Attribute or Abstain: Large Language Models as Long Document Assistants

Anonymous ACL submission

Abstract

LLMs can help humans working with long documents, but are known to hallucinate. *Attribution* can increase trust in LLM responses: The LLM provides evidence that supports its response, which enhances verifiability. Existing approaches to attribution have only been evaluated in RAG settings, where the initial retrieval confounds LLM performance. This is crucially different from the long document setting, where retrieval is not needed, but could help. Thus, a long document specific evaluation of attribution is missing. To fill this gap, we present LAB, a benchmark of 6 diverse long document tasks with attribution, and experiment with different approaches to attribution on 4 LLMs of different sizes, both prompted and fine-tuned. We find that `citation`, i.e. response generation and evidence extraction in one step, mostly performs best. We investigate whether the "Lost in the Middle" phenomenon exists for attribution, but do not find this. We also find that evidence quality can predict response quality on datasets with simple responses, but not so for complex responses, as models struggle with providing evidence for complex claims. We release code and data for further investigation[1].

1 Introduction

Recent LLMs can process long documents (Shaham et al., 2023; Li et al., 2023b), showing great potential as *long document assistants*. For example (Fig. 1), such an assistant could answer a researcher's questions about a paper. However, due to LLM hallucinations (Slobodkin et al., 2023), the researcher must verify responses, which is difficult with lengthy papers. To improve verifiability and trust, the assistant should either *attribute* (Rashkin et al., 2023) or *abstain* (Slobodkin et al., 2023): If it finds the necessary information, it should provide a response and point to the evidence in the paper

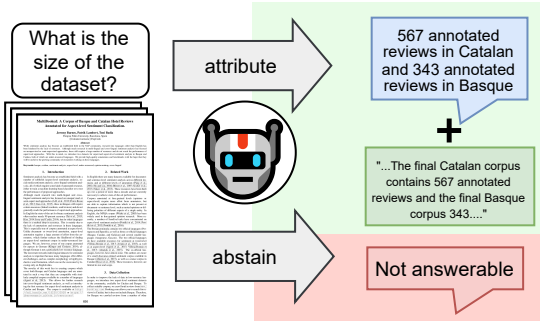

Figure 1: Long document assistants should *attribute*, i.e. provide responses with evidence, or abstain. Example from QASPER (Dasigi et al., 2021). Figure requires emojis to display correctly.

(attribute). If not, it should clearly communicate this (abstain). We investigate the capabilities of LLMs to fulfill these requirements, and the relation between response quality and evidence quality.

Formally, we assume an instruction I (e.g. a question) and a document D consisting of segments d (e.g. paragraphs)[2]. There are two subtasks that can be solved jointly or independently: One is to generate a *response* R (e.g. an answer) containing statements r. If the response is not abstained (e.g. by saying "unanswerable"), the other subtask is to retrieve *evidence* $E_i \subset D$ for each r_i such that r_i, is *attributable* to E_i, i.e. "according to E_i, r_i" is true (Rashkin et al., 2023).

Different approaches to attribution can be defined based on the subtask order (Fig. 2): (1) `post-hoc`: an LLM generates a response R, and evidence is retrieved from D based on R. (2) `retrieve-then-read`: Evidence E is retrieved from D, and an LLM generates a response based on E. (3) `citation`: Based on D, an LLM generates a response and retrieves evidence in one step. To decrease the input length in `post-hoc` and `citation`, D can be reduced to $D' \subset D$ via and additional initial retrieval step. This re-

[1] Anonymous link, code under Apache 2.0, dataset licenses depend on original license, see §B.

[2] In other scenarios, D can be a large corpus, see below.

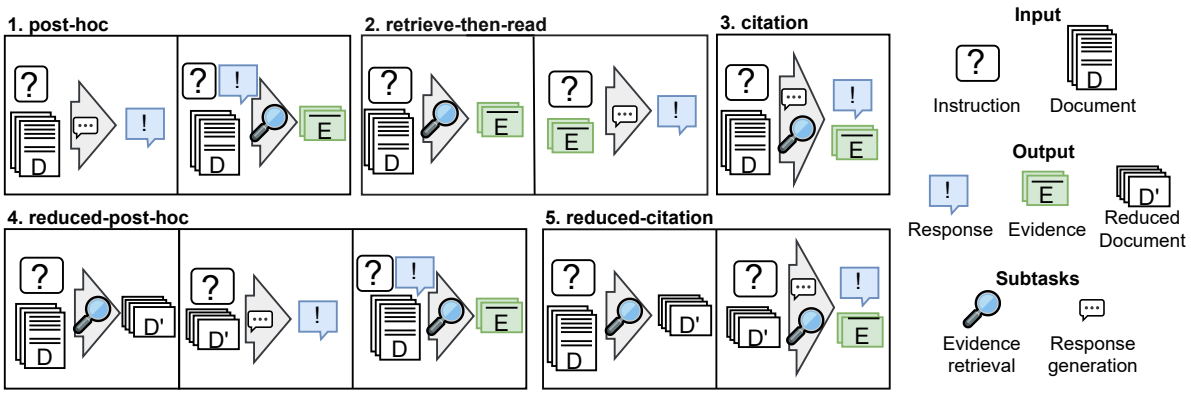

Figure 2: The approaches to attribution in long document scenarios analyzed in this work.

sults in two further attribution approaches: (4) `reduced-post-hoc` and (5) `reduced-citation`.

Attribution has only been investigated in retrieval augmented generation (RAG) settings, where D is not a document, but a large corpus (e.g. Wikipedia) that does not fit the LLM context. This means that only approaches with initial retrieval, `retrieve-then-read`, `reduced-post-hoc` and `reduced-citation`, can be used. Performance depends on retrieval quality, and the best-performing approach is unclear (Bohnet et al., 2022; Gao et al., 2023b; Malaviya et al., 2024).

In contrast, when D is a long document that fits the LLM context, the confounding initial retrieval can be omitted. Still, it is possible that separating generation and retrieval improves performance, as shown in related works on task decomposition (Sun et al., 2023) and reducing long LLM inputs, where Agrawal et al. (2024) found positive effects, while Xu et al. (2024) and Bai et al. (2023) did not. These works did not consider attribution, so there is a lack of knowledge on the effect of separating response generation and evidence retrieval, and the optimal approach to attribution on long documents.

The document length poses additional challenges: Recent works have found that LLM performance on long-input tasks depends on the position of the information in the context (Liu et al., 2024; Staniszewski et al., 2024; Ravaut et al., 2024). Whether this can also be observed for attribution has not yet been investigated.

Evidence quality (attributability) can be evaluated without external reference (Honovich et al., 2022; Yue et al., 2023; Tang et al., 2024). If evidence quality were positively correlated with response quality, bad responses could be abstained from by filtering responses with bad evidence quality. If not, this would lead to abstaining from potentially helpful responses with insufficient evidence. Current research on the relation of response quality and evidence quality is inconclusive: Bohnet et al. (2022) and Gao et al. (2023b) reranked multiple sampled responses by evidence quality. While the former found an improvement in response quality, the latter did not. Neither provide an analysis, so we lack understanding if and how evidence quality correlates with response quality.

To close these gaps, we compile LAB, a **L**ong-document **A**ttribution **B**enchmark of 6 long-document datasets with diverse tasks (QA, classification, NLI and summarization) and domains (science, law, governmental and Wikipedia). We conduct experiments using the outlined attribution approaches on 4 LLMs of varying sizes, prompted and fine-tuned, to answer three research questions:

RQ1: What are optimal approaches for attribution in long document tasks? We find that large and fine-tuned models reach best evidence quality via `citation`, but small prompted LLMs can benefit from `post-hoc` evidence retrieval.

RQ2: Do LLMs exhibit positional biases in evidence attribution? Concerning evidence retrieval, except for GovReport, we find no particular bias, as the predicted and gold evidence distributions are mostly similar. However, we find that response quality generally decreases as evidence appears later in the document.

RQ3: What is the relation between evidence quality and response quality? We find that evidence quality can predict response quality on datasets with single-fact responses, but not so for multi-fact responses, as models struggle with providing evidence for complex claims.

2 Related Work

Attribution Current research in attribution is done in three strains: First, some works evaluate `post-hoc`, `retrieve-then-read` and `citation` approaches in their ability to produce attributed responses (Bohnet et al., 2022; Liu et al., 2023), some proposing new datasets (Malaviya et al., 2024) or benchmarks (DeYoung et al., 2020a; Gao et al., 2023b). Second, methodological works propose new fine-tuning (Schimanski et al., 2024; Huang et al., 2024) and prompting (Berchansky et al., 2024; Fierro et al., 2024) methods to improve the `citation` capabilities of language models, but do not compare to `retrieve-then-read` or `post-hoc` approaches. Both of these strains have focused on open domain QA, and neglected the long document scenario. Here, we close these gaps by providing a comprehensive investigation of attribution for long documents.
To evaluate attributability automatically, most works have used TRUE, Flan-T5-XXL fine-tuned on several NLI datasets (Honovich et al., 2022; Bohnet et al., 2022; Gao et al., 2023b; Fierro et al., 2024; Huang et al., 2024). More recently, Attrscore (Yue et al., 2023) and Minicheck (Tang et al., 2024) were proposed specifically for the evaluation of attributability. We compare these models and employ the best-performing for attributability evaluation.

LLMs for long documents While there is no universal definition of "long documents", existing long document benchmarks contain documents of 1500 to 50000 words average length (Shaham et al., 2023; Dong et al., 2024; An et al., 2023; Li et al., 2023b). Initial LLMs were limited to contexts of less than 2000 tokens (Brown et al., 2020; Touvron et al., 2023), but recent advances in hardware and efficiency (Dao, 2023) have spurred the development of models with context $\geq$8000 tokens, e.g. Longchat (Li et al., 2023a), Mistral (Jiang et al., 2023), GPT-3.5-16K [3] or GPT-4-Turbo-128K.[4] We add to this line of research by evaluating a range of models in their attribution capabilities.

3 Methods

3.1 Datasets

The datasets in LAB are shown in Table 1. All datasets are in English. GovReport (Huang et al., 2021) is the only dataset without annotated evidence. To simulate gold evidence, we used BM25 to find the 2 best-matching paragraphs from a document for each sentence in the gold summary similar to Ravaut et al. (2024). Due to limited resources, we use at most 2000 test instances from any dataset (100 for GPT-4). For details and examples see §B.

3.2 Evaluation

Response parsing For all datasets except GovReport, responses consist of single statements. For GovReport, we split responses into statements (sentences) using NLTK (Bird, 2006). For `citation`, LLMs are expected to generate segment identifiers ("[1] [2]") at the end of each statement.

Evaluation of response quality For comparability with related work, we used established metrics to evaluate response quality: Exact match F1 (EM)[5] on QASPER (Dasigi et al., 2021; Shaham et al., 2023) and Natural Questions (Kwiatkowski et al., 2019; Bohnet et al., 2022), classification macro F1 (CF1) for Evidence Inference (DeYoung et al., 2020b), Wice (Kamoi et al., 2023) and ContractNLI (Koreeda and Manning, 2021) and ROUGE-L[6] (RL, Lin 2004) for GovReport (Huang et al., 2021; Shaham et al., 2023).

Evidence F1 (EF1) For datasets that define a *fixed* response vocabulary (i.e. Evidence Inference, Wice and ContractNLI), we compute evidence quality as evidence F1, comparing the predicted evidence with annotated ground truth evidence. If there is no annotated evidence, evidence F1 is 1 if no evidence was predicted and otherwise 0.

Attributability (ATT) For the other datasets, evidence F1 is insufficient: GovReport does not come with annotated evidence, and for datasets with *free-form* responses (QASPER and Natural Questions), evidence F1 is too rigid: A model might produce a response different from the ground truth, but supported by the retrieved evidence. Even though evidence quality is high, evidence F1 might be low. For these datasets, we evaluate evidence quality as attributability[7], (Gao et al., 2023b; Huang et al., 2021; Schimanski et al., 2024). We assume an attributability evaluation model $\mathrm{M_a}(E, r) \rightarrow \{1, 0\}$.

[3] `https://platform.openai.com/docs/models/gpt-3-5-turbo`

[4] `https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`

[5] Found to correlate better than BERTScore (Zhang* et al., 2020) with human judgment in our preliminary experiments.

[6] ROUGE has shown good correlation with human relevance judgments in Wu et al. (2024).

[7] Also known as citation recall (Gao et al., 2023b).

	Domain	Task	#Inst Train/Dev/Test	Doc #W	Res $\|R\|$/#W	#Evi	Evi Lvl
QASPER (QSP) [1]	Science	QA	2675/1005/1451	3937	1/12	1.7	para
Natural Questions (NQ) [2]	Wiki	QA	232191/6205/7307	4693	1/3	1	para
Evidence Inference (EI) [3]	Science	Cls	18545/1232/1218	3962	1/1	1.1	para
Wice (WIC) [4]	Web	FC	4234/349/358	1339	1/1	3.7	sent
ContractNLI (CNLI) [5]	Legal	NLI	7191/1037/2091	1697	1/1	1.5	para
GovReport (GR) [6]	US Gov.	Sum	15107/964/969	8464	20/517	N/A	N/A

Table 1: The datasets in LAB span multiple domains and task types. The numbers for Natural Questions differ from the original publication, as we filtered the instances (§B). Column names: Doc #: Average number of words per document. Res $|R|$/#W: Number of statements / Average number of words per response. #Evi: Number of annotated evidence segments per instance. Evi Lvl: Level of annotated evidence. Tasks: QA: Question answering. Cls: Classification. FC: Fact checking. NLI: Natural Language Inference. Sum: Summarization. [1] Dasigi et al. (2021) [2] Kwiatkowski et al. (2019) [3] DeYoung et al. (2020b) [4] Kamoi et al. (2023) [5] Koreeda and Manning (2021) [6] Huang et al. (2021)

The attributability score (AS) of a response is computed as the proportion of attributable statements.

$$\mathrm{ATT} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{M_a}(E_i, r_i)$$

When a response is abstained, we do not evaluate attributability, as no evidence is required. We additionally evaluate Unanswerable F1 for QA datasets:

Unanswerable F1 (UF1) We set this up as a classification task similar to Slobodkin et al. (2023). Gold labels are determined depending on the number of annotations (Kwiatkowski et al., 2019): For 3 annotations or less: "Unanswerable" if all annotators annotated unanswerable, else "answerable". More than 3 annotations: "Unanswerable" if at most one annotator did not annotate unanswerable, else "answerable". If a model abstains, its prediction is set to "unanswerable", else "answerable". To detect abstaining, we compiled a list of keywords based on (Slobodkin et al., 2023) and manual inspection. If a response contained a keyword, any predicted evidence was removed, and the response was set to "unanswerable" (see §G).

3.2.1 Attributability Evaluation Model Selection

To select a model for attributability evaluation, we created test datasets for QASPER, Natural Questions and GovReport, and evaluated TRUE (Honovich et al., 2022), Attrscore (Yue et al., 2023) and Minicheck (Tang et al., 2024). For Attrscore, we map "Contradictory" and "Extrapolatory" predictions to a single "not attributable" class.

Human annotation We generated attributed responses using GPT 3.5-`post-hoc` with BM25 for evidence retrieval. Two authors of this study, both holders of a Master's degree and fluent in English, annotated attributability (attributable or not attributable) for 200 responses (200 sentences for GovReport) and reached an agreement of 0.74 (QASPER), 0.77 (Natural Questions) and 0.76 (Govreport) Cohen's κ.

Model	**QSP**	**NQ**	**GR**
TRUE	79/80	**83/83**	**79/78**
AttrScore	76/71	68/67	76/52
Minicheck	**83/82**	82/82	79/70

Table 2: Attributability evaluation model selection. Metrics: accuracy / balanced accuracy (Tang et al., 2024)

Results We report accuracy (Gao et al., 2023b) and balanced accuracy (Tang et al., 2024) scores in Table 2. The scores are comparable to Gao et al. (2023b), who reported 85% accuracy and Tang et al. (2024), who reported between 59% and 84% balanced accuracy on their respective benchmarks. We select the best-performing model to evaluate attributability: Minicheck for QASPER, and TRUE for Natural Questions and GovReport.

3.3 Generation

Model selection We focus on two groups of LLMs with at least 8K tokens input length: (1) The large state of the art models GPT-3.5[8] and GPT-4[9], as they hold top positions in other long document benchmarks (Li et al., 2023b; Shaham et al., 2023; An et al., 2023) (2) Small (~3-7B) models that are accessible with limited resources. We tested a range of models with `citation` on QASPER and GovReport and selected Longchat-7B and Flan-T5-XL as the best performing for prompting and

[8] gpt-35-turbo-0613-16k
[9] gpt-4-turbo-128k

fine-tuning, respectively. For complete selection results and hyperparameters see §E.

Prompts We employ separate prompt sets for `citation` and `non-citation`. Similar to Shaham et al. (2023), we keep instructions short, including guidance on the expected responses and output format. Prompts contained three in-context examples, where documents were shortened to title, section headings and annotated evidence. For details and prompt optimization experiments, see §C.

3.4 Retrieval

Retriever selection For retrieval in `post-hoc`, `retrieve-then-read` and `reduced`, we employ sparse and dense retrievers that showed good performance in related work[10] (Thakur et al., 2021): BM25 (Robertson and Zaragoza, 2009; Gao et al., 2023b), GTR (Ni et al., 2021; Gao et al., 2023b; Bohnet et al., 2022), Contriever (Izacard et al., 2022; Xu et al., 2024; Bai et al., 2023), Dragon (Lin et al., 2023; Xu et al., 2024) and the best-performing Sentence Transformer "all-mpnet-base-v2"[11] (Reimers and Gurevych, 2019). For each combination of approach and task, we selected the best-performing retriever using GPT-3.5 as response generator (see §F).

Details In `retrieve-then-read` and `reduced`, queries were constructed based on information available in the instruction (e.g. question or a claim). For GovReport, similar to Zhang et al. (2020), we created queries from all document segments and retrieved paragraphs based on self-similariy. In `post-hoc`, queries were constructed based on the instruction and the generated response. For both `post-hoc` and `retrieve-then-read`, we set $k = 5$ for Wice, and $k = 2$ for all other datasets based on evidence statistics (see Table 1). In `reduced` approaches we set $k = 10$ based on (Xu et al., 2024) (see §F).

4 Experiments

4.1 RQ1: What are optimal approaches to attribution in long document tasks?

Table 3 shows the results from all combinations of selected models. Due to the large number of experiments, all results are from single runs.

[10]For efficiency reasons, we do not use LLMs for retrieval.

[11]https://sbert.net/docs/sentence_transformer/pretrained_models.html#semantic-search-models

Which approach produces the highest evidence quality? Flan-T5-XL has higher average scores than GPT-3.5 and GPT-4, while the Longchat scores are lower. For GPT-3.5, GPT-4 and Flan-T5-XL, `citation`/`reduced-citation` results in the best evidence quality on average and most datasets, and `retrieve-then-read` performs worst. `Post-hoc` works best for GovReport and Longchat.

Does `citation` hurt response quality? It could be assumed that `post-hoc` results in better response quality than `citation`, as task decomposition can improve performance (Gao et al., 2023a). We compare average response quality between (`reduced-`)`citation` and (`reduced-`)`post-hoc` for GPT-3.5, GPT-4 and Flan-T5-XL. In no case, the response quality for `citation` is more than 0.5 points lower than for `post-hoc`, showing that `citation` has a minimal effect on response quality.

Does reduction of the input document help? Comparing `reduced-post-hoc` to `post-hoc` and `reduced-citation` to `citation`, we find that response quality is mostly better for the non-`reduced` variant. Regarding evidence quality, our findings are not as clear, as `reduced-citation` results in the best average evidence quality for GPT-3.5.

Discussion `Citation` or `reduced-citation` result in the best average evidence quality, while not hurting response quality, in line with recent work showing LLM capabilities for retrieval (Ma et al., 2023). The GovReport task and the Longchat model are exceptions to this, as `post-hoc` results in better evidence quality in these cases. For GovReport, the higher evidence quality with `post-hoc` can be explained with the "repetitive" nature of the summarization task, since the high overlap between response statements and document provides good conditions for retrievers to find evidence. For small models such as Longchat, related work has shown that they lack instruction following capability to perform evidence extraction (Gao et al., 2023b; Schimanski et al., 2024), making `post-hoc` the better approach for the model.

Comparing models, fine-tuned Flan-T5-XL has higher average scores for response and evidence quality than the large prompted models GPT-3.5 and GPT-4. This could also be observed in related work (Huang et al., 2024; Schimanski et al., 2024). Similar to Xu et al. (2024) and Bai et al. (2023), we have not found a general beneficial effect of

		QSP			NQ			EI		WIC		CNLI		GR		Average	
		EM	ATT	UF1	EM	ATT	UF1	CF1	EF1	CF1	EF1	CF1	EF1	RL	ATT	RQ	EQ
GPT-3.5	`p-h`	45	62	65	**51**	44	**58**	77	22	**52**	48	44	40	**26**	**74**	49.17	48.33
	`rtr`	42	**78**	51	50	42	57	73	25	50	44	47	41	21	40	47.17	45.00
	`cit`	**52**	60	**67**	47	38	53	**78**	50	**52**	59	43	53	**26**	59	**49.67**	53.17
	`r-p-h`	48	71	58	50	**45**	57	**78**	22	**52**	48	45	40	22	64	49.17	48.33
	`r-cit`	50	70	64	46	38	53	**78**	**52**	48	**63**	**48**	**57**	22	58	48.67	**56.33**
GPT-4	`p-h`	65	68	68	52	42	56	**87**	24	32	36	**66**	50	**27**	**63**	54.83	47.17
	`rtr`	56	**83**	56	**58**	36	**61**	74	31	22	28	64	59	20	27	49.00	44.00
	`cit`	**68**	76	**71**	51	**49**	57	86	**64**	**35**	**47**	63	64	**27**	62	**55.00**	**60.33**
	`r-p-h`	62	71	69	51	43	57	84	24	28	30	65	50	22	54	52.00	45.33
	`r-cit`	63	76	66	54	**49**	58	83	49	33	46	64	**67**	22	56	53.17	57.17
Flan-T5	`p-h`	**55**	67	**58**	81	61	74	86	22	44	46	**77**	57	**27**	**90**	61.67	57.17
	`rtr`	43	**74**	55	79	62	74	77	25	43	43	64	58	19	67	54.17	54.83
	`cit`	53	57	56	**84**	**75**	71	83	**59**	**53**	**70**	**77**	**78**	25	71	**62.50**	**68.33**
	`r-p-h`	53	72	**58**	80	62	**75**	**87**	22	45	46	76	58	23	84	60.67	57.33
	`r-cit`	52	62	55	**84**	**75**	73	84	57	**53**	68	75	75	22	68	61.67	67.50
Longchat	`p-h`	23	57	**62**	21	33	38	**66**	22	23	**31**	38	33	**24**	**67**	**32.50**	**40.50**
	`rtr`	29	56	49	**27**	32	**43**	43	**25**	19	25	**39**	**39**	21	41	29.67	36.33
	`cit`	22	53	52	17	2	38	**66**	13	**24**	19	36	14	23	9	31.33	18.33
	`r-p-h`	**33**	**58**	54	25	**35**	41	61	22	20	27	33	30	22	63	32.33	39.17
	`r-cit`	26	48	53	21	4	42	65	13	20	25	24	12	22	24	29.67	21.00

Table 3: Evaluation on LAB, all scores show percentages. `Citation` / `reduced-citation` mostly perform best, with notable exceptions for GovReport and Longchat (see §4.1). `p-h`: `post-hoc`. `rtr`: `retrieve-then-read`. `cit`: `citation`. `r-`: `reduced-`. EM: Exact match F1, ATT: Attributability, UF1: Unanswerable F1, CF1: Classification F1, EF1: Evidence F1, RL: Rouge-L, RQ (EQ): Average response (evidence) quality, mean of all scores with blue (green) shade.

context reduction on response quality. The positive results from Agrawal et al. (2024) were obtained in a multi-document setting, where there is no logical coherence between the input documents. In contrast, the logical coherence in long documents can be disrupted through reduction, which can make processing of the reduced document more difficult.

4.2 RQ2: Do LLMs exhibit positional biases in attribution?

Several works have shown that LLM performance depends on the position of information in the input (Liu et al., 2024; Staniszewski et al., 2024; Ravaut et al., 2024). We investigate whether this phenomenon exists for attribution.

Do predicted and gold evidence distributions agree? Figure 3 (top) shows the predicted evidence distributions from `citation`[12] and the gold distributions.[13] The models generally follow the gold evidence distribution, with Longchat showing the strongest deviation. GovReport is an exception: all models show a higher focus on the beginning of the document, especially Flan-T5-XL.

[12]We focus on `citation` because only in this approach the LLM performs evidence retrieval.

[13]For GovReport, we matched summary sentences to paragraphs using BM25 to obtain gold evidence, see §3.1

Does response quality depend on the position of gold evidence? We grouped instances by evidence position and evaluated the approaches with full document input (`citation` and `post-hoc`) separately for each group (Fig. 3, bottom). The strong fluctuations of GPT-4 can be explained by the fact that it was evaluated on 100 samples only. It seems that response quality decreases as evidence appears later in the document. Similar to Ravaut et al. (2024), we computed Spearman correlation between response quality and evidence position, and find that correlation is mostly negative, but not always significant (Table 5).

Discussion Except for GovReport, we could not find positional biases in LLM evidence retrieval. From the results of Liu et al. (2024), we expected to find a "Lost in the Middle" effect, i.e. a reduction of retrieved evidence or performance in the middle of documents. Rather, we found a decrease in response quality towards the end of the document similar to Ravaut et al. (2024). We and Ravaut et al. (2024) work with coherent long documents, while Liu et al. (2024) work in a RAG-like setup, where the input are multiple documents without coherence, and models might expect an ordering by relevance. This might explain the different results.

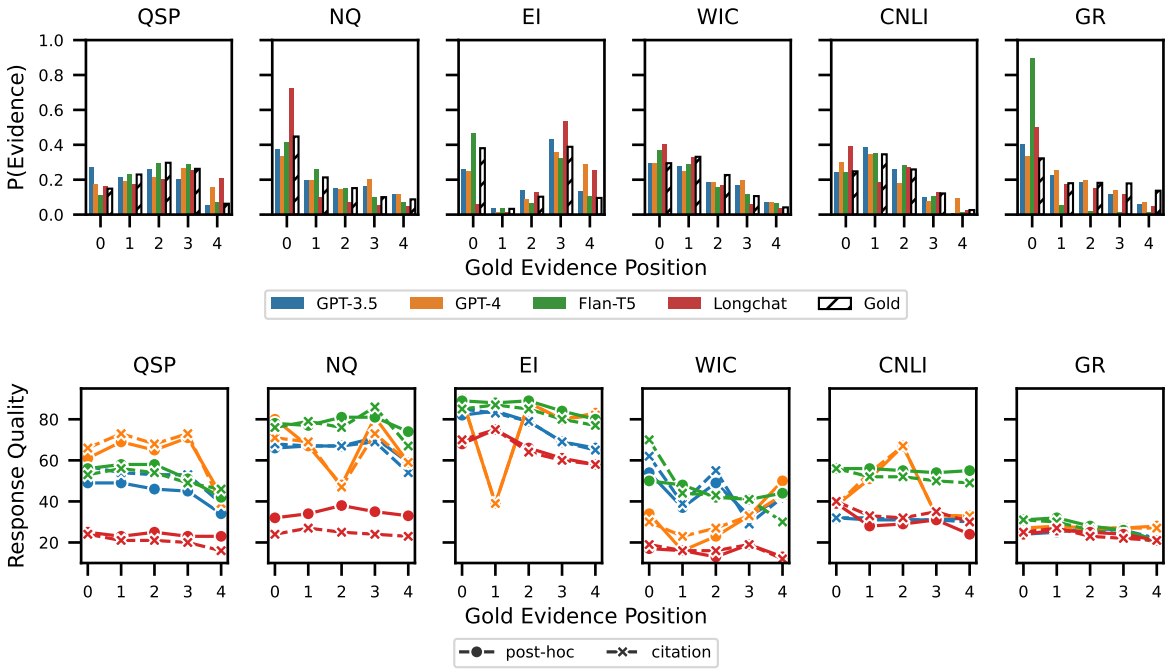

Figure 3: Top: Evidence distribution (predicted via `citation`) by position in the document. Except for GovReport, no positional bias is visible. Bottom: Response quality by position of gold evidence in the document. Negative correlation between evidence position and response quality is visible in several cases (Table 5) (see §4.2).

4.3 RQ3: What is the relation between evidence quality and response quality?

Attributability evaluation models (Honovich et al., 2022; Yue et al., 2023; Tang et al., 2024) can evaluate evidence quality without external reference. If evidence quality were positively correlated with response quality, this could be used to abstain from low-quality responses. We test this with *selective prediction* (El-Yaniv and Wiener, 2010), using attributability scores as an estimate of *confidence*.

Selective Prediction We begin by filtering responses that don't require evidence (e.g. "unanswerable" on QASPER or "not mentioned" on ContractNLI), as attributability cannot be evaluated on these. We sort the remaining l predictions by attributability,[14] obtaining the set of responses $\mathcal{R}_{sel}$ in descending order. For each $t \in \{0...l\}$, we compute average response quality[15] on the subset of $\mathcal{R}_{sel}$ up to t and the coverage (the proportion of responses evaluated). We evalate the confidence estimation by the area under the response quality-coverage curve (AUC) (Chen et al., 2023).

[14]We used Minicheck(Tang et al., 2024) for QASPER and TRUE (Honovich et al., 2022) for all other datasets (§3.2.1)

[15]As filtering responses that do not require evidence produces a strong class imbalance, we use micro F1 instead of macro F1 for Wice and ContractNLI

Is evidence quality a good estimate of confidence in selective prediction? Table 4 shows the AUC difference between ordering predictions by attributability and a random order baseline (mean of 10 repeats). We see that attributability is an effective estimate of confidence on Natural Questions, Evidence Inference and ContractNLI, and, to some extent on Wice. On QASPER, and GovReport, the difference to the random baseline is small.

Do unanswerable instances have lower evidence quality? For instances annotated as "unanswerable", "not supported" or "not mentioned", there is no annotated evidence. If models give different responses, the evidence quality should be low, and these should be filtered in selective prediction. Fig. 4 shows the distribution of gold responses in selective prediction. The proportion of such instances without sufficient evidence decreases with lower coverage for Natural Questions, Wice and ContractNLI but, surprisingly, not for QASPER.

Why does evidence quality fail to predict response quality? We consider GovReport a special case, as its long responses are evaluated in their entirety, which might be too coarse-grained to reflect the per-statement evaluation of evidence quality. This is corroborated by the fact that system-level correlation between evidence quality and

		QSP EM	QSP UF1	NQ EM	NQ UF1	EI CF1	WIC CF1	CNLI CF1	GR RL
GPT-3.5	**post-hoc**	3	2	17	20	12	4	19	0
	citation	0	0	14	17	14	4	24	-1
GPT-4	**post-hoc**	4	1	24	26	9	11	16	0
	citation	3	5	9	20	11	3	18	1
Flan-T5	**post-hoc**	-1	2	14	13	8	0	8	0
	citation	2	1	11	11	11	4	9	-1
Longchat	**post-hoc**	6	2	13	20	18	16	26	0
	citation	1	1	1	2	10	13	15	2

Table 4: Difference in response quality-coverage AUC between responses ordered by evidence quality (attributability) and random ordering. Evidence quality predicts response quality on several datasets (see §4.3)

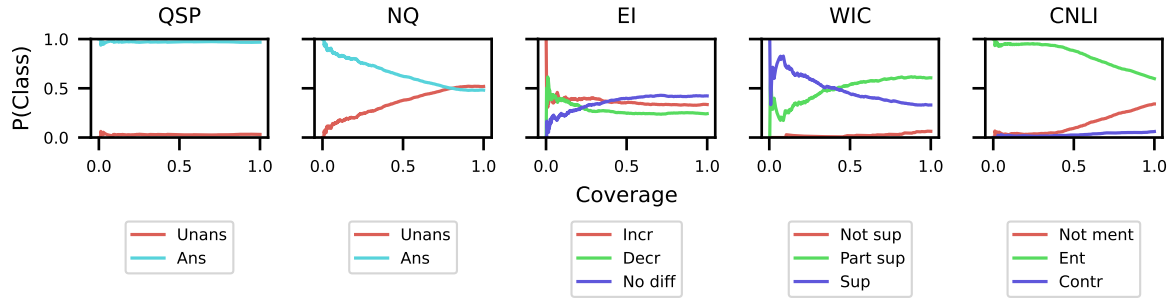

Figure 4: Gold class distribution in selective prediction for GPT-3.5-`citation`.

response quality is significantly positive for all datasets except GovReport (Table 6).

For Wice and QASPER, the possible causes are: (1) Responses are correct, but the evidence is insufficient. (2) Responses are incorrect, but the evidence is sufficient. (3) The attributability scores are wrong. We performed a manual analysis on the 50 responses with the lowest attributability from GPT-3.5-`citation`: For QASPER, the responses were often correct (answer F1 of 66), but the evidence was insufficient in 46 cases. For Wice, this could be observed in 39 cases. This implies that the LLM's failure to extract sufficient evidence is the main reason for low correlation between evidence and response quality.

For Wice, the attribution evaluation model failed to recognize evidence for "partially supported" claims in 8 cases, as it is only trained to distinguish "supported" and "not supported". This can be seen in Fig. 4, where the proportion of "partially supported" decreases with lower coverage.

Discussion We explain the differences in the relationship between evidence quality and response quality by the varying dataset complexity. While Natural Questions, ContractNLI and Evidence Inference focus on single facts (e.g. a single entity or a specific contractual obligation), Wice and QASPER instances often contain multiple facts (e.g. an enumeration or multiple subclaims), which is also reflected in the number of evidence segments per instance (Table 1). Models respond correctly, but fail to point to all necessary evidence. This is in line with related work on attribution: While Bohnet et al. (2022) found that response quality can be improved by attributability-based reranking on Natural Questions, Gao et al. (2023b) did not find this on their more complex benchmark.

5 Conclusion

In our experiments on LAB, we found that `citation` is a promising approach to attribution for large or fine-tuned models, while for small prompted models, `post-hoc` extraction can improve performance. We did not find a "Lost in the Middle" effect, but negative correlation between evidence position and response quality in some cases. Finally, we showed that evidence quality can predict response quality for responses with low complexity. We hope that our results, code and data spur further research on long document assistants, most prominently: (1) Improving the `citation` capabilities of LLMs for complex responses. (2) Combining attributability evaluation models and iterative self-refinement (Gao et al., 2023a) to try to improve abstained responses.

6 Limitations

Using LLMs for retrieval LLMs have shown good performance in reranking tasks (e.g. Ma et al. 2023). For efficiency reasons, we did not employ LLMs for retrieval. Instead, we employed state-of-the-art retrievers and implemented a rigorous selection procedure, elucidating the best retriever for each combination of task and approach. In the case of `reduced` approaches, we deem it unlikely to see large beneficial effects: The "pressure" on the retrievers was already low, as they only had to retrieve 1-3 relevant segments among 10 retrieved in total.
For `post-hoc` and `retrieve-then-read`, it could be that using LLMs for evidence retrieval improves performance. However this would not change our main claims: Our experiments with the `citation` approach already show that using LLMs for evidence retrieval works best. Therefore, we found the trade-off between increased computational cost and additional insights not favorable towards employing LLMs for evidence retrieval in `post-hoc` and `retrieve-then-read`. This might be different for researchers or practicioners interested in maximal performance.

Evaluation of evidence quality While our attributability model selection experiments showed that these models obtain good accuracy, our analysis in 4.3 showed that edge cases are not yet handled well. Research into solving such edge cases is a promising direction for future work.

Datasets We compiled a benchmark with diverse tasks, domains, response and document lengths, but naturally, we were not able to cover all variations of these properties. Many long document datasets and benchmarks are available (e.g. Li et al. 2023b; An et al. 2023; Shaham et al. 2023), but only few contain annotated evidence, which we required for the positional bias analysis in our paper (not finding a long document summarization dataset with annotated evidence, we resorted to GovReport). Given that we found only limited positional biases, extending our work to more datasets with and without annotated evidence is an interesting direction for future work.

7 Potential Risks

The goal of this work is to promote the development of more trustworthy long document assistants, which we deem a promising, but low-risk research goal. Similarly, we deem our research methodology to be of low risk: All datasets used were created for NLP research, do not contain personal, harmful or sensitive data and were published under permissive licenses (Table 7). We did not employ human annotators other than the authors of the study themselves.

8 References

References

Devanshu Agrawal, Shang Gao, and Martin Gajek. 2024. Can't remember details in long documents? you need some r&r. *arXiv:2403.05004*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv:2307.11088*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv:2308.14508*.

Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, and Peter Izsak. 2024. Cotar: Chain-of-thought attribution reasoning with multi-level granularity. *arXiv:2404.10513*.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv:2212.08037*.

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.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with self-evaluation to improve selective prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213, Singapore. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *lmsys.org/blog*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv:2307.08691*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020a. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020b. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099, Torino, Italia. ELRA and ICCL.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. *arXiv:2404.03381*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training language models to generate text with citations via fine-grained rewards. *arXiv:2402.04315*.

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.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *arXiv:2112.09118*.

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. *arXiv:2310.06825*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review. *Computational Linguistics*, 48(4):949–986.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023a. How long can open-source llms truly promise on context length? *lmsys.org/blog*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023b. Loogle: Can long-context language models understand long contexts? *arXiv:2311.04939*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv:2302.07452*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv:2305.02156*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. Expertqa: Expert-curated questions and attributed answers. *arXiv:2309.07852*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv:2402.09906*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *arXiv:2112.07899*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840.

Mathieu Ravaut, Aixin Sun, Nancy F. Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. *arXiv:2310.10570*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust llm specialists for evidence-based question-answering. *arXiv:2402.08277*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Konrad Staniszewski, Szymon Tworkowski, Yu Zhao, Sebastian Jaszczur, Henryk Michalewski, Łukasz Kuciński, and Piotr Miłoś. 2024. Structured packing in llm training improves long context utilization. *arXiv:2312.17296*.

Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, and Mohit Iyyer. 2023. Pearl: Prompting large language models to plan and execute actions over long documents. *arXiv:2305.14564*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv:2404.10774*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *arXiv:2403.08295*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Additional Results

A.1 Correlation between Evidence Position and Response Quality

See Table 5.

A.2 System-level Correlation of Attributability and Response Quality

.

See Table 6

		QSP	NQ	EI	WIC	CNLI	GR
GPT-3.5	post-hoc	-0.76*	-0.03	-0.77*	-0.49	0.42	-0.45
	citation	-0.26	-0.44	-0.62	-0.41	-0.02	-0.59
GPT-4	post-hoc	0.12	0.02	0.03	0.35	0.04	0.12
	citation	0.12	0.04	0.03	0.09	-0.19	0.03
Flan-T5	post-hoc	-0.32	0.05	-0.75*	-0.37	-0.07	-0.92*
	citation	-0.19	-0.13	-0.81*	-0.76*	-0.26	-0.95*
Longchat	post-hoc	-0.43	0.44	-0.33	-0.32	-0.58	-0.64*
	citation	-0.26	-0.28	-0.42	-0.30	-0.54	-0.81*

Table 5: Correlation between response quality and position of annotated evidence. See §4.2 for details.

	Pearson r (p)
QSP	0.86 (6.8×10^{-3})
NQ	0.99 (2.7×10^{-6})
EI	0.97 (3.9×10^{-5})
WIC	0.90 (2.3×10^{-3})
CNLI	0.93 (9.4×10^{-4})
GR	0.18 (6.7×10^{-1})
Avg	0.98 (1.2×10^{-5})

Table 6: System-level Pearson correlation between response quality and evidence quality for `citation` and `reduced-citation` (8 score pairs per dataset). For Avg, correlation was computed over the average scores (rightmost columns in Table 3).

QASPER	CC-BY 4.0
Natural Questions	CC BY SA 3.0
Evidence Inference	MIT
Wice	CC BY SA 4.0 (Text), ODC-BY (annotations)
ContractNLI	CC BY 4.0
GovReport	No copyright

Table 7: Licenses of the datasets in LAB.

B Datasets

See Table 8 for examples from the datasets in LAB

QASPER (Dasigi et al., 2021) is a dataset of NLP papers and questions about them. Answers can be extractive, abstractive, "Yes", "No" or "unanswerable". Evidence is annotated on paragraph level. We remove instances with evidence in tables or figures.

Natural Questions (Kwiatkowski et al., 2019) is a dataset of genuine questions from Google search logs and Wikipedia pages that may or may not contain the answers. We removed all annotations with answers in tables and those that only have a *long* answer, keeping only the annotations that have short answers (i.e. extractive spans), "Yes"/"No" answers or "unanswerable". All non-unanswerable annotations have a single evidence paragraph. As the official test set is hidden, we used the dev set for testing and a part of the train set for development.

Evidence Inference (DeYoung et al., 2020b) consists of reports from clinical studies, "prompts" in the form of *intervention*, *comparator*, and *outcome*, one or multiple labels for the prompt ("significantly increased", "significantly decreased", or "no significant difference") and corresponding evidence spans. We map the annotated evidence spans to paragraphs.

Wice (Kamoi et al., 2023) is a dataset of claims from Wikipedia and referenced webpages. Claims are annotated as "supported", "partially supported" or "not supported". The referenced webpages are annotated with evidence on sentence level. We use the full-claim subset.

ContractNLI (Koreeda and Manning, 2021) is a dataset of non-disclosure agreement contracts and claims about these agreements. The relation between contract and claim is annotated as "entailment", "contradiction" or "not mentioned". We split the contract documents into paragraphs at newline symbols to obtain paragraphs, and map the sentence-level annotated evidence to these paragraphs.

GovReport (Huang et al., 2021) is a dataset of reports from US-American governmental institutions and their executive summaries.

Dataset format All datasets were converted to the Intertext Graph format (Kuznetsov et al., 2022) to enable shared processing and the use of document structure (where available).

C Prompts

The prompts used in our experiments can be divided into two building blocks: (1) Instructioncand (2) instance specific input. The instruction further

consists of (a) task explanation and (b) format explanation. We explain the building blocks in the following. For a complete prompt example, see Table 13.

C.1 Instruction

C.1.1 Task explanation

Task explanations give information on the type of document used as input, the type of task to be solved, and possible labels. Table 9 shows the task explanations used.

C.1.2 Format explanation

Format explanations were only used for `citation` approaches, explaining the expected format of pointers to evidence segments. Table 10 shows the format explanations used for single-statement responses and multi-statement responses. The single statement explanation was used for all datasets except GovReport, where the multi statement explanation was used.

C.2 Instance Specific Input

Instance specific input consisted of an input document and task-dependent additional information, such as a question or a claim. Table 11 shows the formatting of instance specific input.

C.3 Example document formatting

We shortened the documents in examples to document title, section headings (where available) and annotated evidence segments. If there were no annotated evidence segments (e.g. because an example instance is unanswerable) we selected 2 random segments from the document (5 for Wice).

C.4 Prompt Selection

We optimized two prompt properties: The position of the instruction (i.e. task explanation and format explanation) and the number of few-shot examples. We ran experiments employing GPT-3.5 on QASPER and GovReport under the `citation` approach, the results are shown in Table 12. We first varied the position of the instruction, finding that the instruction before the instance specific input resulted in best performance. Next, we experimented with using 1, 2 or 3 few shot examples, finding that 3 examples resulted in best performance. We limited the number of few-shot examples to leave enough space for the input document.

C.5 Complete prompt example

See Table 13.

D Attributability Evaluation

To evaluate attributability, we experimented with TRUE (Honovich et al., 2022), Attrscore (Flan-T5-XXL version) (Yue et al., 2023) and Minicheck (Flan-T5-Large version) (Tang et al., 2024). These models expect a *claim* and *evidence* as input. In the following we explain the construction of claims, evidence formatting and model-specific prompts.

D.1 Claim Construction

Claims were constructed based on the task specific inputs and outputs. Table 14 shows examples.

QA datasets For QASPER and Natural Questions, question and answer were concatenated to get the claim.

Evidence Inference If the response was "no significant difference", the claim was formulated as "There was no significant difference between the effect of {intervention} and the effect of {comparator} on {outcome}.". If the response was "significantly increased" or "significantly decreased" were predicted, the claim was formulated as "The {intervention} {response} {outcome} in comparison to {comparator}".

Wice If the response was "supported", the input claim was used as the claim for attributability evaluation. If the response was "partially supported", the claim for attributability evaluation was formulated as "The claim {claim} is partially supported." "Not supported" responses did not require attributability evaluation.

ContractNLI If the response was "entailment", the input claim was used as the claim for attributability evaluation. As there are only 20 claims in the complete dataset, we formulated an *inverse* version of each claim. This was used as the claim when the response was "contradiction". See Table 14 for an example.

GovReport The generated summary sentences were used as claims.

D.2 Input Formatting

Evidence formatting Predicted evidence segments were ordered by occurence in the document, joined by newline symbols, and prepended with the document title.

Model-specific formatting The prompt templates for attributability evaluation can be seen in Table 15. They were taken from the respective original publications.

D.3 Annotation Instructions

As mentioned in 3.2.1, we manually annotated predictions on QASPER, Natural Questions and GovReport. The annotation instructions are shown in Fig. 5.

E Generation

E.1 Model Selection

To select open source models for prompting and fine-tuning, we compared their performance in preliminary experiments. Table 16 shows all open source models considered.

Fine-tuning We fine-tuned all candidate models for 1000 steps on QASPER and evaluated them on the dev set. As the results in Table 17 show that Flan-T5-XL has a clear advantage over the other models, we used it in all further fine-tuning experiments.

Prompting We evaluated all candidate models on 100 instances from the QASPER and GovReport dev sets, using the `citation` approach with the prompts described in §C. Table 18 shows the results. As the Longchat model obtained the highest average score, we used it in all further experiments.

E.2 Hyperparameters

Generation We set the maximum input length to 16K, truncating the input document if needed. We performed greedy decoding and temperature 0 for best reproducibility (§G).

Fine-tuning We employed LoRA fine-tuning (Hu et al., 2022) in a `citation` setting and a non-`citation` setting for 1000 steps. We set $r = 64$, $\alpha = 16$, a dropout rate of 0.1, a learning rate of $10^{-}4$, effective batch size of 8, and employed an AdamW optimizer (Loshchilov and Hutter, 2019)

F Retrieval

F.1 Retriever Selection

Candidates We experimented with BM25 (Robertson and Zaragoza, 2009), GTR (Ni et al., 2021), Contriever (Izacard et al., 2022), Dragon (Lin et al., 2023) and the best-performing Sentence Transformer "all-mpnet-base-v2"[16] (Reimers and Gurevych, 2019).

Results We tested all combinations of `post-hoc` (Table 21), `retrieve-then-read` (Table 19) and `reduced-citation` (Table 20) approaches, tasks and retrievers, using GPT-3.5 to generate responses. We selected the best-performing retriever for each combination

F.2 Query Construction

`Post-hoc` query construction `Post-hoc` queries were constructed by combining instance specific inputs and outputs. The exact query construction depended on the task. For QASPER and Natural Questions, question and generated response were concatenated. For Evidence Inference, ContractNLI and GovReport, `post-hoc` queries were constructed in the same manner as claims for attributability evaluation (see §D). For Wice, the input claim was used as the query. See Table 22 for examples.

`Reduce` and `retrieve-then-read` query construction Queries for `Reduce` and `retrieve-then-read` were constructed based on instance specific input. For QASPER and Natural Questions, this was the question. For Evidence Inference, this was the question formed out of intervention, comparator and outcome. For Wice and ContractNLI, this was the claim. For GovReport, these were the document paragraphs. See Table 23 for examples.

`Reduce` and `retrieve-then-read` for GovReport To find the most relevant paragraphs from documents in the GovReport dataset, we used each paragraph as a query and computed the retrieval score for all paragraphs (including the paragraph itself), resulting in n^2 scores $s_{i,j}$ for a document with n paragraphs. Each $s_{i,j}$ is the score for retrieving p_j with query p_i. We compute a single score for each paragraph as $s^*_j = \sum_{i=0}^{n} s_{i,j}$, i.e. the sum of scores to retrieve p_j. The paragraphs with the highest s^* are then selected.

G Technical Details

Technical setup GPT-35 and GPT-4 were accessed via the Azure OpenAI API [17], all other mod-

[16] https://sbert.net/docs/sentence_transformer/pretrained_models.html#semantic-search-models

[17] https://azure.microsoft.com/en-us/products/ai-services/openai-service

els were downloaded and run locally via Huggingface Transformers (Wolf et al., 2020) on NVIDIA A100 and H100 GPUs.

Rouge Scoring We used the `rouge-score` package[18] to evaluate ROUGE-L

Compute Budget We spent around $400 to access OpenAI models, including preliminary experiments. We estimate to have spent 300 GPU hours on all experiments, including fine tuning, inference and attributability evaluation.

Use of AI assistants We used Github Copilot[19] for coding assistance.

[18] https://pypi.org/project/rouge-score/
[19] https://github.com/features/copilot

	Input	Output
QASPER	Question: Which domains do they explore?	news articles related to Islam and articles discussing Islam basics
Natural Questions	Question: who won the 2017 ncaa mens basketball tournament	North Carolina
Evidence Inference	Question: With respect to Quality of life, characterize the reported difference between patients receiving good motivation/capability and those receiving inadequate motivation/capability.Choose between 'significantly decreased', 'no significant difference', and 'significantly increased'.	no significant difference
Wice	Claim: Having over 3,000 animals of nearly 400 different species, the zoo has slowly increased its visitors and now ranks as the number one outdoor tourist attraction in the state. Additional Info: The Sedgwick County Zoo is an AZA-accredited wildlife park and major attraction in Wichita, Kansas. Founded in 1971 with the help of the Sedgwick County Zoological Society, the zoo has quickly become recognized both nationally and internationally for its support of conservation programs and successful breeding of rare and endangered species.	Partially Supported
ContractNLI	Claim: Receiving Party shall not reverse engineer any objects which embody Disclosing Party's Confidential Information.	not mentioned
GovReport	Question: Write a one-page summary of the document. Structure your summary according to the following questions: 1. Why GAO Did This Study 2. What GAO Found 3. What GAO Recommends	{summary}

Table 8: Dataset examples

QASPER	You are given a Scientific Article and a Question. Answer the Question as concisely as you can, using a single phrase or sentence. If the Question cannot be answered based on the information in the Article, answer "unanswerable". If the question is a yes/no question, your answer should be "yes", "no", or "unanswerable". Do not provide any explanation. (If the question can be answered, provide one or several evidence paragraphs that can be used to verify the answer. Give as few paragraphs as possible.)
Natural Questions	You are given a Wikipedia page and a question about that page. Answer the question as concisely as you can, using at most five (5) words. If the question cannot be answered based on the information in the article, write "unanswerable". If the question is a yes/no question, answer "yes", "no", or "unanswerable". Do not provide any explanation. (If the question can be answered, provide one evidence paragraph that can be used to verify the answer.)
Evidence Inference	You are given a clinical study report and a question. Assess the effect of a treatment on a clinical outcome, compared to a control treatment. The options are "significantly increased", "significantly decreased" and "no significant difference". Do not provide any explanation. (Provide one or several evidence paragraphs that can be used to verify the answer. Give as few paragraphs as possible.)
Wice	You are given a document and a claim. Evaluate if the claim is supported by the document. You can choose between "supported", "partially supported" and "not supported". Do not add any explanation. (If you answer "supported" or "partially supported", provide the evidence sentences from the document that can be used to verify the answer. Give as few sentences as possible.)
ContractNLI	You are given a non disclosure agreement contract and a statement. Determine the relation between the contract and the statement. You can choose between "entailment", "contradiction" and "not mentioned". Do not add any explanation. (If you answer "entailment" or "contradiction", provide the evidence paragraphs from the contract that can be used to verify the answer. Give as few paragraphs as possible.)
GovReport	You are given a government report document. Write a one page summary of the document. (Each sentence in your summary should reference the source paragraphs from the document that can be used to verify the summary sentence.)

Table 9: Task explanations for the datasets in LAB. Text in parentheses at the end was only shown when `citation` approaches were used.

Annotation Instructions

Instructions

You have received a spreadsheet with several columns. Depending on the dataset, only some of the columns might be relevant for annotation.

- Question answering: “label”, “question”, “answer” and “predicted_evidence”.
- Summarization: “label”, “answer”, “predicted_evidence”.

Your job is to decide whether the predicted evidence fully supports the predicted answer. If it does, put 2 into the label column. If it partly supports the answer (there is evidence for only some of the facts in the answer), put 1 in the label column. If it does not support the answer, put 0 in the CR column

If there is no evidence, move to the next row.

To put a 2 in the CR column, the evidence should contain all necessary information in the answer.
Example 1 (fully supports) → CR = 2

- Question: “what ner models were evaluated?”
- Predicted Answer: “Stanford NER, spaCy 2.0, and a recurrent model similar to BIBREF13, BIBREF14”
- Predicted Evidence: “"In this section we describe a number of experiments targeted to compare the performance of popular named entity recognition algorithms on our data. We trained and evaluated Stanford NER, spaCy 2.0, and a recurrent model similar to BIBREF13 , BIBREF14 that uses bidirectional LSTM cells for character-based feature extraction and CRF, described in Guillaume Genthial's Sequence Tagging with Tensorflow blog post BIBREF15 .”

Example 2: (fully supports, but answer is incomplete) -> CR = 2

- Question: “What is the baseline?”
- Predicted Answer: “Nearest Number”
- Predicted Evidence: “Apart from learning-based baselines, we also create two naive baselines, one each for the Dosage and Frequency extraction tasks. For Dosage extraction, the baseline we consider is `Nearest Number', where we take the number nearest to the Medication Name as the prediction, and `none' if no number is mentioned or if the Medication Name is not detected in the input. For Frequency extraction, the baseline we consider is `Random Top-3' where we predict a random Frequency tag, from top-3 most frequent ones from our dataset - {`none', `daily', `twice a day'}.”

Example 3 (partially supports) → CR = 1

- Question: “How do they match words before reordering them?”
- Predicted Answer: “They use a dictionary to translate the sentences from English to the target language before reordering them”
- Predicted Evidence: 'Since the source language and the assisting language (English) have different word order, we hypothesize that it leads to inconsistencies in the contextual representations generated by the encoder for the two languages. In this paper, we propose to pre-order English sentences (assisting language sentences) to match the word-order of the source language and train the parent model on this pre-ordered corpus. In our experiments, we look at scenarios where the assisting language has SVO word order and the source language has SOV word order.’

Example 4 (does not support) → CR = 0

- Question: “Which information about text structure is included in the corpus?”
- Predicted Answer: “number of paragraphs”
- Predicted Evidence: “For the webpages, a static dump of all documents was created. Following this, the documents were manually checked to verify the language. The main content was subsequently extracted, i.e., HTML markup and boilerplate removed using the Beautiful Soup library for Python. Information on text structure (e.g., paragraphs, lines) and typography (e.g., boldface, italics) was retained. Similarly, image information (content, position, and dimensions of an image) was preserved”

If the answer is not a viable answer to the question, put 0 in the CR column.

Example 1 (off topic) → CR = 0

- Question: “How is model compactness measured”
- Predicted Answer: “15.4 MB”
- Predicted Evidence: 'Even if LangID-High does not present a more accurate result, it does present a more compact one: LangID-High is 15.4 MB, while the combined wFST high resource models are 197.5 MB.’

Example 2 (wrong answer) → CR = 0

- Question: "What datasets did they use?”
- Predicted Answer: “Carnegie Mellon Pronouncing Dictionary”
- Predicted Evidence: 'In order to train a neural g2p system, one needs a large quantity of pronunciation data. A standard dataset for g2p is the Carnegie Mellon Pronouncing Dictionary BIBREF12 . However, that is a monolingual English resource, so it is unsuitable for our multilingual task. Instead, we use the multilingual pronunciation corpus collected by deri2016grapheme for all experiments. This corpus consists of spelling–pronunciation pairs extracted from Wiktionary. It is already partitioned into training and test sets. Corpus statistics are presented in Table TABREF10 .’

Example 3 (non-sensical answer) → CR = 0

Figure 5: Annotation instructions for attributability model evaluation.

Single statement	Your reply must have the following format: "<answer> [X] [Y]" In your reply, replace <answer> with your solution to the task. Your reply must be followed by the ids of the relevant segments from the document.
Multi statement	Your reply must have the following format: "<answer_sentence_1>[X] [Y] <answer_sentence_2>[Z]..." In your reply, replace <answer_sentence_1> with your first sentence, <answer_sentence_2> with your second sentence, and so forth. Each sentence must be followed by the ids of the segments relevant to the sentence.

Table 10: Format explanations. Multi statement was used for GovReport, Single Statement for all other datasets.

QSP	Scientific Article: {document} [End of Document] Question: {question}
NQ	Document: {document} [End of Document] Question: {question}
EI	Document: {document} [End of Document] Question: {question}
WIC	Document: {document} [End of Document] Claim: {statement} Additional Info: {additional_info}
CNLI	Contract: {document} [End of Document] Statement: {statement}
GR	Document: {document} [End of Document] {question}

Table 11: Formatting of instance specific input. See Table 8 for examples of task specific inputs.

	QSP				GR		
	AF1	ATT	UF1	Avg	RL	ATT	Avg
Inst before and after	42	75	14	43.77	22	43	32.60
Inst before	47	70	33	50.31	25	10	17.62
Inst after	40	60	20	40.16	21	14	17.85
Inst before, 1 example	53	67	43	54.21	26	36	30.97
Inst before, 2 examples	50	59	42	50.40	26	44	35.11
Inst before, 3 examples	50	66	67	**60.86**	28	54	**41.08**

Table 12: Prompt optimization results on GPT-3.5 under the `citation` approach. Inst before / after refers to the position of the instruction being before / after the task specific input. For complete instructions, see C Avg: Average of scores for one task.

Task Explanation	You are given a Scientific Article and a Question. Answer the Question as concisely as you can, using a single phrase or sentence. If the Question cannot be answered based on the information in the Article, answer "unanswerable". If the question is a yes/no question, your answer should be "yes", "no", or "unanswerable". Do not provide any explanation. (If the question can be answered, provide one or several evidence paragraphs that can be used to verify the answer. Give as few paragraphs as possible.)
Format explanation	Your reply must have the following format: "<answer> [X] [Y]" In your reply, replace <answer> with your solution to the task. Your reply must be followed by the ids of the relevant segments from the document.
Example 1	Scientific Article: Automated Hate Speech Detection and the Problem of Offensive Language Abstract {omitted} [End of Document] Question: What type of model do they train?
Example 2	{omitted}
Example 3	{omitted}
Instance specific input	Scientific Article: Combining Thesaurus Knowledge and Probabilistic Topic Models Abstract {omitted} [End of Document] Question: Which domains do they explore?

Table 13: Example of the final prompt format used for `citation` on QASPER.

	Input	Claim
QASPER	Question, response	"The answer to the question 'Which domains do they explore?' is 'news articles related to Islam and articles discussing Islam basics"'
Natural Questions	Question, response	"The answer to the question 'who won the 2017 ncaa mens basketball tournament?' is 'North Carolina"'
"Evidence Inference	Question, response	"There was no significant difference between the effect of good motivation/capability and the effect of inadequate motivation/capability on quality of life." "Good motivation/capability [significantly increased/significantly decreased] quality of life compared to inadequate motivation/capability"'
"Wice	Claim, response	"Having over 3,000 animals of nearly 400 different species, the zoo has slowly increased its visitors and now ranks as the number one outdoor tourist attraction in the state." "The claim 'Having over 3,000 animals of nearly 400 different species, the zoo has slowly increased its visitors and now ranks as the number one outdoor tourist attraction in the state' is partially supported."
"ContractNLI	Claim, response	"Receiving Party shall not reverse engineer any objects which embody Disclosing Party's Confidential Information." "Receiving Party may reverse engineer any objects which embody Disclosing Party's Confidential Information."'
GovReport	Response statement	"Improper payments in Medicaid increased from $29.1 billion in fiscal year 2015 to $36.7 billion in fiscal year 2017."

Table 14: Examples for claims constructed for attributability evaluation.

TRUE	premise: {evidence} hypothesis: {claim}
Attrscore	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Verify whether a given reference can support the claim. Options: ""Attributable, Extrapolatory or Contradictory. ### Input: Claim: {claim} Reference: {evidence} ### Response:
Minicheck	predict: {evidence}</s>{claim}

Table 15: Prompts for attributability evaluation models based on the respective publications.

	#Params	Reference
Gemma-7b-it	7B	Team et al. (2024)
GritLM-7B	7B	Muennighoff et al. (2024)
Mistral-7B-Instruct-v0.2	7B	Jiang et al. (2023)
Vicuna-7B-v1.5-16K	7B	Chiang et al. (2023)
Flan-T5-XL/XXL*	3B/11B	Longpre et al. (2023)
LongChat-7B-v1.5-32K	7B	Li et al. (2023a)
Llama3-8B-Instruct	8B	See [20]

Table 16: Open source models considered in selection experiments. *: Flan-T5-XL was used in fine-tuning, Flan-T5-XXL was used in prompting experiments.

	AF1	ATT	UF1	Avg
Gemma	0.41	0.28	0.25	0.27
GritLM	0.42	0.30	0.22	0.26
Mistral	**0.46**	0.34	**0.26**	0.30
Vicuna	0.42	0.31	0.24	0.28
Flan-T5-XL	0.45	**0.61**	0.22	**0.42**
LongChat	0.42	0.29	0.24	0.26
Llama3	0.44	0.37	0.25	0.31

Table 17: results of Fine-tuning Model Selection

Model	QASPER			GovReport		Avg
	Answer F1	Attr	Unans F1	R-L	Attr	
`Gemma`	22	22	0	20	1	12.51
`GritLM`	26	46	0	22	0	17.44
`Mistral`	**31**	47	0	**0.25**	0	19.33
`Vicuna-7B`	21	42	0	23	2	16.67
`Flan-T5-XXL`	26	28	**36**	12	1	18.32
`LongChat`	21	52	0	**0.25**	**16**	**22.31**
`Llama3`	17	**60**	0	23	0	18.78

Table 18: results of OS Prompt Model Selection

	QASPER				**Natural Questions**			
	AF1	ATT	UF1	Avg	AF1	ATT	UF1	Avg
BM25	32	60	42	44.49	44	33	59	45.17
SBERT	29	69	39	45.56	47	**53**	60	**53.36**
Contriever	37	75	50	54.20	47	51	61	52.89
Dragon	**39**	**79**	**52**	**56.65**	**48**	48	**63**	52.93
GTR	36	70	46	50.54	46	48	61	51.47
	Evidence Inference				**Wice**			
	CF1	EF1	Avg	-	CF1	EF1	Avg	-
BM25	77	16	46.46	-	29	36	32.71	-
SBERT	78	23	50.51	-	33	43	**42.50**	-
Contriever	**83**	29	**55.71**	-	26	42	33.87	-
Dragon	78	**33**	55.23	-	31	41	36.29	-
GTR	78	**33**	55.58	-	**33**	**43**	38.20	-
	Contract NLI				**Govreport**			
	CF1	EF1	Avg	-	RL	ATT	Avg	
BM25	43	34	38.35	-	27	33	29.96	-
SBERT	42	35	38.21	-	**28**	36	**31.79**	-
Contriever	38	37	36.80	-	24	30	26.92	-
Dragon	39	37	37.64	-	21	19	20.16	-
GTR	**44**	**39**	**41.39**	-	23	**37**	30.00	-

Table 19: Retriever selection for `retrieve-then-read`. Retrievers were combined with GPT-3.5 and were evaluated on 100 dev instances per dataset. The retriever that resulted in the best average score was used in all further experiments for the respective combination of `retrieve-then-read` and task.

	QASPER				Natural Questions			
	AF1	ATT	UF1	Avg	AF1	ATT	UF1	Avg
BM25	45	67	**63**	58.33	**42**	36	**59**	45.69
SBERT	40	65	51	51.79	41	39	55	45.13
Contriever	48	72	65	**61.64**	40	**42**	55	45.83
Dragon	**49**	**73**	56	59.20	**42**	**42**	58	**46.95**
GTR	47	71	**63**	60.17	**42**	41	58	46.63

	Evidence Inference				Wice			
	CF1	EF1	Avg	-	CF1	EF1	Avg	-
BM25	83	49	66.12	-	**44**	61	**52.38**	-
SBERT	82	50	65.75	-	40	**65**	52.13	-
Contriever	83	57	70.42	-	36	60	48.14	-
Dragon	**87**	54	70.51	-	37	64	50.49	-
GTR	86	**62**	**73.66**	-	37	62	49.75	-

	Contract NLI				Govreport			
	CF1	EF1	Avg	-	RL	ATT	Avg	
BM25	42	49	45.25	-	26	51	38.29	-
SBERT	44	56	50.11	-	**27**	**60**	**43.38**	-
Contriever	45	**58**	51.42	-	23	45	33.93	-
Dragon	**52**	55	**53.29**	-	22	43	32.12	-
GTR	46	54	49.94	-	23	51	37.23	-

Table 20: Retriever selection for `reduced` approaches. Retrievers were combined with GPT-3.5-`reduced-citation` and were evaluated on 100 dev instances per dataset. The retriever the resulted in the best average score was used in all further experiments for the respective combination of `reduced-citation` / `reduced-post-hoc` and task.

	QASPER				Natural Questions			
	AF1	ATT	UF1	Avg	AF1	ATT	UF1	Avg
BM25	52	65	73	63.25	41	40	57	45.92
SBERT	52	55	73	64.12	41	50	57	53.35
Contriever	52	63	73	68.42	41	**54**	57	**55.28**
Dragon	52	**69**	73	**71.11**	41	**54**	57	**55.28**
GTR	52	59	73	66.27	41	**54**	57	**55.28**

	Evidence Inference				Wice			
	CF1	EF1	Avg	-	CF1	EF1	Avg	-
BM25	86	25	55.60	-	86	25	55.60	-
SBERT	86	20	52.94	-	86	20	52.94	-
Contriever	86	27	56.60	-	86	27	56.60	-
Dragon	86	**28**	**57.27**	-	86	**28**	**57.27**	-
GTR	86	24	55.27	-	86	24	55.27	-

	Contract NLI				Govreport			
	CF1	EF1	Avg	-	RL	ATT	Avg	
BM25	46	36	41.14	-	28	**73**	**50.55**	-
SBERT	46	36	41.24	-	28	60	44.14	-
Contriever	46	39	42.69	-	28	65	46.63	-
Dragon	46	40	43.16	-	28	68	48.22	-
GTR	46	**41**	**43.57**	-	28	61	44.61	-

Table 21: Retriever selection for `post-hoc` approaches. Retrievers were combined with GPT-3.5-`post-hoc` and were evaluated on 100 dev instances per dataset. The retriever the resulted in the best average score was used in all further experiments for the respective combination of `post-hoc` and task.

	Input	Query
QASPER	Question, response	"Which domains do they explore? news articles related to Islam and articles discussing Islam basics"
Natural Questions	Question, response	"who won the 2017 ncaa mens basketball tournament? North Carolina"
Evidence Inference	Question, response	"There was no significant difference between the effect of good motivation/capability and the effect of inadequate motivation/capability on quality of life." "Good motivation/capability [significantly increased/significantly decreased] quality of life compared to inadequate motivation/capability"
Wice	Claim	"Having over 3,000 animals of nearly 400 different species, the zoo has slowly increased its visitors and now ranks as the number one outdoor tourist attraction in the state"
ContractNLI	Claim, response	"Receiving Party shall not reverse engineer any objects which embody Disclosing Party's Confidential Information." "Receiving Party may reverse engineer any objects which embody Disclosing Party's Confidential Information."
GovReport	Response statement	"Improper payments in Medicaid increased from $29.1 billion in fiscal year 2015 to $36.7 billion in fiscal year 2017."

Table 22: Examples for queries for `post-hoc` evidence retrieval.

	Input	Query
QASPER	Question	"Which domains do they explore?"
Natural Questions	Question	"who won the 2017 ncaa mens basketball tournament?"
Evidence Inference	Question	"With respect to Quality of life, characterize the reported difference between patients receiving good motivation/capability and those receiving inadequate motivation/capability. Choose between 'significantly decreased', 'no significant difference', and 'significantly increased'."
Wice	Claim	"Having over 3,000 animals of nearly 400 different species, the zoo has slowly increased its visitors and now ranks as the number one outdoor tourist attraction in the state"
ContractNLI	Claim	"Receiving Party shall not reverse engineer any objects which embody Disclosing Party's Confidential Information."
GovReport	Paragraph	"Medicaid has been on our high-risk list since 2003, in part, because of concerns about the adequacy of fiscal oversight and the program's improper payments—including payments made..."

Table 23: Examples for queries for `retrieve-then-read` and `reduced` retrieval.