT-4o) along with strong open-source LLMs including Llama-3 (Touvron et al., 2023) and Mistral (Jiang et al., 2023) to ensure a robust conclusion. Empirical results suggest that our proposed methods surpass all baselines, including 7 non-LLM baselines and 4 LLM baselines, showing that zero-shot LLM inference could benefit from a grounded solution space by depicting the task taxonomy in the instruction. Our proposed methods, termed Factuality with Taxonomy (FACTAX), achieve the best overall performance across five domains; especially, FACTAX with ChatGPT outperforms previous state-of-the-art zero-shot baseline G-Eval with GPT-4 (Liu et al., 2023a; Qiu et al., 2024) on the AGGREFACT-FTSOTA benchmark (Tang et al., 2023).

We then further seek to distill a model that fuses the task taxonomy into model parameters through supervised training. By unifying the error types of previous independently introduced datasets, we regard them jointly as training resources. Llama3-8B models are trained to learn binary decisions as well as to recognize specific error types on summaries, through our designed completions and training strategies. The resulting trained model outperforms previous supervised baselines, and is able to match the best zero-shot inference performance, effectively acting as an efficient alternative to zero-shot reasoning with much larger LLMs.

Overall, our key contributions in this work are:

- We underscore the importance of a fine-grained task taxonomy for the inference of summary inconsistencies, leading to enhanced performance and interpretability upon vanilla reasoning.
- We pinpoint key error types and incorporate them into our designed zero-shot prompting schemes, anchoring LLM reasoning within an explicit solution space. Experiments on diverse datasets and LLMs demonstrate its efficacy over baselines.
- We further distill a model that rationalizes the task taxonomy into parameters through our supervised training strategies, practically substituting zero-shot reasoning with SOTA performance.

2 Related Work

Factual Inconsistency Evaluation Datasets Numerous datasets for evaluating factual inconsistencies in summaries have been independently introduced in recent years. Among these, many focus on the news domain, primarily addressing CNN/DailyMail summaries (Nallapati et al., 2016), such as FactCC (Kryscinski et al., 2020), FRANK (Pagnoni et al., 2021) and SummEval (Fabbri et al., 2021); others addressing XSum summaries (Narayan et al., 2018) constructed upon BBC news, such as XSumFaith (Maynez et al., 2020b) and DeFacto (Liu et al., 2023b); some also addressing both, such as CLIFF (Cao and Wang, 2021) and Goyal and Durrett (2021).

Apart from news, several datasets focus on dialogue summaries, especially daily dialogues from SAMSum (Gliwa et al., 2019), such as DiaSummEval (Gao and Wan, 2022), FactEval (Wang et al., 2022b), DiaSummFactCorr (Gao et al., 2023). DiaSummFact (Zhu et al., 2023) also assesses meeting summaries from QMSum (Zhong et al., 2021).

Recent datasets have also been proposed to address more domains, e.g. Koh et al. (2022) evaluates factual consistency on official reports from GovReport (Huang et al., 2021); LongEval (Krishna et al., 2023) addresses story summaries from SQuALITY (Wang et al., 2022a).

In this work, we aim for robust evaluation across diverse domains under the same task requirement, especially for zero-shot methods that should generalize across different types of documents.

Non-LLM Approaches State-of-the-art models prior to LLMs mainly focus around two directions. The first is to effectively leverage NLI models to assess the entailment between the document-

	Domain	Doc Len	Summ Len	# Summ	Ent.	Pred.	Circ.	Coref.	AddE.
Polytope (**Huang et al., 2020**)	CNN/DM	573.2	64.8	1268	-	-	-	-	-
SummEval (**Fabbri et al., 2021**)	CNN/DM	363.6	62.8	1698	-	-	-	-	-
FRANK (**Pagnoni et al., 2021**)	CNN/DM	476.2	40.6	2246	✓	✓	✗	✗	✓
BUMP (**Ma et al., 2023**)	CNN/DM	686.4	52.5	1087	✓	✓	✓	✓	✓
CLIFF (**Cao and Wang, 2021**)	CNN/DM & XSum	453.4	35.6	600	✓	✓	✗	✗	✓
XsumFaith (**Maynez et al., 2020b**)	XSum	381.1	19.2	2353	✓	✓	✗	✗	✓
QAGS/Wang'20 (**Wang et al., 2020**)	XSum	324.5	33.3	474	-	-	-	-	-
Goyal'21 (**Goyal and Durrett, 2021**)	XSum	430.3	21.8	150	✓	✓	✗	✗	✓
Cao'22 (**Cao et al., 2022**)	XSum	349.4	25.3	696	-	-	-	-	-
DiaSumFact (**Zhu et al., 2023**)	Dialogues	187.0	43.7	475	✓	✓	✓	✓	✓
DiaSummEval (**Gao and Wan, 2022**)	Dialogues	109.5	22.6	474	-	-	-	-	-
DiaSummFactCorr (**Gao et al., 2023**)	Dialogues	113.1	20.8	4000	✓	✓	✓	✓	✓
FacEval (**Wang et al., 2022b**)	Dialogues	98.5	19.6	750	✓	✓	✓	✓	✓
GovReport (**Koh et al., 2022**)	Reports	3884.5	397.2	204	✓	✓	✓	✓	✓
SQuALITY (**Krishna et al., 2023**)	Stories	4795.7	376.9	60	-	-	-	-	-

Table 1: Datasets utilized in this work with statistical details: averaged document length, summary Length, number of all available summaries; and the unified error type taxonomy described in Section 3: "-" means no error types originally annotated; ✓ and ✗ represent whether the corresponding error type is available after label conversion.

summary pair, such as Falke et al. (2019) and SummaC (Laban et al., 2022). Within this direction, several works focus on improving the NLI model itself through methods such as synthetic data construction (Kryscinski et al., 2020; Yin et al., 2021; Utama et al., 2022; Qiu et al., 2024) or multitask learning (Zha et al., 2023a,b). The second direction employs QA-based models, such as QuestEval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2022), where they generate questions regarding explicit entities in the summary, then verify upon the source document. Besides the two main directions, other works have also explored to recognize factual errors through methods such as syntactic dependencies (Goyal and Durrett, 2020) or information extraction (Nan et al., 2021).

LLM Approaches The capability of LLMs on detecting inconsistencies have been studied by several recent works. Most of them resolve this task in the zero-shot or few-shot prompting manner (Shen et al., 2023; Luo et al., 2023; Wang et al., 2023; Liu et al., 2023a). Other utilization of LLMs have also been proposed, such as synthetic data generation with LLMs (Gekhman et al., 2023).

3 Task Taxonomy

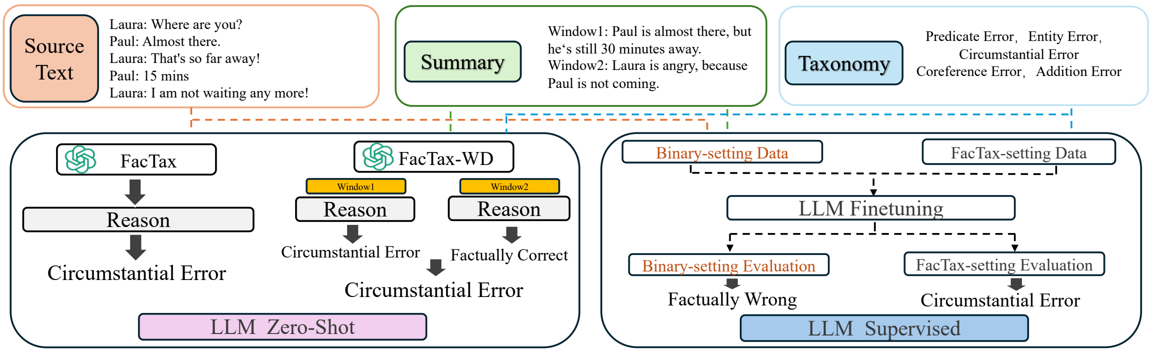
Establishing what constitutes inconsistent facts in a summary is a fundamental aspect of this task. In this work, we target to consolidate key error types that are salient for inconsistent fact detection in general, instead of building fine-grained complex taxonomy, for two reasons. First, a simple taxonomy is easier to be consumed by models than a complex one, aligning with our goal of practical utilization during inference. Second, a more fine-grained taxonomy may be of greater noises, as the annotated types from different datasets can vary significantly in their standards.

Based on the annotation schemas of previously introduced datasets and the aggregation of factual errors by AGGREFACT (Tang et al., 2023), we identify the following five salient error types:

- *Predicate Error*: the semantics expressed by a predicate in the summary are not consistent with those in the source document.
- *Entity Error*: any core arguments or attributes (e.g. subjects and objects in semantic frames) in the summary are not consistent accordingly.
- *Circumstantial Error*: Time, duration, or the location of an event or action is not consistent.
- *Coreference Error*: a pronoun or a reference mention in the summary cannot be resolved to refer to the correct entity.
- *Addition Error*: the summary expresses facts or events that have no grounding sentences in the document, thus cannot be verified (unless clearly extrapolatable by common sense).

These five error types focus on the "factuality" aspect that reflects semantic frames not aligned with the source document, which have been partially or entirely adopted in previous datasets. To unify labels across datasets by the above taxonomy, we conduct the following steps:

1. For datasets originally without error types annotated, no label mapping is performed.
2. For datasets addressed by AGGREFACT, we uti-



Laura: Where are you?
Paul: Almost there.
Laura: That's so far away!
Paul: 15 mins
Laura: I am not waiting any more!

Window1: Paul is almost there, but he's still 30 minutes away.
Window2: Laura is angry, because Paul is not coming.

Figure 1: Illustration of our proposed approaches that ground the task inference of factual inconsistency by its taxonomy (Sec. 3), via either the zero-shot paradigm (Sec. 4) or the supervised paradigm (Sec. 5) with LLMs.

lize their unified labels from AGGREFACT, then perform a heuristic conversion: NP → *Entity Error*; Pred → *Predicate Error*; Sent → *Addition Error*. For all extrinsic errors, we also mark them as *Addition Error*.

3. For datasets not included in AGGREFACT, we manually perform the label conversion per dataset (details provided in Appx. A).

Table 1 shows the resulting conversion as well as more statistical details. We do notice that neither our adopted taxonomy nor the fine-grained one in AGGREFACT is completely free from noises, due to different annotation standards across datasets.

4 Approach: Zero-Shot Paradigm

To incorporate the error type taxonomy, we first propose zero-shot prompting methods that leverage the general comprehension capability of LLMs, aiming to depict the explicit solution space to facilitate the zero-shot inference.

4.1 FACTAX

Our first designed prompting scheme, dubbed Factuality with Taxonomy (FACTAX), follows the standard zero-shot procedure: for a document-summary pair, we instruct a LLM to determine whether the summary is factually correct, as in previous works utilizing LLMs (Luo et al., 2023; Wang et al., 2023; Liu et al., 2023a). For each error type, we handcraft its explanation along with an optional example, and we ask the LLM to reason in a Chain-of-Thought (CoT) style (Wei et al., 2022): whether there are any specific error types present in the summary, instead of generating a binary decision directly. A summary is thus recognized as factually correct through a rationalization stage, when no specified error types are present.

The resulting zero-shot inference to this end, is regularized by the underlying task taxonomy, so to achieve a comprehensive task reasoning. We provide our full prompt in Appx. B.

4.2 FACTAX-WD

Since summaries often extend beyond a single sentence, prior works adapted NLI models such as SummaC (Laban et al., 2022) conduct inference on each sentence independently, which helps mitigate degradation that may occur when inferring over long summaries. As LLMs are susceptible to degradation over long sequences as well (Hsieh et al., 2024), certain errors scattered across many sentences may be overlooked by the model. Thus, we further introduce a second prompting scheme intuitively: rather than processing the entire summary at once, we divide it into separate windows that are individually processed. The final result is then aggregated across windows, such that a summary is factually correct only if each window possesses no errors. The second method is thereby termed FACTAX by Windows (FACTAX-WD).

After our preliminary experiments, we manually set the window size as roughly 30 words to balance between the performance and efficiency. It is worth noting that smaller window size (i.e. one sentence per window) does not necessarily lead to higher performance, as we observe that pronouns in the summary could often introduce false negatives when inferred without its broader context.

4.3 Zero-Shot Experiments

Datasets For comprehensive evaluation, we adopt diverse document types of five domains: CNN/DM, XSum, dialogues, reports, and stories. Each domain consists of one or multiple datasets from Table 1, with 10 datasets evaluated by the zero-shot paradigm in total. Notably, we only evaluate

	CNN/DM					XSum					Dialogues	Reports	Stories	
	Polytope	**SummEval**	**Frank**	**CLIFF**	Avg.	**Wang'20**	**CLIFF**	**Goyal'21**	**Cao'22**	Avg.	**DiaSumFact**	**GovReport**	**SQuALITY**	MACRO
QuestEval	17.60	64.90	62.60	**74.00**	70.20	56.00	61.90	**81.40**	60.10	69.50	57.03	26.90	42.11	51.15
QAFactEval	32.40	65.20	54.70	71.60	67.80	**75.60**	62.60	75.40	61.30	65.85	65.91	40.59	44.79	56.60
SUMMAC-ZS	97.10	62.20	57.00	65.60	64.00	69.80	59.60	46.60	49.00	56.40	58.81	35.19	15.00	45.88
AlignScore	94.12	43.40	53.65	67.61	64.04	65.52	74.68	52.63	65.70	67.59	68.93	37.07	43.77	56.28
Align	91.18	44.92	55.48	58.30	69.49	68.09	**74.82**	68.06	65.34	68.41	**69.22**	35.05	46.26	57.47
FalseSum	-	-	-	-	50.50	-	-	-	-	54.70	-	-	-	-
AMRFact	**100.00**	**80.70**	**72.40**	71.00	**72.30**	59.50	66.70	59.10	64.50	64.10	-	-	-	-
ChatGPT-ZS	90.19	79.78	54.82	65.13	60.03	71.82	74.01	63.38	68.82	69.39	66.85	41.40	44.63	56.46
ChatGPT-CoT	89.22	66.64	51.94	62.20	56.20	68.30	66.27	63.85	65.98	66.21	61.59	40.73	42.64	53.47
ChatGPT-Star	41.17	54.57	51.27	57.91	55.30	56.72	56.61	65.25	54.89	55.89	62.86	35.90	25.62	47.11
G-Eval	99.02	48.98	54.18	56.25	55.04	51.05	56.61	53.08	52.36	51.57	51.73	15.77	35.86	41.99
FacTax	78.44	67.43	62.82	68.71	68.97	74.06	70.25	74.08	**71.65**	**72.21**	62.76	40.54	45.93	58.08
FacTax-WD	85.31	72.98	67.09	70.71	68.92	71.46	70.81	68.85	69.49	69.82	64.15	**48.36**	**48.06**	**59.94**

Table 2: Evaluation results for the zero-shot paradigm (Section 4.3). Five domains (10 datasets in total) are evaluated, where the setting for CNN/DM and XSum is kept consistent and comparable with AggreFact-FtSota (Tang et al., 2023) using thresholds per dataset. MACRO is the final evaluation metric that computes the macro-average score across each domain. FacTax methods are our proposed approaches that ground the zero-shot inference by the task taxonomy. All LLM-based methods are shown the averaged scores of three repeated runs for robust evaluation, and they are directly comparable due to adopting the same underlying model (*gpt-3.5-turbo-0125*).

upon summaries generated by state-of-the-art models specified by each dataset, in coordination with AggreFact-FtSota (Tang et al., 2023).

For GovReport and SQuALITY, documents are long articles that exceed certain models' length limit. We follow Wu et al. (2023) that for each document, top sentences that maximize ROUGE scores towards the summary are retrieved as a condensed context (details in Appx. C).

Metrics As in previous works, we use Balanced Accuracy for all datasets that offer classification labels. For GovReport and SQuALITY whose labels are consistency scores, we use Pearson Correlation that aligns with prior works. The performance of each domain is either from the standalone dataset (e.g. Dialogues), or represented by the micro average score of all datasets within this domain, aligning with AggreFact evaluation. We further introduce a single metric to evaluate the overall performance, termed MACRO, which takes the macro average scores across each domain.

Non-LLM Methods We adopt strong non-LLM models as baselines, including QuestEval (Scialom et al., 2021), QAFactEval (Fabbri et al., 2022), SummaC (Laban et al., 2022), AlignScore (Zha et al., 2023a) and Align (Zha et al., 2023b). For each, we either take the evaluation scores from prior works, or run the code released by the authors on new datasets not evaluated previously. Additionally, we also include scores of FalseSum and AMRFact from their original papers. More details on non-LLM baselines are provided in Appx. C.

LLM Methods We use ChatGPT-ZS, ChatGPT-CoT (Luo et al., 2023), ChatGPT-Star (Wang et al., 2023), and G-Eval (Liu et al., 2023a) as the LLM baselines, which have achieved strong performance in AggreFact. For direct comparison, we use the same ChatGPT (*gpt-3.5-turbo-0125*) for all prompts. In later analysis (Sec. 4.4), we vary LLMs for more insights on model comparison.

For FacTax, we adapt the prompts to additionally yield a score on GovReport and SQuALITY summaries to enable evaluation with their score-based labels.

Results Due to the variation from sampled generation of LLMs, we run all LLM-based methods three times for robust conclusions, and show the averaged scores of each dataset in Table 2, along with evaluation results of non-LLM baselines. Several observations from Table 2 can be made as follows.

• Corroborating previous works, **LLM zero-shot inference is capable to identify factual errors directly** with decent performance, matching or exceeding strong non-LLM baselines specialized for factual inconsistency detection. Specifically, FacTax methods and ChatGPT-ZS achieve 56.5 - 59.9 MACRO scores, on par with 56.3 - 57.5 obtained by QAFactEval, AlignScore and Align. The lowest score of LLM baselines is 41.2, which only lags behind SUMMAC-ZS by 3.9%.

• Comparing among LLM-based approaches, FacTax-WD achieves the best overall performance, surpassing the best LLM baseline ChatGPT-ZS by 3.5%, also outperforming all non-LLM baselines. As the main difference between FacTax and LLM baselines is the incorporation of given task taxonomy in the prompt, the empirical result suggest that **LLM inference can indeed benefit from a grounded solution space**.

	CNN/DM	XSum	Dialogues	MACRO
G-Eval (GPT-4)	**69.90**	65.80	-	-
ChatGPT	68.97	**72.21**	62.76	67.98
GPT-4o	68.56	71.20	73.08	70.93
Llama3-8B	50.38	61.17	62.81	58.12
Llama3-70B	69.6	71.05	**74.63**	**71.76**
Mistral-7B	51.95	57.8	58.31	56.02

Table 3: Evaluation results using FacTax on three domains by varying LLMs. FacTax with ChatGPT achieves SOTA performance on AggreFact-FtSota benchmark, surpassing previous zero-shot SOTA G-Eval with GPT-4 (from Qiu et al. (2024)).

• The gap between FacTax and FacTax-WD is relatively trivial. The **window-based inference is shown effective on long summaries**, demonstrated by the significant performance raise on GovReport and SQuALITY.

4.4 Zero-Shot Analysis

We focus on three common domains with shorter summaries: CNN/DM, XSum, Dialogues, and perform further analysis for more regarding insights.

Impact of LLMs and Sizes Apart from ChatGPT, we also employ GPT-4o from OpenAI, as well as strong open-source LLMs including Llama3-8B/70B and Mistral 8B. Table 3 provides the model comparison by adopting the same FacTax prompts. Notably, FacTax with ChatGPT achieves SOTA performance on AggreFact-FtSota benchmark (Tang et al., 2023), outperforming G-Eval with GPT-4 (Qiu et al., 2024).

Just by switching to GPT-4o, there comes a direct boost upon ChatGPT by 3 MACRO score overall. There is still quite a gap of almost 10% between the smaller 7B/8B models and the larger OpenAI models. By Table 3, it is evident that increasing the model size significantly improves the task reasoning, as switching Llama3-8B to 70B obtains performance gain by an impressive 13.6%.

Impact of Examples Our default FacTax setting grounds the inference by depicting error type definitions in the prompt, without supplying any examples. After conducting multiple rounds of experiments by adding crafted examples per type, we are not able to obtain stable improvement, as adding a few examples could lead to biases towards errors. The averaged improvement after adding examples is only 0.06 MACRO score; thus, we keep FacTax off examples in this work.

Impact of Summary Lengths Figure 2 plots the performance curve on different lengths of summaries using ChatGPT. FacTax-WD is shown more robust against long summaries, due to its length-agnostic scoring mechanism. For long summaries, false positives become more often for FacTax, which is alleviated by the window-based inference of FacTax-WD.

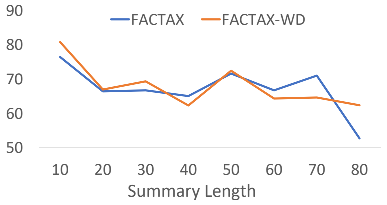


Figure 2: Accuracy of FacTax methods for different summary lengths using ChatGPT.

5 Approach: Supervised Paradigm

As Section 4 has demonstrated the strengths of grounding zero-shot LLM inference by its task taxonomy, we further seek to distill a model that absorbs the taxonomy into LLM parameters through supervised training. Two advantages could come with such distillation. First, by learning the taxonomy from examples, the model gains real-world distribution of each error type, rather than relying on shallow comprehension through prompt instructions. Second, it is more efficient and practical compared to zero-shot methods, avoiding the need for large model sizes and lengthy generation due to zero-shot CoT reasoning.

With above motivations, we utilize previous datasets that were proposed independently, and regard them jointly as training resources. To this end, we prepare our training and test set as follows:

- **Training Set I**: FRANK, Polytope, BUMP, CLIFF, Goyal’21, DeFacto, XSumFaith, DiaSummEval, DiaSummFactCorr, FactEval
- **Training Set II**: DocNLI, FalseSum
- **Test Set**: SummEval, Wang’20, Cao’22, DiaSumFact

Concretely, the test set is formed to cover at least one dataset per CNN/DM, XSum, and Dialogues domain. Training Set I encompasses datasets with human-annotated labels, and we use all available examples of each dataset in training, while keeping the test set only containing summaries from state-of-the-art models, such that the evaluation of the supervised paradigm is directly comparable with zero-shot results in Table 2&3.

		CNN/DM	XSum		Dialogues	
		SummEval	**Wang'20**	**Cao'22**	**DiaSumFact**	MACRO
Zero-Shot	ChatGPT	73.0	71.5	69.5	64.2	69.2
	GPT-4o	77.0	**76.4**	70.3	72.9	74.4
Zero-Shot	Llama3-8B	64.4	59.1	61.1	62.8	62.4
	Llama3-70B	79.5	74.1	70.1	74.6	**75.4**
Supervised	`I-Binary + INF-Binary`	80.1	62.7	**72.5**	72.6	73.4
	`I-Taxonomy + INF-Binary`	80.3	63.0	67.4	76.1	73.9
	`I&II-Taxonomy + INF-Binary`	79.0	68.0	70.9	**76.7**	75.1
	`I&II-Taxonomy + INF-Taxonomy`	**81.1**	67.7	71.8	75.2	**75.4**

Table 4: Evaluation results of the supervised paradigm, which are directly comparable with zero-shot results in Table 2. Llama3-8B models are trained through supervised finetuning by three training settings (Sec. 5.1). For the `I&II-Taxonomy` setting, we apply both binary inference and error type inference for the best performance.

Training Set II includes two publicly released large-scale datasets constructed via synthetic data generation. The more recent AMRFACT is excluded, since its data has not been released as of this writing. For efficiency, Training Set II retains randomly sampled 50k examples from DocNLI and FalseSum respectively that do not overlap with any source documents in the test set.

The resulting training resources thereby have 16k examples in Training Set I, 100k examples in Training Set II, and 1k examples in Test Set.

5.1 Training Strategy

With our identified task taxonomy, we unify the error types for all training examples whenever applicable according to Table 1. The training is conducted through LLM supervised finetuning, where each example is converted into pairs of prompts and completions. To fully utilize the available resources, we design two types of prompt-completion pairs, according to if error type labels are available:

- *Error Type Completion*: if a dataset has error type labels available after the label conversion, the prompt then lists error type candidates, instructing the model to generate specific error types if present any in the summary.
- *Binary Completion*: for an example, the prompt can ask to directly classify if the summary is factually correct. The completion is then a binary label.

Note that for those examples with error type labels, both two types of completions can be created, which inflates training size, and also anchors different error types towards "factually wrong".

5.2 Supervised Experiments

Training Settings We employ Llama3-8B as the backbone LLM model for supervised training. Three training settings are experimented, based on different training sets and completion types:

- `I-Binary`: Training Set I that only adopts *Binary Completion* for all examples.
- `I-Taxonomy`: Training Set I with both *Binary* and *Error Type Completion* when applicable.
- `I&II-Taxonomy`: adding Training Set II (only *Binary Completion* is applicable), in addition to all prompt-completion pairs in `I-Taxonomy`.

Particularly, the performance difference between `I-Binary` and `I-Taxonomy` could directly reflect the impact of incorporating the task taxonomy into model parameters. `I&II-Taxonomy` further explores the extent to which synthetic data can complement human annotations.

For the latter two settings, the INFerence of trained models is also flexible, which could either opt to determine the factual consistency directly (`INF-Binary`), or to yield fine-grained error types (`INF-Taxonomy`), according to specific types of prompts given. For `I&II-Taxonomy`, we evaluate both inference for the best performance.

In our experiments, we adopt common hyperparameters for LLM finetuning, described in Appx. D, without requiring a development set due to limited resources. Detailed statistics of three settings are shown in Table 5. Specifically for `I&II-Taxonomy`, we boost the ratio of *Error Type Completion* to 20% in training by adjusting the data sampling strategy, to facilitate model learning of the task taxonomy.

Results Table 4 shows the evaluation results of our supervised paradigm, along with comparison by various zero-shot results. Unsurprisingly, trained Llama3-8B models of any settings outperform its zero-shot inference by large margins, up to 13 MACRO score. More impor-

	# Train	Length	T-Ratio	# Test
I-Binary	16393	648.5	0%	1033
I-Taxonomy	26315	634.2	37.7%	1033
I&II-Taxonomy	116393	783.3	20.0%	1033

Table 5: Statistics of three supervised training settings: number of prompt-completion pairs in training; averaged length of prompts; ratio of prompts with Taxonomy provided; number of prompts for evaluation.

	Ent.	Pred.	Circ.	Coref.	AddE.
ChatGPT	41.1	28.8	**28.2**	21.9	45.1
GPT-4o	60.4	34.4	18.2	10.0	**46.5**
Llama3-8B	32.7	34.2	23.1	10.4	35.8
Llama3-70B	58.5	32.6	27.8	3.4	43.0
I-Taxonomy	61.1	**37.1**	25.3	**28.9**	21.9
I&II-Taxonomy	**62.6**	29.0	23.0	26.3	31.0

Table 6: F1 of five error types on DiaSumFact evaluation, with both zero-shot and supervised paradigm.

	(i)	(ii)	(iii)	(iv)
ChatGPT	9.7	33.4	63.4	28.5
GPT-4o	9.8	35.6	76.3	37.6
Llama3-8B	4.6	27.8	62.0	24.0
Llama3-70B	9.0	36.5	73.5	36.0
I-Taxonomy	15.5	31.8	**85.4**	40.4
I&II-Taxonomy	**15.8**	**39.4**	83.8	**41.1**

Table 7: Percentage of correct error type predictions on DiaSumFact by four different criteria: i) exact match by gold error types; ii) predicted types are a subset of gold types; iii) predicted types contain one of gold types; iv) predicted types contain all gold types.

tantly, I&II-Taxonomy + INF-Taxonomy achieves the best performance, matching the best existing approaches by using FACTAX with GPT-4o and Llama3-70B. Our trained model can effectively serve as **an efficient alternative to zero-shot inference by much larger LLMs**, which we will publicly release for the research community.

Comparing I-Binary and I-Taxonomy, there is an enhancement of 0.5 MACRO score by adopting *Error Type Completion* in training; indeed, utilizing large-scale synthetic data brings more improvement by 1.2 MACRO score. By reasoning via error types rather than binary decisions, I&II-Taxonomy receives further 0.3 gain, validating the benefit of fusing task taxonomy into model parameters.

5.3 Supervised Analysis

Fine-Grained Evaluation Table 6 shows the F1 score of each error type with zero-shot and supervised paradigms. Among five types, most methods suffer on *Circumstantial Error* and *Coreference Error*, while performing the best on *Entity Error*. The two trained models surpass zero-shot methods on three error types. However, they perform worse on *Addition Error*. We attribute the degradation to different annotation standards across datasets, which may become noisy even after label unification. Nevertheless, as we have already seen improvement with the current taxonomy, future works with cleaner labels have good potentials to further boost the supervised performance.

Error Type Predictions As models often predict partially correct error types, Table 7 shows the percentage of correct type predictions by four criteria, from strict to relaxed. As the results suggest, either zero-shot or supervised methods could recognize at least one gold error type on most of the factually incorrect cases, by up to 85.4% achieved by the trained model I-Taxonomy. Whereas for exact match, even the trained models could only obtain 15% accuracy. The best performance by either criterion is achieved by the supervised paradigm, as expected, since the model learns the real-world distribution from training examples.

6 Conclusion

We highlight the importance of task-specific taxonomy for factual inconsistency detection, where we consolidate salient error types, and incorporate them to facilitate LLM inference with both zero-shot and supervised paradigms. Extensive experiments on ten datasets of five domains demonstrate the efficacy of depicting task taxonomy to ground the zero-shot inference, achieving state-of-the-art performance compared with respective baselines. We further distill models that fuse the given error taxonomy into parameters through our designed training completions and strategies, effectively serving as an efficient alternative to state-of-the-art zero-shot reasoning by much larger LLMs.

7 Ethical Statements

This work does not pose direct ethical concerns, as we utilize existing datasets on a well-studied task. However, we do emphasize that factual predictions by our proposed approaches should not be taken without verification. We provide failed examples by zero-shot LLM inference in Appx. F.

Limitations

While our study demonstrates the effective utilization of task-specific taxonomy for detecting factual inconsistencies, it is important to acknowledge certain limitations.

First, as discussed in Section 5.3, the unified labels after conversion can contain noises, due to the different annotation standards across previous independently introduced datasets. The resulting converted error type labels may hinder the supervised training process. Further consolidation may be conducted for a cleaner realization of the error type taxonomy.

Second, both the zero-shot paradigm and supervised paradigm may not fully capture the nuances of complex summaries. We list concrete qualitative examples in Appendix F on the failed cases by LLMs. Specifically for zero-shot paradigm, the failed cases could come from imperfect instruction following, as well as ambiguous descriptions of the task taxonomy that are not fully comprehensive. For the supervised paradigm, it indeed requires either human annotated examples, or synthetic data generation, which may not generalize as well as the zero-shot inference.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. 2022. Toward unifying text segmentation and long document summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 106–118, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang, and Baoxing Huai. 2023. Reference matters: Benchmarking factual error correction for dialogue summarization with fine-grained evaluation framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13932–13959, Toronto, Canada. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models?

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023b. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Liang Ma, Shuyang Cao, Robert L Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. BUMP: A benchmark of unfaithful minimal pairs for meta-evaluation of faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12788–12812, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020b. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. 2024. Amrfact: Enhancing summarization factuality evaluation with amr-driven negative samples generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022a. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022b. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2023. Less is more for long document summary evaluation by llms. *arXiv preprint arXiv:2309.07382*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023a. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023b. Text alignment is an efficient unified model for massive NLP tasks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6825–6845, Toronto, Canada. Association for Computational Linguistics.

A Taxonomy Conversion

For datasets not included in AggreFact, we manually perform the error type conversion as follows:

- BUMP: Authors of original dataset manually edit reference summaries to constructs an unfaithful summary and classified error type into Extrinsic Entity Error, Intrinsic Entity Error, Intrinsic Predicate Error, Extrinsic Circumstance Error, Intrinsic Circumstance Error, Coreference Error and Other. We mapped Extrinsic Entity Error and Intrinsic Entity Error to Entity Error; Extrinsic Predicate Error and Intrinsic Predicate Error to Predicate Error; Extrinsic Circumstance Error and Intrinsic Circumstance Error to Circumstantial Error; Coreference Error to Coreference Error and Extrinsic-related Error to Addition Error. We also manually mapped five edited summaries with other types of errors to the types we have set according to our own judgment.
- DiaSumFact: Authors of original dataset classified error types into Ex-EntE, In-EntE, Ex-PredE, In-PredE, Ex-CirE, In-CirE, CorefE, LinkE ande Others. We mapped Ex-EntE and In-EntE to Entity Error; Ex-PredE, In-PredE and LinkE to Predicate Error; Ex-CirE and In-CirE to Circumstantial Error; CorefE to Coreference Error; Ex-Error to Addition Error and manually mapped Others base on the comment given by annotators.
- DiaSummFactCorr: The error types of summmaries in this dataset were classified into EntE, PredE, CircE, CorefE, LinkE, GramE, OutE and OthE. We mapped EntE to Entity Error; PredE, GramE and LinkE to Predicate Error; CircE to Circumstantial Error; CorefE to Coreference Error;OutE to Addition Error and mapped each summary with OthE manually according to our own judgment.
- FacEval: Authors of original dataset classified error types into Subject Object Error, Pronoun Error, Negation Error, Particulars Error, Hallucination Error and Other Error. We mapped Subject Object Error to Entity Error; Pronoun Error to Coreference Error; Negation Error to Predicate Error, Particulars Error to Circumstantial Error; Hallucination Error to Addition Error and mapped each summary with Other Error manually according to our own judgment.
- GovReport: Authors of original dataset classified each summary sentence's factuality based on seven types of errors: PredE, EntityE, CircE, CorefE, LinkE, OutE and GramE. We mapped EntityE to Entity Error; PredE, GramE and LinkE to Predicate Error; CircE to Circumstantial Error; CorefE to Coreference Error and OutE to Addition Error.

B Full Prompts

We provide the full prompt for FacTax in Figure 3.

C Zero-Shot Experimental Settings

Long Document Alignment As documents in both GovReport and SQuALITY have long length of thousands of tokens, alignment is firstly performed, such that for each summary or summary window, related sentences from the document are retrieved, which will be used as a shorter context for factual error evaluation. Though past work has proposed techniques for long context segmentation (Cho et al., 2022), in this work, we opt for the common approach via retrieval for simplicity.

For FacTax, top sentences from the document that maximize the recall of ROUGE-1 and ROUGE-2 towards the summary are retrieved until the total length reaches a certain threshold. These sentences are concatenated as the new context, which is shorter but has a higher information density than the original document.

For FacTax-WD that operates on summary windows, n important sentences are independently extracted to maximize the recall of ROUGE-1 and ROUGE-2 towards the summary. Table 8 shows the alignment thresholds we adopted for the two datasets.

	Summ-Alignment	Window-Alignment (n=5)
GovReport	1024	102.31
SQuALITY	1024	28.50

Table 8: The maximum length of aligned context for FacTax, and the averaged length of aligned context per summary window for FacTax-WD. n is the number of sentences extracted for each summary window. For SQuALITY, some of the retrieved sentences can be quite short.

Evaluation for Baselines Five non-LLM baselines, QuestEval, QaFactEval, SummaC-ZS, AlignScore and Align produce a consistency score for each summary, which requires a threshold to convert to the classification label. For each dataset and origin split in AggreFact-FtSota test set, following Tang et al. (2023), we tune the threshold to reach the best balanced accuracy on the corresponding part of the dataset or origin split in AGGREFACT-FTSOTA validation set.

For DiaSumFact, we use validation split of AGGREFACT-FTSOTA dataset to tune the threshold for the baselines For GovReport and SQuALITY, the evaluation metric is Pearson Correlation, thus not requiring any thresholds.

D Supervised Experimental Settings

We perform full finetuning of Llama3-8B with flash attention enabled on 8 Nvidia A100 GPUs (40GB memory each). Each training setting runs for 8 epochs, taking around 6 hours to finish `I-Binary` and `I-Taxonomy`, and 24 hours to finish `I&II-Taxonomy`. We use a batch size of 1 per GPU, learning rate of 1×10^{-5}, and cosine learning rate schedule with warm up rate 0.05. Loss is only enabled on the completion during training.

E AI Assisted Writing

ChatGPT from OpenAI is used in this writing for revising certain expressions.

F Failure Cases by LLMs

Failed to grasp the implied information For some cases, LLM failed to grasp the implicit content in the text, leading to an error in judgment.

[Part of the source text]
...
Ana: You sleeping?
Catherine: Not yet.
Ana: Wanna go visit grandma tomorrow? I miss her.
*Catherine: **Yeah that would be nice** :) I'll call you when I wake up*
...

[Part of the summary]
Ana and Catherine are going to see their grandma tomorrow.

[Human comment]
None

[Error type]
Factually Correct

[Response of LLM]
*- Reasoning: The statement is not consistent with the source document as it is **only Ana** who expresses the desire to visit grandma, not Catherine.*
- Label: Entity_Error

Inconsistent reasoning with humans For some cases, even though LLM makes the correct label prediction, its reasoning does not correspond to what humans think.

[Part of the source text]
...
phd d: Uh working in the cepstral domain ? So there is one guy in Grenada
phd b: Yeah, in Grenada one of my friend.
phd d: and another in uh, Lucent that I met at ICASSP.
...

[Part of the summary]
The team said that the person in grenada was working in the cepstral domain.

[Human comment]
the team -> phd d

[Error type]
Entity Error

[Response of LLM]
*- Reasoning: The summary states that **the person in Grenada was working in the cepstral domain.** However, there is no mention in the source text about the person in Grenada working in the cepstral domain. This indicates a predicate inconsistency.*
- Label: Predicate_Error

```
### Background

Now you are one of the greatest English linguist experts and Logician. Below is an
instruction for evaluating the factual consistency of a given statement derived from the
source document. Consistency measures whether the statement is correctly aligned with the
facts presented in the source document, without contradicting or misrepresenting any
information. The goal is to determine the consistency label of the statement defined as below
, identifying whether the statement contains factual errors, or is factually correct towards
the source document.


### Consistency Labels

i) Predicate_Error: a predicate from the statement (usually a verb of a semantic frame) is
not consistent with the source document.
ii) Entity_Error: a core argument or attribute in a semantic frame (usually a subject or
object) from the statement is not consistent with the source document (capitalization should
not matter).
iii) Circumstantial_Error: Time, duration, or the location of an event from the statement is
not consistent with the source document.
iv) Coreference_Error: a pronoun or a reference (e.g., this picture) from the statement is
wrong, or cannot be resolved to refer to the correct entity.
v) Addition_Error: the statement introduces facts that cannot be verified from the source
document.
vi) Factually_Correct: the statement is factually correct without above factual errors. Note
that it is allowed for the statement to miss important information from the source document;
it is considered factually correct as long as the statement can be verified from the source
document.

### Task

Please try your best to firstly reason in few sentences on whether the statement has factual
errors or is factually correct, then determine the consistency label(s) from the above 6
types of labels in the end.


--- Your_Task ---

### Source Document

{Document}

### Statement

{Entire Summary or Summary Window}

### Output Format

- Reasoning: ...
- Label: ...

### Your Output

Please refer to the above instruction, return Reasoning and Label in a Markdown list, to
evaluate the factual consistency of the statement.
```

Figure 3: Prompt for FacTax described in Section 4. Slots in blue refer to the input document and summary.