

INFOLOSSQA: Characterizing and Recovering Information Loss in Text Simplification

Jan Trienes^{1,2,4*} Sebastian Joseph³ Jörg Schlötterer^{4,5}

Christin Seifert⁴ Kyle Lo⁶ Wei Xu⁷ Byron C. Wallace⁸ Junyi Jessy Li³

¹University of Duisburg-Essen ²Institute for AI in Medicine, University Hospital Essen

³The University of Texas at Austin ⁴University of Marburg ⁵University of Mannheim

⁶Allen Institute for AI ⁷Georgia Institute of Technology ⁸Northeastern University

jan.trienes@uni-marburg.de jessy@utexas.edu

Abstract

Text simplification aims to make technical texts more accessible to laypeople but often results in deletion of information and vagueness. This work proposes INFOLOSSQA, a framework to characterize and recover simplification-induced information loss in form of question-and-answer (QA) pairs. Building on the theory of Questions Under Discussion, the QA pairs are designed to help readers deepen their knowledge of a text. First, we collect a dataset of 1,000 linguist-curated QA pairs derived from 104 LLM simplifications of English medical study abstracts. Our analyses of this data reveal that information loss occurs frequently, and that the QA pairs give a high-level overview of what information was lost. Second, we devise two methods for this task: end-to-end prompting of open-source and commercial language models, and a natural language inference pipeline. With a novel evaluation framework considering the correctness of QA pairs and their linguistic suitability, our expert evaluation reveals that models struggle to reliably identify information loss and applying similar standards as humans at what constitutes information loss.¹

1 Introduction

Technical texts, many of which exist in high-stake domains (e.g., medicine), are often written in a language incomprehensible to lay readers. Improving the accessibility of such texts may help address wider social issues, e.g., disinformation (UNSDG, 2021) and access to higher education (Goff et al., 2004). Automatic text simplification that rewrites text into plain language may therefore be a technology for good. With the adoption of LLMs, document-level text simplification has significantly progressed in recent years (August et al., 2023; Laban et al., 2023; Agrawal and Carpuat, 2024).

*Work done while visiting UT Austin.

¹Code, dataset and an interactive data viewer is available at <https://jantrienes.github.io/ts-info-loss/>.

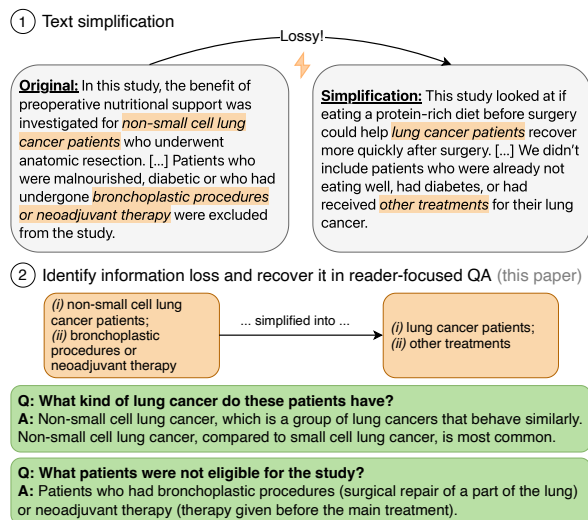


Figure 1: The goal of INFOLOSSQA is to generate a series of QA pairs that reveal to lay readers what information a simplified text lacks compared to its original.

Simplification is an inherently *lossy* process: Even when done by professional editors, the resulting plain language tends to lose details and become more generic (Li and Nenkova, 2015), and some content is omitted (Zhong et al., 2020). The *over-simplification* of content—including excessive deletion and vagueness—may lead to reduced comprehension (Agrawal and Carpuat, 2024) or in its worst case to misinterpretations and factual errors (Devaraj et al., 2022). Therefore, we consider the following question: *How can we characterize information loss and help readers recover what is lost in an intuitive and understandable manner?*

This paper presents a question generation and answering framework to characterize and recover information loss during simplification in a reader-centric way. Inspired by the theoretical work of “potential questions” (Onea, 2016) and expectation-driven Questions Under Discussion (QUD, Kehler and Rohde, 2017), information loss is **characterized** in the form of pragmatically felicitous ques-

tions rooted in simplified texts, without assuming access to the original text. Information loss is then **recovered** by answering these questions in plain language, based on the original (technical) text (Figure 1). Informed by QUD, the questions are curiosity-driven and information-seeking, and are subject to well-defined linguistic constraints (Riester et al., 2018; Wu et al., 2023). While prior work used factoid QAs to describe specific factual changes (Cole et al., 2023), the QUD framework suggests higher-level questions, and allows us to characterize general yet subtle language changes pervasive in factually equivalent texts. These include lower specificity (Louis and Nenkova, 2011) and greater degrees of vagueness (van Deemter, 2012). We envision the resulting QA pairs being directly incorporated in interactive simplification tools (Fok et al., 2023).

Contribution 1: Task formulation and dataset.

We first introduce INFOLOSSQA, a linguistically curated dataset grounded in the scenario of lay readers digesting simplified abstracts of randomized controlled trials (RCTs) in medicine. We focus on RCTs because they combine important and illustrative properties: They use highly technical/inaccessible language; lay audiences often having keen interest in their content; and misinterpretations are risky in the healthcare domain. INFOLOSSQA consists of 1,000 QA pairs across 104 RCT abstracts simplified by GPT-4 (prior work has shown that GPT-3.5 is a capable simplifier of medical texts; Shaib et al., 2023). The QA pairs provide a rich description of what information is lost both on a pragmatic and semantic level. Our analyses show that questions often elaborate about *why* and *how* an RCT was conducted, going beyond frequently asked questions in this domain (August et al., 2023).

Contribution 2: Automatic methods.

We then present two methods to perform the question generation and answering task. First, an end-to-end model, where we prompt several open-source and API-based LLMs with detailed task instructions and a one-shot example. Second, a natural language inference (NLI) pipeline which reasons over entailment of atomic facts (Stacey et al., 2023).

Contribution 3: Evaluation framework and human quality judgments.

Finally, we design a comprehensive framework to evaluate models at this task. The framework includes 10 quality desiderata evaluating the correctness of QAs,

their linguistic suitability, and the recall of human-written QAs. Expert judgments are collected on more than 400 generated QA pairs. We find that models have good question generation and answering capabilities but fall short at reliably identifying information loss. In this regard, the NLI pipeline is more effective than open-source LLMs, but produces QA pairs with a smaller granularity.

2 The INFOLOSSQA Task

Motivation. Information loss is inherent to simplification, insofar as the task typically demands producing outputs which are both comparatively uncomplicated *and* reasonably concise. Characterizing and revealing the elided content is critical to (1) provide a path for users to achieve better understanding than only seeing the simplified text, supporting users’ growth (Gooding, 2022), and (2) to allow researchers to evaluate the quality of a simplified text, given that omitting or oversimplifying key information may yield misleading texts (Devaraj et al., 2022).

Task formulation. Given a pair of documents, the original text X_{src} and the simplified text X_{tgt} , the task entails generating QA pairs that each reveal to lay readers one instance of information loss. We define an instance of information loss as a tuple $(X_{src}, X_{tgt}, Q, A, E, R)$. A potential question Q asks for information absent from X_{tgt} . The answer A provides the missing information. We additionally require two forms of localization or grounding to support richer analysis and scrutiny of errors in the generated QA pairs. First, each answer A must be supported by *evidence* $E \in X_{src}$ extracted from the original text. Second, each QA pair may be supported by an extractive *rationale* $R \in X_{tgt}$ which localizes where the information loss or vagueness occurs within the simplified text. Note that while E always exists, R may not (some information in X_{src} may be completely absent from X_{tgt}).

This formulation permits multiple information loss instances for a given X_{src} and X_{tgt} pair. We do not constrain the length or linguistic unit(s) of the extractive spans (E and R): they can comprise words, phrases, sentences, or entire paragraphs, and a single E or R can be one or more spans.

Types of information loss. For the purposes of this study, we define two types of information loss:

1. **Deletion.** Pieces of information which were not included in the simplification.

2. **Oversimplification.** Pieces of information that were simplified to the extent that they are vague or devoid of their original meaning. This is where the rationale R is identified.

These categories are meant to be flexible to capture a variety of information transformations, but also concrete to be consistently applied in an annotation protocol (Section 3).

Linguistic suitability. One of our primary goals is to generate QA pairs that may enhance users' comprehension of a text going beyond its simplified version (Fok et al., 2023). To this end, we pose two requirements for the QA pairs:

(1) *Readability.* The language level of the questions and answers should match that of the simplified text. That is, while the QA must discuss technical material from the original text, it should be explained at a level appropriate for the reader.²

(2) *Givenness.* Questions should be pragmatically felicitous to be understood by a reader without having seen the answer or the original text. The theory of QUD formalizes this through the Givenness constraint (Riester et al., 2018; Wu et al., 2023) which specifies that Q should not contain concepts that are hearer-new (Markert et al., 2012) with respect to a common ground. Here, the common ground is the simplified text X_{tgt} . Intuitively, this means that it should be clear from reading the question how the answer would expand on what a reader already knows from the simplified text.

Figure 2 illustrates the different elements and challenging nature of the task. To identify information loss, models cannot rely on lexical overlap, ordering, or other surface-level properties.

3 Data Collection

3.1 Simplification Corpus

We focus on medical texts, which provide an important and illustrative case where lay individuals may be keenly interested in understanding newly published science. Specifically, we consider abstracts of medical journal articles that describe the conduct and findings of RCTs. RCTs are the foundation of evidence-based medicine that informs patient care (Sackett, 1998). But those articles use technical language, which is effectively inaccessible to lay patients (Zuccala, 2010; August et al., 2023). This means that individuals may already interact

²It is possible that X_{tgt} itself is not at the right level for a given reader. However, this is out of scope of our study.

Original (X_{src}): These results indicate that acute/chronic endurance (E_1) arm-cranking with EMS applied to the lower limbs improves the brachial artery endothelial function (E_2) more markedly than the same exercise without EMS.

Simplified (X_{tgt}): The study concluded that doing the arm-cranking exercise with EMS on the lower body can improve arm function (R_2) more than without it.

Q_1 : On what timeframes does EMS improve training? A_1 : It helps in single (acute) or repeated applications (chronic).

Q_2 : How did the researchers measure how well EMS works? A_2 : They measure to what extent the main artery of the arm widens, which is called flow-mediated vasodilation.

Q'_1 (ℓ): Did EMS improve training in acute and chronic applications?

Figure 2: Example with a **Deletion** (“acute/chronic”) and an **Oversimplification** (“improve arm function” is too broad given that EMS improves “artery function”). These give rise to two QA pairs (Q_1 and Q_2) which fulfill the Readability and Givenness constraints. For contrast, Q'_1 violates (ℓ) givenness. Q_1 is likely more natural to lay readers because it could be asked without having seen the original text (no presupposition that the study looked at short-term and long-term effects).

with LLMs to simplify such texts. But automated simplification of RCTs may result in information loss, which risks readers misinterpreting findings.

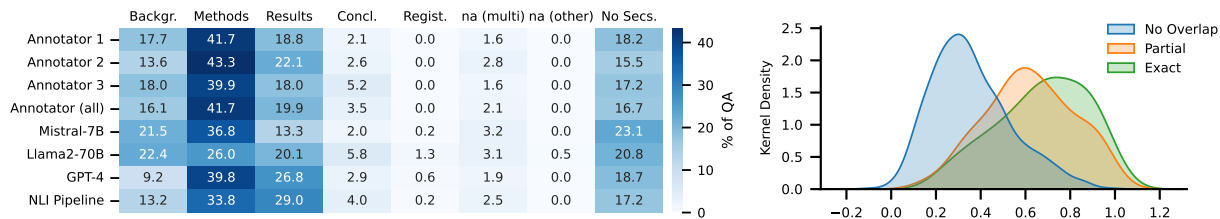
To better quantify information loss in LLM-simplifications, we create a parallel simplification corpus with GPT-4 (gpt-4-0613).³ On the related task of plain language summarization of full RCTs, Shaib et al. (2023) observed low error rates (<10% of summaries affected), but we do not yet have a quantitative understanding of information loss in LLM-simplifications. We sample RCT abstracts from the Evidence Inference v2.0 dataset (DeYoung et al., 2020b; Lehman et al., 2019), originally sourced from PubMed. Table A.1 provides the simplification prompt and example outputs.

3.2 Annotation Process

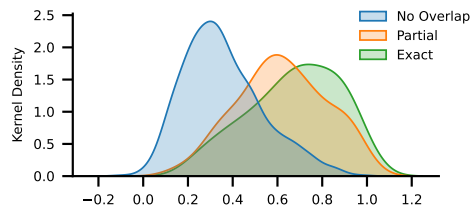
The process for annotating INFOLOSSQA is as follows. First, annotators examine both the original and simplified versions of the text. Next they identify and highlight information loss spans. For each highlighted instance, annotators compose a QA pair. We used the Thresh toolkit for annotation (Heineman et al., 2023). See Appendix B for annotation guidelines and interface screenshots.

Scenario grounding. Technical texts like RCTs have a long tail of information that can be lost in simplification (e.g., numerical results and significance values). Annotating *all* information loss

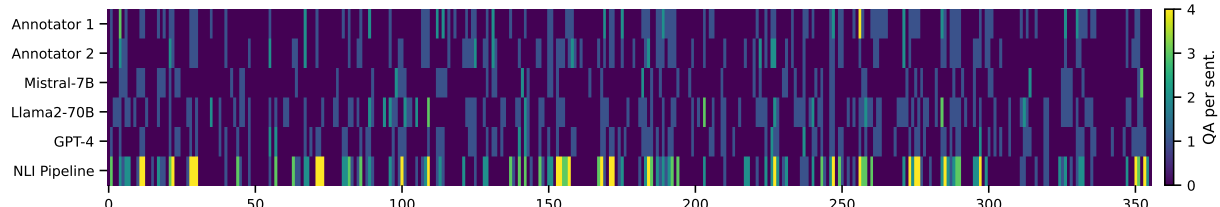
³No curated parallel corpus exists for RCT simplification.



(a) Distribution of QA pairs over RCT abstract sections. Some QAs are localized in multiple sections (multi.), in a rare section (other) or their abstracts do not have a section structure (19.2% of abstracts; no secs.).



(b) Cosine similarity of annotators' questions (Q) by whether their evidence spans (E) have no/partial/complete overlap.



(c) Number of QA pairs per sentence for a random selection of 25 documents. The x-axis are sentence indices.

Figure 3: Distribution of information loss. Humans produce a similar distribution of questions by section (a), but the questions differ in their localization (c). A similar localization results in more similar questions (b). Comparing humans to models, we see differences where questions are localized, and by extension also what they are about.

is intractable, and the resulting QAs would likely overwhelm readers. Therefore, we grounded annotation in a scenario. Annotators were asked to imagine a patient who could benefit from better understanding the RCT at hand, and to focus on information that may be salient in this scenario.

This notion of saliency is subjective as people have different information needs. Similarly, prior work in QUD (Ko et al., 2022) showed that question generation by human annotators is also subjective: People ask different questions even when the answer to these questions is the same. In our dataset, each document was annotated by two annotators, independently. This allows us to evaluate the extent to which people ask similar questions (Section 3.3).

Annotators. We hired three expert annotators who are native English speakers, majoring in linguistics and experienced in annotating medical texts. To ensure high annotation quality, we conducted a pilot phase with written feedback and held frequent review meetings to discuss questions. Annotators were compensated \$15/hr.⁴ The median duration per document was 22 minutes.

Dataset statistics. We collected annotations on 104 abstract-simplification pairs (average number of tokens: $|X_{\text{src}}| = 312$; $|X_{\text{tgt}}| = 271$) yielding 1,000 questions and answers (average number of tokens: $|Q| = 11.4$; $|A| = 30.3$).⁵ Each annotator wrote on average between 4.1 and 5.7 QA pairs

(i.e., information loss instances) per document. We observe that oversimplification is more prevalent (74.1% of questions) in this corpus than complete deletion of information (25.9% of questions). Additional statistics are given in Table A.2.

3.3 Analysis of INFOLOSSQA

We analyze the fine-grained annotation of questions and their localization in form of evidence spans to get a better understanding of how models preserve (or omit) content during simplification.

Where in the text is most information lost? We combine the precise localization of a QA pair with structural elements of the abstracts to identify areas where the most information loss occurs. By convention, most RCT abstracts are structured, i.e., having section headings for Background, Methods, Results and Conclusion (most commonly). We use a regular expression to identify the start and end boundaries of sections. Because the specific naming of such sections differs across articles, we collected a lookup dictionary to normalize synonymous headers.⁶ Afterwards, we align QA pairs to sections based on their localization.

Figure 3 shows the distribution of QAs over RCT sections. Most information loss occurs in the Methods section (41.7% of QAs). This is intuitive as Methods sections are likely to be particularly technical. The least information loss occurs in the Conclusion (3.5% of QAs) section. A small percentage

⁴Higher than local minimum wage at the time of study.

⁵Tokenizing with NLTK (Bird et al., 2009).

⁶For example, “Design,” “Methods” and “Materials and Methods” are all mapped to “Methods.”

Question Type	% All	% O	% D	Examples
Procedural. Asking for the procedures, tools, or methods by which a certain outcome is achieved.	34.3	34.1	34.7	• How did they measure the patients’ opioid medication needs? • How did the study control for bias?
Concept. Asking for a definition of an event or a concept.	25.7	28.1	18.9	• What kind of hip surgery were patients undergoing? • What type of mental illnesses are being studied?
Extent. Asking for the extent or quantity of an event or a concept.	17.0	19.2	10.8	• How many patients were in each group? • How much lurasidone was given to the patients in the trial?
Comparison. Asking for comparison among multiple events or concepts.	8.3	8.9	6.6	• How much did the abnormal blood vessels reduce in group 1 compared to group 2? • How did pre-meal insulin affect platelet activation compared to the placebo?
Consequence. Asking for the consequences or results of an event.	5.7	3.0	13.5	• What else does increased blood flow do? • What was the correlation between measures for total daily calcium intake?
Cause. Asking for the cause or reason for an event or a concept.	4.3	1.6	12.0	• What was a motivation behind the study? • Why are soy products being looked at in this study?
Example. Asking for example(s) or instance(s) of an event or a concept.	3.4	4.3	0.8	• What are the other brain functions that were tested? • What medication is typically used for outpatients?
Disjunction. Asking for the true one given multiple events or concepts, where comparison among options is not needed.	0.6	0.5	0.8	• Which one of the three components did this group’s supplement contain? • Which gel saw the increase in beneficial microbes after 24 hours?
Verification. Asking for the truthfulness of an event or a concept.	0.7	0.3	1.9	• Were there any other changes in blood sugar level observed in the study? • Were both eyes of each patient tested?
Judgmental. Asking for the opinions of the answerer’s own.	Not observed.			

Table 1: Example questions in INFOLOSSQA classified according to the question taxonomy by Cao and Wang (2021). Percent of all questions is given by category (O = Oversimplification; D = Deletion).

Granularity (Measure)	Del.	Over.	All
Span, exact (F1)	0.20	0.13	0.17
Span, partial (F1)	0.34	0.34	0.34
Sentence (Krippendorff’s α)	–	–	0.50

Table 2: Evidence span (E) overlap between annotators. Del = Deletions; Over = Oversimplification.

of QAs spans multiple sections. These tend to be about concepts that were repeatedly oversimplified.

What types of questions are asked? To characterize the information that is typically lost, we categorize questions according to the 10 pragmatic function types defined by Cao and Wang (2021) using a few-shot prompting classifier.⁷ Table 1 shows the question distribution across types, their definition and examples. Most questions are of type *procedural*, requesting specific details about how the study was conducted (34.3%). This aligns with the earlier observation that most information loss occurs in Methods sections. The second most common type of questions seek more concrete definitions for overgeneralized *concepts* (25.7%), such as the health condition or clinical procedure. Comparing deletions and oversimplifications, we observe a similar distribution, with the exception of questions asking for the *consequences* (e.g., experimental re-

⁷Prompting gpt-4-0613 with annotation guidelines of Cao and Wang (2021) and few-shot examples achieves an F1 score of 0.87 on a validation set (Table A.1).

sults) or *causes* (e.g., what motivated the study) of an event being more prevalent among deletions.

Appendix A.1 also provides a domain-specific classification of QAs in terms of frequently asked questions about RCTs (August et al., 2023).

How similar are information needs across annotators? As discussed in Section 3.2, the INFOLOSSQA task is subjective. Enlisting two annotators per document allows us to quantitatively study information needs of different individuals.

First, we analyze to what extent annotators highlight the same evidence spans E . We calculate three measures: (1) **strict span-level F1**, where boundaries and type (deletion vs. oversimplification) have to match exactly, (2) **partial span-level F1** where two spans are considered a match if their token-level intersection over union exceeds a threshold (DeYoung et al., 2020a),⁸ and (3) **sentence-level Krippendorff’s α** where we project span-level annotations to a binary label indicating if a sentence has at least one span (Goyal et al., 2022). The sentence-level α shows that the annotators do tend to pick up information loss localized in the same evidence *sentence* to a moderate degree (Artstein and Poesio, 2008). However, the precise evidence span (in smaller linguistic units) is more subjective, as indicated by the lower span-level scores (Table 2).

⁸We tokenize on whitespace and use a threshold of $t = 0.5$.

Second, we analyze the similarity of questions as a function of whether their underlying evidence spans E have complete, partial, or no overlap. We use cosine similarity over question embeddings (Reimers and Gurevych, 2019). Generally, questions become more similar when the evidence overlaps (Figure 3b). However, subjectivity remains: Some questions are different although the evidence is the same. Similar observations were made for QUD generation (Ko et al., 2022).

4 Automatic Information Loss Detection

No system in literature is directly applicable to solve all parts of INFOLOSSQA. Therefore, we propose and evaluate two systems: (1) End-to-end prompting, contrasting several strong LLMs; (2) An NLI pipeline. We describe the methods below and provide detailed parameters in Table A.1.

Method 1: LLM prompting. We design a one-shot prompt that specifies the task, closely resembling the manual annotation guidelines. We ask all models to localize information loss instances, to provide a short motivation for why it is considered information loss, akin to Chain-of-Thought prompting (Wei et al., 2022), and finally to generate the QA pairs. We benchmark three LLMs: **Mistral-7B**, **Llama2-70B**, and **GPT-4**.⁹ All prompts were manually tuned for each model, though this was a heuristic process and better prompts may exist.

Method 2: NLI pipeline. To contrast with end-to-end LLMs, we implement a three-step pipeline: (i) fact extraction, (ii) entailment classification, (iii) QA generation for facts with a *neutral* relation. NLI assesses if a “hypothesis” is inferable from a “premise,” categorizing it as *entailed*, *contradictory*, or *neutral*. We treat the original text X_{src} as hypothesis, and the simplification X_{tgt} as premise. Intuitively, a neutral relation indicates information loss, where at least one piece of information in X_{src} cannot be inferred from X_{tgt} . In contrast, entailment indicates faithful conveyance of all information.

Information loss can be as granular as a word or phrase, and one sentence may contain multiple distinct instances. Therefore, we adopt NLI over fine-grained facts (Stacey et al., 2023). Formally, we decompose each sentence in X_{src} into atomic facts using an LLM (Stacey et al., 2023). Then, a standard NLI model (Laurer et al., 2024) classifies

⁹Specifically, we use gpt-4-0613, llama-2-70b-chat, and mistral-7b-instruct-v0.1.

entailment of each fact against X_{tgt} . Finally, we prompt an LLM to generate QA pairs given X_{src} , X_{tgt} , and the list of neutral facts. The sentence that generated the fact serves as evidence E .¹⁰

5 Evaluation Framework

We next propose a comprehensive framework to evaluate automatic systems for INFOLOSSQA. All evaluation is done manually by trained annotators. We describe the evaluation criteria below and Appendix C provides full annotator guidelines.

Aspect 1: Correctness of generated QA pairs. Our first criterion, **Answerability** (3-point), states that the generated *questions* (Q s) should be about an information loss. Valid questions must be answerable with the original text (i.e., there exists an evidence E answering Q), but unanswerable or only vaguely answerable with the simple text.

The second criterion, **Answer Accuracy**, states that questions should be correctly answered by the extracted and generated answers. We evaluate this in three parts. First, **Accuracy - Evidence** (3-point) captures how well the highlighted evidence E answers Q . Second, **Accuracy - Answer** (3-point) evaluates how well A answers Q . Finally, we ask if A contains any **Hallucinations** (binary).

Aspect 2: Linguistic suitability. We aim for QAs that are *pragmatically felicitous* and *comprehensible* for readers (cf. Section 2). We propose evaluating this through four linguistic criteria. First, we evaluate **Givenness** (binary), closely following the constraints of QUD (Riester et al., 2018; Wu et al., 2023): A question should only contain concepts that are directly mentioned in the question context, inferable, or generally known. We define the question context as the entire simple text X_{tgt} for QAs on Deletions, and everything up to and including the rationale span R for QAs on Oversimplifications.¹¹ Second, when a question seeks clarification about an oversimplified concept, there should be a rationale R in X_{tgt} as an opportunity to show users that the question addresses said vagueness (**Rationale**, 3-point). Third, we evaluate if A contains any unexplained **Jargon** (4-point). Finally, **Standalone** (binary) states that answers must be de-

¹⁰We do not implement localization R in the simple text, but note that interpretable NLI models could provide such a localization (Stacey et al., 2022; Rodriguez et al., 2023).

¹¹With this definition we intend to enable interactions where users can inspect oversimplification QAs *as they read* the text, e.g., visualized as expandable highlights (Fok et al., 2023).

contextualized (Choi et al., 2021), i.e., they should not contain unresolved coreferences or acronyms that would require reading the original text.

Aspect 3: Recall of human QA. Lastly, we evaluate if models identify similar information loss as the references in INFOLOSSQA. A challenge in this evaluation is that both humans and models produce a *set of QAs* without a strict 1-to-1 correspondence. We found that directly comparing two question sets is too cognitively demanding with typical sets of more than five QAs. Therefore, we opt for a pairwise setting, comparing each reference QA with one model QA at a time, and marking the reference as either *fully recalled*, *partially recalled* or *not recalled*. As comparing all possible pairs is intractable, we exclude unlikely matches following a paraphrase mining approach (Wright et al., 2022). We report the detailed procedure to estimate a minimum similarity threshold in Appendix A.3.

Evaluation process. We sample 100 QAs per model for aspect 1+2, and 50% of the documents for aspect 3. All samples were independently evaluated by two of the experts described in Section 3.2. Evaluation took 64 hours for aspect 1+2, and 20 hours for aspect 3, with additional training hours. We measure Randolph’s Kappa (Randolph, 2005) for inter-evaluator agreement and observe a substantial average agreement of $\kappa = 0.70$ (Table 5). Appendix A.3 provides details on sampling and exclusion of invalid LLM generations.

6 Results and Discussion

Overall we find that INFOLOSSQA is a challenging task. While all models have good question-generation/answering abilities, the main difficulty lies in reliably identifying information loss and ensuring the answers are comprehensible for readers. Furthermore, models identify different information loss than humans. Quantitative results are given in Tables 3 and 4 and qualitative examples in Figure 4.

Models generate valid questions but not all are about an information loss (Table 3). The majority of questions is answerable with X_{src} , showing that all models are good at document-grounded question generation. However, a sizable number of questions lack specificity as they have multiple valid answers (7.8%–19.6%). For example, “*What are the details of the experiments?*” (Ex. 1 in Figure 4). Furthermore, a small number of questions is

unanswerable with X_{src} . Qualitatively, these questions tend to ask for lay explanations of technical terms which is out of scope of information loss.

While most questions are valid, all models could improve at generating questions that are *not* or only *vaguely* answerable with X_{tgt} . See example 2 (Figure 4), where the generated QA does not add any new information. Notably, the fact-based NLI pipeline outperforms open-source LLMs in identifying unanswerable questions, indicating a promising direction for addressing information loss.

Considering answer accuracy, GPT-4 provides near perfect results both in extracting the correct evidence span (98% correct) and in rewriting it into a full answer (94% correct). Surprisingly, Mistral and Llama2 more often generate correct abstractive answers than extracting the right evidence, implying internally inconsistent LLM responses. Lastly, across all models, we only observe a small rate of hallucinations (worst: Llama2, 7.8%).

Models often fail to localize QAs in the simple text (Table 4). Too often models do not produce a rationale span in X_{tgt} , which shows a general limitation of models to correctly discern between deletion and oversimplification (42.2%–82.4% missing rationale). For reference: 25% of human QAs are deletions, whereas the models classify more than 50% of QAs as deletions (Table A.2). Annotators have substantial agreement for when there should be a rationale span ($\kappa = 0.80$, Table 5), so this could be alleviated by model fine-tuning.

QAs are well-framed, with room for improvement in terms of jargon (Table 4). Considering the Givenness constraint, the end-to-end models provide well-framed questions that only contain concepts familiar to readers and do not have any answer leakage. This result is on par or slightly better than recent QUD parsers (Wu et al., 2023).

Concerning the answers, most can be understood without referring to the original text, i.e., they are Standalone. Qualitatively, undefined abbreviations are frequent reasons for answers to not be standalone (Ex. 3 in Figure 4) which could be solved with a dedicated decontextualization step (Choi et al., 2021). Lastly, many answers contains unexplained Jargon (--, best NLI 28.4%, worst Mistral 55.9%). Despite instructing all models to generate simple answers, the task likely has too many constraints for a single end-to-end prompt. By decoupling information loss detection and QA generation, the NLI pipeline can better follow this constraint.

Model	Q is Answerable w/ X_{src}			Q is Answerable w/ X_{tgt}			Accuracy - Evidence (E)			Accuracy - Answer (A)		
	Yes	Multi. [↓]	No [↓]	No	Vague	Yes [↓]	Correct	Part. [↓]	Wrong [↓]	Correct	Part. [↓]	Wrong [↓]
Mistral-7B	72.5	19.6	7.8	32.4	30.4	37.3	74.5	12.7	12.7	84.8	12.3	2.9
Llama2-70B	83.0	9.7	7.3	21.4	32.5	46.1	77.7	11.7	10.7	85.9	10.7	3.4
GPT-4	87.5	12.5	0.0	58.5	33.0	8.5	98.0	2.0	0.0	94.0	6.0	0.0
NLI Pipeline	89.7	7.8	2.5	49.5	25.0	25.5	77.5	4.4	18.1	77.9	20.6	1.5

Table 3: Correctness of generated QA pairs as percent of ratings given by two independent annotators over 100 QA pairs per model. Multi. = Multiple valid answers; Part. = Partially complete answer; [↓] = lower is better.

Model	Given. (Q)	Rationale (R)			Jargon (A)				Stand. (A)	Hallu. (A)	Recall		
	% Good	Good	Unrel. [↓]	Miss. [↓]	(++)	(+)	(-) [↓]	(--) [↓]	% Good	% Good	Full	Partial	No
Mistral-7B	74.5	52.0	5.9	42.2	21.6	9.3	13.2	55.9	84.3	96.1	14.2	4.1	81.8
Llama2-70B	75.2	34.5	5.3	60.2	26.2	7.8	17.5	48.5	79.6	92.2	16.2	4.4	79.3
GPT-4	82.0	52.0	1.5	46.5	15.0	24.0	11.0	50.0	90.5	99.0	28.1	3.4	68.4
NLI Pipeline	58.3	17.6	0.0	82.4	28.9	29.9	12.7	28.4	86.3	98.0	22.9	20.0	57.2

Table 4: Linguistic suitability of generated QA pairs and recall of human QAs. Given. = Givenness; Stand. = Standalone; Hallu. = Hallucinations; Unrel. = Unrelated; Miss = Missing. [↓] = lower is better.

Models have a low recall of human QAs (Table 4). The NLI pipeline is most effective in this criterion (42.9% of QAs at least partially recalled), but also generates substantially more QAs per document than humans (16.5 vs. 9.6, Table A.2). Qualitatively, the generated questions each cover a smaller part of missing content, whereas humans aggregate related missing information into one larger question (see Figure A.4 for an example). This is corroborated by the NLI pipeline having the highest *partial* recall of human QAs. Exciting future directions will be to (1) get models to apply similar standards as humans at what constitutes an information loss, and (2), to study how to aggregate smaller-scope questions into broader questions.

Automatic evaluation: prompting LLMs to evaluate linguistic suitability is difficult (Table 5). The previous results rely on manual evaluation by trained annotators, which is expensive and not scalable. Therefore, we run a preliminary investigation of automatic LLM-based evaluation. For each of the 10 evaluation criteria, we created a prompt with evaluation instructions, the generated QA pair, and additional context for assessment. The model outputs a brief rationale followed by the rating, akin to chain-of-thought (Wei et al., 2022). No further prompt engineering was done. Agreement between human ratings and ratings by representative LLMs (GPT-4o and Llama3-70B)¹² are given in Table 5.

On average, Llama3-70B and GPT-4o agree with humans evaluators to a *fair* and *moderate* de-

¹²Specifically, we use the checkpoint gpt-4o-2024-05-13 and meta-llama/Meta-Llama-3-70B-Instruct. Full evaluation suite is provided in code release.

Criterion	Human	GPT-4o	Llama3
<i>Aspect 1: Correctness of QA</i>			
Q is Answerable w/ X_{src}	0.61	0.68	0.66
Q is Answerable w/ X_{tgt}	0.50	0.33	0.24
Accuracy - Evidence (E)	0.71	0.65	0.52
Accuracy - Answer (A)	0.68	0.63	0.57
Hallucinations (A)	0.89	0.81	0.72
<i>Aspect 2: Linguistic Suitability</i>			
Givenness (Q)	0.79	0.30	0.24
Rationale Localization (R)	0.80	0.18	0.06
Jargon (A)	0.57	0.34	0.43
Standalone (A)	0.69	0.21	-0.30
<i>Aspect 3: Recall of Human QA</i>			
Recall	0.70	0.67	0.47
Average	0.70	0.48	0.36

Table 5: Inter-evaluator agreement as measured by Randolph’s Kappa (Randolph, 2005). Agreement of models with humans is the average of pairwise model-human agreements.

gree (Artstein and Poesio, 2008). GPT-4o matches human agreement on criteria related to the correctness of QAs and recall of human QA, suggesting it could substitute human evaluation. In contrast, LLM-based evaluation of the linguistic suitability of generated QAs appears unreliable and likely requires a more sophisticated method to achieve good agreement with human rating, corroborating results on general QUD evaluation (Wu et al., 2023).

7 Related Work

Deletion in text simplification. Professional editors commonly use deletion to make text more accessible (Petersen and Ostendorf, 2007; Xu et al., 2015; Zhong et al., 2020; Yamaguchi et al., 2023). Devaraj et al. (2022) found that even professional

Example 1: Vague question with multiple valid ways of answering; The answer is not standalone because acronym “VS” is only defined in X_{src}.								
X_{src} : ... Two experiments were conducted. In Experiment 1, COCs (n = 420) were randomly assigned to four groups: ... In Experiment 2, COCs (n = 581) were assigned to the same groups, but those in VS1, VS1 + VS2 and Vitrified groups were sub-divided and exposed to the warming solution for either 1 or 5 min. ... In Experiment 2, there was no effect of time in the warming solution...								
X_{tgt} : ... We did two experiments. In the first experiment, we had 420 [cow egg cells (also called COCs)] separated into four groups: ... In the second experiment, we treated 581 COCs the same way but also had some of the COCs from groups 2, 3 and 4 spend either 1 or 5 minutes in a warming solution...								
Q : What were the details of Experiment 2? A : In Experiment 2, 581 COCs were assigned to the same groups as in Experiment 1, but those in VS1, VS1 + VS2 and Vitrified groups were sub-divided and exposed to the warming solution for either 1 or 5 minutes.								
Llama-2	~ Answerable with X_{src} (Multi.)	✓ Vaguely Answerable with X_{tgt}	~ E is partial	~ A is partial	✓ Givenness	~ R missing	✗ Jargon (-)	✗ Standalone
Example 2: Not an information loss: A does not add any information to X_{tgt}. Givenness: “HbA1c” is not known from the simple text or general knowledge.								
X_{src} : One-hundred participants with a mean±s.d. age of 55.6±10.6 year, body weight of 102.9±18.4 kg and HbA1c of 7.7±1.3% were randomly ...								
X_{tgt} : We had 100 participants on average about 55 years old, weighing about 102.9 kilos, with a blood sugar level of 7.7%. They were placed randomly ...								
Q : What were the ages, weights, and HbA1c levels of the participants in the study?								
A : The participants had a mean age of 55.6 years, a mean weight of 102.9 kg, and a mean HbA1c level of 7.7%.								
Llama-2	✓ Answerable with X_{src}	✗ Answerable with X_{tgt}	✓ E is correct	✓ A is correct	✗ Givenness	~ R missing	✗ Jargon (-)	✓ Standalone
Example 3: Both evidence and answer are incomplete because they miss the loading dose of one group.								
X_{src} : ... Group L received a loading dose of 0.5 $\mu\text{g}/\text{kg}$ for 10 minutes while group H received 1.0 $\mu\text{g}/\text{kg}$...								
X_{tgt} : ... We divided these patients into two groups. One received a lower dosage of the drug, while the other received a higher dosage.								
Q : What was the specific loading dose used in the study? A : The patients in group H received a loading dose of 1.0 $\mu\text{g}/\text{kg}$.								
Mistral	~ Answerable with X_{src} (Multi.)	✓ Not Answerable with X_{tgt}	~ E is partial	~ A is partial	✗ Givenness	✓ R is related	✗ Jargon (-)	✗ Standalone

Figure 4: Qualitative examples demonstrating error cases. More examples in Figure A.1.

simplifications include *over*-deletions, and that models are prone to mimic this behavior. In user studies, Agrawal and Carpuat (2024) found that deletions are a major factor for diminished reading comprehension. These studies highlight the importance of detecting and mitigating deletions. We contribute to this area by providing the first annotated dataset of information loss and a QA-based framework for addressing it.

Question generation (QG). While early work considered QG with factoid answers, the focus shifted to more natural, information-seeking and inquisitive questions (Kwiatkowski et al., 2019; Ko et al., 2020; Scialom and Staiano, 2020; Dasigi et al., 2021; Meng et al., 2023). We consider questions with a similar pragmatic goal as clarification questions (Rao and Daumé III, 2018; Majumder et al., 2021) and gap-focused questions in dialogue (Rabin et al., 2023): Asking about information which is missing or vague in a context. Newman et al. (2023) demonstrated the merits of QG to represent missing information for decontextualization. But we draw attention to the particularity of QG in text simplification. Compared to experts which have expectations of what information texts typically include, lay readers have difficulty asking these clarification questions due to their lacking “disciplinary knowledge” (August et al., 2023), i.e., they have *unknown unknowns*.

Closely related is DIFFQG (Cole et al., 2023) which uses QG to describe *factual* changes in two revisions of a Wikipedia passage. In contrast, we consider simplification-induced changes where answers to questions are not necessarily different but vague. Also, our task requires document-level com-

parisons, simple factoid to complex multi-sentence answers, and tailoring the QA to laypeople.

QA for evaluation and entailment. QA has been used to evaluate summarization and simplification (Mani et al., 2002; Agrawal and Carpuat, 2024). The advances in QG gave rise to the cross-questioning paradigm (Wang et al., 2020; Durmus et al., 2020; Deutsch et al., 2021, *inter alia*). This line of work differs from INFOLOSSQA in two important aspects. First, there is a strong focus on noun-phrase and entity-centric QA. Our proposed task invites QAs that are meant for consumption by end-users, and hence must satisfy additional linguistic criteria (Section 2). Second, while its conceivable to use cross-questioning to get candidate QAs, Kamoi et al. (2023) identified error-propagation in the QG stage as a fundamental limitation making QA-level answerability unreliable. They therefore advocate for NLI-based approaches (Laban et al., 2022; Rodriguez et al., 2023), which our pipeline-system is inspired by.

8 Conclusion

We propose INFOLOSSQA: A task and dataset to describe and recover simplification-induced information loss as reader-centric QAs. Our analyses show that QAs following the QUD theory provide a rich description that can mitigate overdeletion and vagueness in text simplification. We also establish automatic pipelines for the task and propose a rigorous evaluation framework considering correctness of QAs and user-centric constraints. Looking ahead, this work opens new avenues in interactive simplification tools (Fok et al., 2023) and for quality assessment of automatic simplifications.

Limitations

Our motivation and vision for INFOLOSSQA is to help users to deepen their understanding of a text. However, we focus on the technical dimensions of this goal: establishing a dataset, an evaluation framework, and developing and evaluating baselines. Testing the effects of the proposed QA on *end-user* comprehension is an important direction for future work in interactive text simplification.

Furthermore, while we believe that the proposed approach is generalizable, our experiments are confined to one language (English), text genre (abstracts of medical publications) and simplification style (GPT-4 simplification). Future work could assess the versatility of this framework under different conditions. For instance, we considered simplifications at one level of compression, but readers may prefer different degrees of simplification (Xu et al., 2015; August et al., 2024). As we alter the degree of simplification, it becomes important to understand (a) how effective models are at identifying information loss, and (b) how the distribution of information loss and associated QAs changes.

Lastly, our proposed evaluation framework currently relies on human annotators to judge the quality of model outputs. Any kind of human evaluation comes at a significant cost which may limit further studies on better modeling. Therefore, future work could develop automatic metrics for the task. Our initial experiments show that automatic LLM-based evaluation is a promising method, but requires additional work to achieve good correlations with human judgments. To facilitate this direction, we release the full evaluation suite including human judgments and evaluation baselines.

Acknowledgments

We thank Keziah Kaylyn Reina, Kathryn Kazanas and Karim Villaescusa F. for their annotation and evaluation effort, David Heinemann for help with the annotation interface, and Ritvik Renikunta for help with the initial design of the end-to-end prompts. We also thank Eunsol Choi for feedback on this paper, and Juan Diego Rodriguez and Manya Wadhwa for useful discussions. This research is partially supported by NSF CAREER Award IIS-2145479 and Good Systems,¹³ a UT Austin Grand Challenge to develop responsible AI technologies. Trienes was supported by the Cancer

Research Center Cologne Essen (CCCE), the Federal Ministry of Education and Research (BMBF) and by a fellowship within the IFI programme of the German Academic Exchange Service (DAAD). Wallace was supported in this work by the National Institutes of Health (NIH), grant R01LM012086, and by the National Science Foundation (NSF), grant 1750978. Xu is supported in part by NSF awards IIS-2144493 and IIS-2112633.

References

- Sweta Agrawal and Marine Carpuat. 2024. [Do text simplification systems preserve meaning? A human evaluation via reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Ron Artstein and Massimo Poesio. 2008. [Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. [Know your audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience](#). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. [Paper Plain: Making medical research papers approachable to healthcare consumers with natural language processing](#). *ACM Transactions on Computer-Human Interaction*, 30(5):1–38.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O’Reilly Media, Inc.
- Shuyang Cao and Lu Wang. 2021. [Controllable open-ended question generation with a new question type ontology](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6424–6439.
- Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. [Decontextualization: Making sentences stand-alone](#). *Transactions of the Association for Computational Linguistics*, 9:447–461.
- Jeremy R. Cole, Palak Jain, Julian Martin Eisenschlos, Michael J.Q. Zhang, Eunsol Choi, and Bhuwan Dhingra. 2023. [DiffQG: Generating questions to summarize factual changes](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3088–3101.

¹³<https://goodsystems.utexas.edu>

- 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 (NAACL)*, pages 4599–4610.
- Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. [Towards question-answering as an automatic metric for evaluating the content quality of a summary](#). *Transactions of the Association for Computational Linguistics*, 9:774–789.
- Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. [Evaluating factuality in text simplification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7331–7345.
- 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 (ACL)*, pages 4443–4458.
- 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.
- Esin Durmus, He He, and Mona Diab. 2020. [FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5055–5070.
- Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2023. [Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries](#). *CoRR*, abs/2310.07581.
- Brent Goff, Vanessa Patino, and Gary Jackson. 2004. Preferred information sources of high school students for community colleges and universities. *Community College Journal of Research & Practice*, 28(10):795–803.
- Sian Gooding. 2022. [On the ethical considerations of text simplification](#). In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. [SNaC: Coherence error detection for narrative summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–463.
- David Heineman, Yao Dou, and Wei Xu. 2023. [Thresh: A unified, customizable and deployable platform for fine-grained text evaluation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 336–345.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Ryo Kamoi, Tanya Goyal, and Greg Durrett. 2023. [Shortcomings of question answering based factuality frameworks for error localization](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 132–146.
- Andrew Kehler and Hannah Rohde. 2017. [Evaluating an expectation-driven question-under-discussion model of discourse interpretation](#). *Discourse Processes*, 54(3):219–238.
- Wei-Jen Ko, Te-yuan Chen, Yiyang Huang, Greg Durrett, and Junyi Jessy Li. 2020. [Inquisitive question generation for high level text comprehension](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555.
- Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. [Discourse comprehension: A question answering framework to represent sentence connections](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11752–11764.
- 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*, 7:452–466.
- 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.
- Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. [SWIPE: A dataset for document-level simplification of Wikipedia pages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10674–10695.
- Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. [Less annotating, more](#)

- classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.
- Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 3705–3717.
- Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).
- Annie Louis and Ani Nenkova. 2011. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42.
- Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what’s missing and what’s useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4300–4312.
- Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 2002. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.
- Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 795–804.
- Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FollowupQG: Towards information-seeking follow-up question generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, pages 252–271.
- Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3194–3212.
- Edgar Onea. 2016. *Potential questions at the semantics-pragmatics interface*, volume 33. Brill.
- OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.
- Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *Proc. Speech and Language Technology in Education (SLaTE 2007)*, pages 69–72.
- Roni Rabin, Alexandre Djerbetian, Roeel Engelberg, Lidan Hackmon, Gal Elidan, Reut Tsarfaty, and Amir Globerson. 2023. Covering uncommon ground: Gap-focused question generation for answer assessment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 215–227.
- Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss’ fixed-marginal multirater kappa.
- Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2737–2746.
- 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for questions under discussion and information structure. *Information structure in lesser-described languages: Studies in prosody and syntax*, pages 403–443.
- Juan Diego Rodriguez, Katrin Erk, and Greg Durrett. 2023. X-PARADE: Cross-lingual textual entailment and information divergence across paragraphs. *CoRR*, abs/2309.08873.
- David L. Sackett. 1998. Evidence-based medicine. *Spine*, 23(10):1085–1086.
- Thomas Scialom and Jacopo Staiano. 2020. Ask to learn: A study on curiosity-driven question generation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 2224–2235.
- Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron C. Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1387–1407.
- Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2023. Logical reasoning for natural language inference using generated facts as atoms. *CoRR*, abs/2305.13214.

- Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. [Logical reasoning with span-level predictions for interpretable and robust NLI models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3809–3823.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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](#). *CoRR*, abs/2307.09288.
- UNSDG. 2021. Access to information is the cure of disinformation. unsdg.un.org/latest/blog/access-information-cure-disinformation. Accessed: 2024-02-13.
- Kees van Deemter. 2012. *Not Exactly: In Praise of Vagueness*. Oxford University Press.
- 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 (ACL)*, pages 5008–5020.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.
- Lysandre Wolf, Thomas 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 (EMNLP): System Demonstrations*, pages 38–45.
- Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. [Modeling information change in science communication with semantically matched paraphrases](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1783–1807.
- Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. [QUDeval: The evaluation of questions under discussion discourse parsing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5344–5363.
- Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. [Problems in current text simplification research: New data can help](#). *Transactions of the Association for Computational Linguistics*, 3:283–297.
- Daichi Yamaguchi, Rei Miyata, Sayuka Shimada, and Satoshi Sato. 2023. [Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 359–375.
- Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. [Discourse level factors for sentence deletion in text simplification](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):9709–9716.
- Alesia Zuccala. 2010. Open access and civic scientific information literacy. *Information Research: An International Electronic Journal*, 15(1).

A Appendix

A.1 Analysis: Relation to Paper Plain Key Question Index

In addition, to the domain-agnostic question taxonomy (Cao and Wang, 2021) used in Section 3.3, we analyzed the QA pairs through the lens of the Paper Plain Key Question Index which was designed to convey the most important elements of an RCT (August et al., 2023). We manually codify 120 QA pairs (40 per annotator) according to the 8 categories of the question index.

Aligned with findings in Section 3.3, the information which is most likely lost is about *methodological detail* (i.e., study protocol, analysis tools, population; 53% of QA pairs), followed by *results* (18% of QA pairs) and the *goals* of the trial (11% of QA pairs). Critical information like the usual and new treatments is preserved. Most notably, we rarely observe outright omission of *all* information regarding a top-level question in the key question index. Instead, the INFOLOSSQA questions cover information with a higher level of specificity. See Table A.4 for a detailed breakdown and examples.

A.2 Analysis: How are Errors Distributed Across Document Sections?

Given that some RCT sections are more technical than others, it is conceivable that models are better generating information loss QAs in some sections than in others. We plot the percentage of good responses per model, section and evaluation criterion in [Figure A.2](#). Contrary to our expectation, there are no notable differences across sections, with only a slight trend for reduced Givenness and Jargon for QAs localized in the results and conclusion section. Qualitatively, this is often due to questions asking about the statistical significance of the results, which annotators deemed to be an unfamiliar concept for lay readers.

A.3 Experiment Detail: Manual Evaluation

Sampling. For aspect 1+2, we take a stratified sample to preserve the relative frequencies at which models generate QAs per RCT section. As the generations by Mistral-7B and Llama2-70B do not always follow the specified output format, we only sample from QA pairs which could be completely parsed. See [Table A.3](#) for a detailed analysis of generation errors.

Recall evaluation: Finding candidate matches.

For our dataset of 1,000 reference QAs across 104 documents and predictions by four models, the pairwise recall evaluation described in [Section 5](#) results in 33,825 comparisons which is intractable. Therefore, we follow the paraphrase mining approach by [Wright et al. \(2022\)](#) to get *candidate matches*. First, we calculate the cosine similarity of two QA pairs using sentence embeddings ([Reimers and Gurevych, 2019](#)). Both the question and the answer are concatenated before calculating the embedding. Afterwards, we establish a threshold by annotating 400 matches (predicted QA and reference QA) equally sampled from 20 bins in the similarity range of $[0, 1]$. We observe the first bin with a notable number of recalled QA (here: $T = 0.65$). Pairs with similarity $\leq T$ are set to *not recalled*. This process reduced the required manual evaluations by 93% to 2,466.

We aggregate votes as follows. A reference QA is set to *fully recalled* if at least one predicted QA fully recalls it. If a reference QA is not fully recalled but partially recalled by at least one predicted QA, it is set to *partially recalled*. Otherwise it is set to *not recalled*.

Description	Prompt	Decoding Parameters
<i>Simplification.</i> Simplifying technical RCT abstracts (X_{src}) into a plain language version (X_{tgt}). See top part of Figure A.4 for an example simplification.	Figure A.5	model=gpt-4-0613 temperature=1.0 max_tokens=1024 top_p=1 frequency_penalty=0 presence_penalty=0
<i>Few-shot question classifier.</i> Categorizing questions according to the typology of Cao and Wang (2021) . This classifier obtains an F1 score of 0.87 on a manually labeled validation set of 50 questions in INFOLOSSQA.	Figure A.6	model=gpt-4-0613 temperature=0 max_tokens=1024 top_p=1 frequency_penalty=0 presence_penalty=0
NLI pipeline for information loss detection.		
<i>Part 1: fact extraction.</i> Extracting atomic facts from X_{src} . These facts are classified for entailment with X_{tgt} with a standard NLI model (Laurer et al., 2024).	Figure A.7	model=gpt-4-0613 temperature=0 max_tokens=512 top_p=1 frequency_penalty=0 presence_penalty=0
<i>Part 2: QA-generation.</i> Generating QA pairs based on X_{src} , X_{tgt} and the list of facts with <i>neutral</i> outcome of the NLI classifier.	Figure A.8	model=gpt-4-0613 temperature=1 max_tokens=4096 top_p=1 frequency_penalty=0 presence_penalty=0
End-to-end prompts for information loss detection. All prompts include a one-shot example, which we found to substantially improve performance over a zero-shot prompt. For a fair comparison across models, we do not include more than one example as it would exhaust the context windows of some models under investigation.		
<i>GPT-4 (OpenAI, 2023).</i> End-to-end prompt to detect information loss and generate QA pairs. Inference on API of OpenAI.	Figure A.9	model=gpt-4-0613 temperature=0 max_tokens=2048 top_p=1 frequency_penalty=0 presence_penalty=0
<i>Mistral-7B (Jiang et al., 2023).</i> End-to-end prompt to detect information loss and generate QA pairs. Inference with the Huggingface transformers library (Wolf et al., 2020) on one NVIDIA RTX A6000 (48GB) completed in less than one hour.	Figure A.10	model=Mistral-7B-Instruct-v0.1 do_sample=False temperature=0 max_tokens=2048 top_p=1 top_k=1 repetition_penalty=1
<i>Llama2-70B (Touvron et al., 2023).</i> End-to-end prompt to detect information loss and generate QA pairs. Inference on API of Together.AI .	Figure A.11	model=llama-2-70b-chat max_tokens=None temperature=0 top_p=1 top_k=1 repetition_penalty=1

Table A.1: LLM prompts and decoding parameters.

Annotator / Model	Docs.	QA	% Over	% Del	QA/doc	Q	A	E	R
Annotator 1	48	192	90.1	9.9	4.1	10.8	23.8	12.2	7.6
Annotator 2	75	425	71.8	28.2	5.7	12.2	33.4	15.5	9.0
Annotator 3	85	383	68.7	31.3	4.7	10.9	30.2	14.4	8.6
All (micro avg.)	104	1000	74.1	25.9	9.6	11.4	30.3	14.4	8.5
Mistral-7B	104	507	45.2	54.8	4.9	11.6	26.2	22.3	18.1
Llama2-70B	104	681	38.2	61.8	6.7	12.9	30.6	20.4	15.7
GPT-4	104	477	48.4	51.6	4.6	14.4	33.7	25.4	20.6
NLI Pipeline	104	1699	-	100.0	16.5	14.5	24.6	34.7	-

Table A.2: Summary statistics of human-written QAs and model predictions. Over = Oversimplification; Del = Deletion; Length of question Q , answer A , evidence spans $E \in X_{src}$ and rationale spans $R \in X_{tgt}$ is given in tokens.

Error	Mistral-7B	Llama2-70B	GPT-4	NLI Pipeline
Deletion (Total)	278	421	246	1699
% Valid (no error)	61.9	87.2	100	100
% Spurious R	38.1	1.0	-	-
% Invalid E	-	11.9	-	-
Oversimplification (Total)	229	260	231	0
% Valid (no error)	67.2	78.8	100	-
% Invalid R	28.8	5.8	-	-
% Missing R	2.2	6.2	-	-
% Invalid E	1.7	2.3	-	-
% Invalid E + Invalid R	-	5.8	-	-
% Invalid E + Missing R	-	1.2	-	-

Table A.3: Analysis of LLM generation errors grouped by deletion and oversimplification. For each category, the total number of QA pairs is given with the percent of QAs per error. Spurious R : QAs classified as deletions should not have a rationale span. Invalid E/R : the span cited by the model is not a valid substring of X_{src}/X_{tgt} .

Example 4: Good generation.

X_{src} : [The aim of this study is] to compare the efficacy of high and low oral doses of haloperidol in the treatment of acute and transient psychotic disorders...

X_{tgt} : This study is about testing whether high or low doses of a drug called haloperidol are more effective at treating short-lasting mental illnesses...

Q : What specific type of mental illness were the patients suffering from? A : The patients were suffering from acute and transient psychotic disorders, which are characterized by sudden onset of psychotic symptoms that last for a short period of time.

GPT-4 ✓ Answerable with X_{src} ✓ Vaguely Answerable with X_{tgt} ✓ E is correct ✓ A is correct ✓ Givenness ✓ R is related ✓ Jargon (+) ✓ Standalone

Example 5: Missing rationale R : annotator pointed out that “patients” is an oversimplification of “patients with aBTC;” so it should be highlighted.

X_{src} : Since the prognosis of advanced biliary tract cancer (aBTC) still remains very poor... Patients with aBTC refractory to at least one regimen of chemotherapies were randomly assigned to receive PPV with low dose CPA ... (PPV/CPA, n = 24) or PPV alone (n = 25).

X_{tgt} : Biliary tract cancer (aBTC), a type of liver cancer, has a very poor survival rate. To help improve this, we’re researching new treatments ... [W]e gave it to 24 patients and compared them to a group of 25 patients who received only the PPV.

Q : Who were the participants in the study?

A : The participants were patients with advanced biliary tract cancer (aBTC) who had failed to respond to at least one regimen of chemotherapies.

Mistral ✓ Answerable with X_{src} ✓ Not Answerable with X_{tgt} ✓ E is correct ✓ A is correct ✓ Givenness ~ R missing ✓ Jargon (++) ✓ Standalone

Example 6: Even though the QA correctly reveals missing information, the highlighted evidence E does not support the answer.

X_{src} : This study investigated the effect of combining oral dexamethasone with either nebulized racemic epinephrine or salbutamol compared to bronchodilators alone for the treatment of infants with bronchiolitis. ... This was a double-blind, randomized controlled trial on infants (1 to 12 months) who were diagnosed in the emergency department with moderate-to-severe bronchiolitis...

X_{tgt} : This research looked at how well infants with a common lung infection called bronchiolitis responded to different combinations of medicines... The study investigated the effects on babies between 1 to 12 months old...

Q : Where were the infants diagnosed with bronchiolitis? A : The infants were diagnosed in the emergency department of a hospital.

NLI Pipeline ✓ Answerable with X_{src} ✓ Not Answerable with X_{tgt} ✗ E is wrong ✓ A is correct ✓ Givenness ✓ Deletion = no R ✓ Jargon (++) ✓ Standalone

Figure A.1: Qualitative examples demonstrating error cases. Continued from Figure 4.

QA Category	%	Definition and Examples
Motivation [†]	5.0%	Why was the study conducted? Q: What was the motivation for this study to investigate the dosage of haloperidol to address nausea and vomiting after surgery? A: There is evidence that a small dose of haloperidol can help prevent...
Condition	5.8%	What condition does this paper study?
– Inclusion Criteria [†]	3.3%	<i>What were the specific inclusion criteria for participants?</i> Q: What criteria did the researchers use to select eligible participants? A: The researchers gathered participants between the ages of 3-18, ...
– Other [†]	2.5%	<i>Other condition-related question.</i> Q: What kind of hip surgery did the participants receive? A: The participants were getting hip surgery under the subarachnoid block...
Goal	10.8%	What did the paper want to find out? Q: What about the nicotine vaccine could potentially help smokers quit? A: The nicotine vaccine boosts antibody concentrations, which helps to stimulate...
Usual Treatment	1.7%	How is the condition usually treated? Q: What type of active deep brain stimulation was used in previous trials? A: Previous trials found a positive impact of unilateral (only applied to one side of the brain) active deep brain stimulation on symptoms of Parkinson's disease.
New Treatment	4.2%	What were the new treatment(s), if any this paper looked into? Q: What type of formoterol is being observed in the study? A: Long-acting beta2-agonist formoterol, which is a drug used to treat asthma...
Method	53.3%	What did the paper do?
– Outcome (Tool) [†]	14.2%	<i>What tools/procedures were used to measure the effects of interventions?</i> Q: How was sleep quality and life quality measured for participants? A: Sleep quality and life quality were assessed using the Sleep-Apnoea-Quality-of-Life-Index...
– Study Protocol [†]	10.8%	<i>What was the protocol of the study?</i> Q: What was the setting and design of the study? A: The study was an open (both participants and researchers knew who was assigned which drops), cross-over (the participants receive both treatments ...), comparative study (comparing the two drops).
– Quantity [†]	7.5%	<i>With what dosage/quantity/frequency were the interventions performed?</i> Q: How much lurasidone was given to the patients in the trial? A: Patients were given 40 to 80 milligrams of lurasidone per day, given flexibly.
– Population [†]	7.5%	<i>What were the demographics of the patients in the study?</i> Q: What was the age range of the patients? A: 52-85 years.
– Analysis Technique [†]	5.8%	<i>How were the results analyzed?</i> Q: How were the results analyzed? A: The results of the test were analyzed using a t-test and...
– Other [†]	7.5%	<i>Other method-related question.</i> Q: How often were participants checked during the study? A: Participants were assessed on their first visit, on the 7th day or their first follow-up, and on the 14th day or their second follow-up.
Results	18.3%	What did the paper find? Q: How did the number of events in the Tamoxifen group compare to the other group? A: The Tamoxifen group had a total of 39 events, while the other...
Contrasting Results	0.8%	Are the findings different depending on a person's demographics? Q: How did selenium levels of pregnant women in the UK vary from those of pregnant women in other countries? A: The selenium (mineral) levels of pregnant women in the UK were lower compared to those of pregnant women in other countries...

Table A.4: Classification of 120 randomly sampled human QA pairs in INFOLOSSQA according to the pragmatic goals of questions in the Paper Plain Key Question Index developed for lay readers of RCTs (August et al., 2023). Frequency of questions is given in percent (%). (Sub-)Categories marked with [†] were added by us to give a more fine-grained view on the questions types.

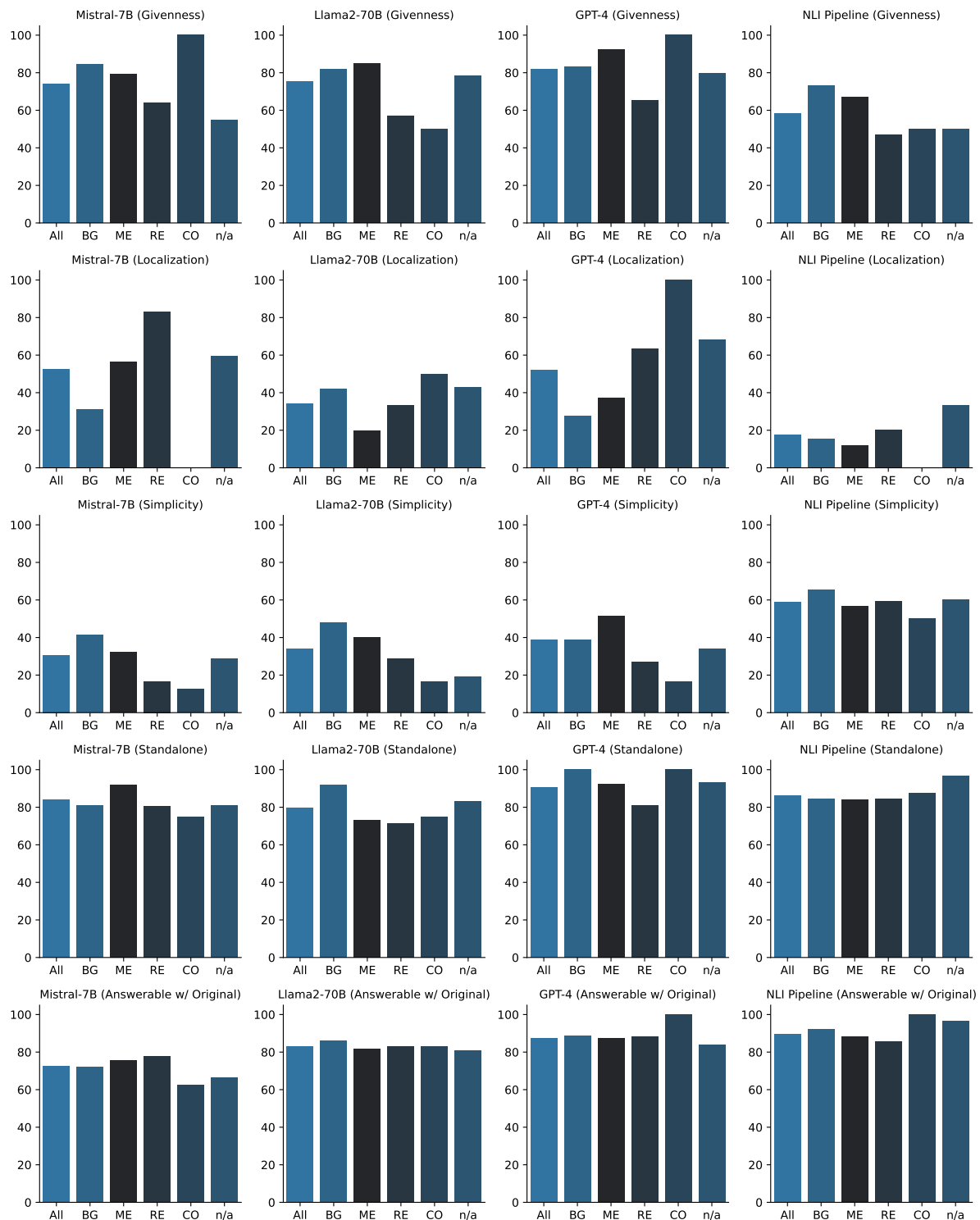


Figure A.2: Percentage of QAs with a good rating grouped by model, evaluation criterion and RCT section (BG = Background; ME = Methods; RE = Results; CO = Conclusion; n/a for RCTs without sections). Criteria with multiple response options were binarized to “good” vs. “not good.” Continued in Figure A.3

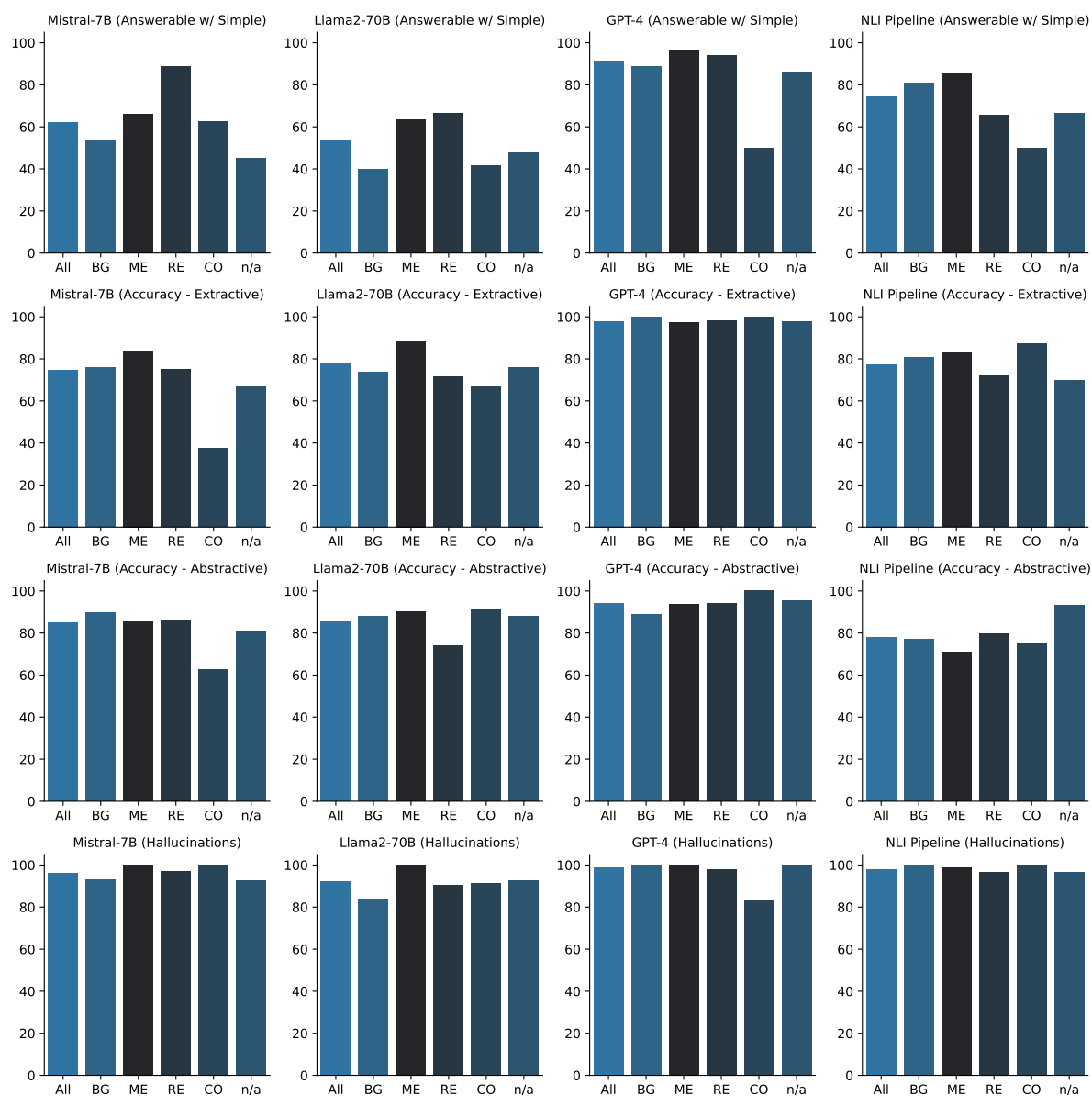


Figure A.3: Percentage of QAs with a good rating grouped by model, evaluation criterion and RCT section (BG = Background; ME = Methods; RE = Results; CO = Conclusion; n/a for RCTs without sections). Criteria with multiple response options were binarized to “good” vs. “not good.” Continued from Figure A.2

Original	Simplification
<p>BACKGROUND. The present study examined the effects of a balance-based exergaming intervention using the Kinect sensor on postural stability and balance in people with Parkinson's disease (PD).</p> <p>METHODS. We conducted a subject-blinded, randomized controlled study. Twenty people with PD (Hoehn and Yahr stages I through III) were recruited and randomly assigned to either a balance-based exergaming group (N = 10) or a balance training group (N = 10) for an 8-week balance training period. Postural stability was assessed using the limits of stability (LOS) and one-leg stance (OLS) tests. Balance was assessed using the Berg Balance Scale (BBS) and the timed up and go (TUG) test. Participants were assessed pre- and post-training.</p> <p>RESULTS. After training, participants in the balance-based exergaming group showed significant improvements in LOS performance, and in the eyes-closed condition of the OLS test. Both training programs led to improvements in BBS and TUG performance. Furthermore, balance-based exergaming training resulted in significantly better performance in directional control in the LOS test ($78.9 \pm 7.65\%$) compared with conventional balance training ($70.6 \pm 9.37\%$).</p> <p>CONCLUSIONS. Balance-based exergaming training resulted in a greater improvement in postural stability compared with conventional balance training. Our results support the therapeutic use of exergaming aided by the Kinect sensor in people with PD.</p>	<p>BACKGROUND. We did a study to see if a special type of exercise game, called exergaming, that uses the Kinect sensor could help improve balance and stability in people with Parkinson's disease.</p> <p>METHODS. We performed a study where the participants didn't know what group they were in, and it was all decided randomly. We had twenty people with Parkinson's disease take part. These people were either at the early, middle or later stages of the disease. Half of them were put into a group that did the Kinect exergame for balance, and the other half were put into a group that did traditional balance training. This went on for 8 weeks. We used certain tests (LOS, OLS, BBS and TUG) to measure their balance and stability before and after the training.</p> <p>RESULTS. The group that did the Kinect exergame had better results in some stability tests. Both groups improved in the BBS and TUG balance tests. In one of the LOS tests, the exergaming group did much better than the traditional training group.</p> <p>CONCLUSIONS. The group that did the Kinect exergame improved more in their stability than the group that did traditional training. Our study suggests that these kind of exercise games can be helpful for people with Parkinson's disease.</p>

NLI Pipeline	Humans	GPT-4
Cluster 1: scope of the study		
What was the main focus of the study? The main focus of the study was to understand if a type of exercise game called exergaming improved balance and stability in people with Parkinson's disease.	What type of exergaming was studied? This study examined balance-based exergaming, meaning it focused on improving balance.	What specific type of exergaming intervention was examined in the study? The study examined the effects of a balance-based exergaming intervention.
Cluster 2: outcome measures		
What tests were used to assess the participants' stability? One of the tests used to assess the participants' stability was the limits of stability (LOS) test.	What do LOS, OLS, BBS and TUG mean? LOS stands for the limits of stability, OLS for one-leg stance, BBS Berg Balance Scale, and TUG, timed up and go test. These are all tests to measure the balance of patients.	How was postural stability assessed in the study? Postural stability was assessed using the limits of stability (LOS) and one-leg stance (OLS) tests.
What other tests besides the LOS test were used to assess stability? Besides the LOS test, the one-leg stance (OLS) test was also used to assess stability in the participants.		Which tests were used to assess balance in the study? Balance was assessed using the Berg Balance Scale (BBS) and the timed up and go (TUG) test.
What tests were used to assess balance in the participants? The Berg Balance Scale (BBS) was used to assess the balance of participants.		
Besides BBS, were there any other tests used to assess balance? Yes, another test called the timed up and go (TUG) test was used along with the BBS to assess balance in the participants.		
Cluster 3: differences between experiment groups		
Was there any improvement seen in specific conditions of the balance and stability tests? Yes, the participants in the exergaming group showed improvements in the eyes-closed condition of the one-leg stance (OLS) test.	How did the exergaming group compare in some of the stability tests compared to the traditional training group? The exergaming group did better in the OLS test, when they had to stand on one leg with their eyes closed, and in the LOS test, particularly in how much the participant can intentionally move their center of gravity in a specific direction. The exergaming group scored an average of 78.9 in the LOS test while the traditional training group scored an average of 70.6.	In what aspect did the group that did the Kinect exergame improve more than the group that did traditional training? The group that did the Kinect exergame showed a greater improvement in postural stability compared to the group that did traditional training.
What was the performance level in the LOS test for directional control in the group that used the Kinect exergame for balance training? In the balance-based exergaming group, the performance in the LOS test for directional control was found to be at an average of 78.9%, with a degree of variation of approximately 7.65%.		
How did the conventional balance training group perform in directional control in the LOS test? The conventional balance training group scored an average of 70.6%, with a degree of variation of approximately 9.37% in the LOS test for directional control.		
Cluster 4: statistical significance of results		
Was there a significant difference in the performance of the two groups in terms of directional control? Yes, the group that did the balance-based exergaming training significantly outperformed the conventional balance training group in terms of directional control.	How reliable are these results? The improvements in the exergaming group compared to the traditional training group in the LOS and OLS scores were found to be significant, or unlikely to occur by chance.	
Other questions, not aligned across models		
Was the study conducted in such a way that the participants didn't know what group they were in? Yes, the study was setup in a way that the participants didn't know whether they were in the exergaming group or in the traditional balance training group. This is known as a subject-blinded study.		What were the stages of Parkinson's disease of the participants in the study? The participants in the study were at Hoehn and Yahr stages I through III of Parkinson's disease.
Was the study conducted under controlled conditions? Yes, the study was a controlled one. This means that the researchers could accurately compare the effects of the exergaming and the traditional balance training.		What was the design of the study? The study was a subject-blinded, randomized controlled study. This means that the participants did not know which group they were assigned to (exergaming or traditional balance training), and the assignment to the groups was done randomly to avoid bias.
In what areas were improvements seen after the training? There were improvements seen in the performance in the Berg Balance Scale (BBS) test after the training.		
Apart from BBS, where else were improvements seen in participant performance? Apart from the BBS test, improvements were also seen in the timed up and go (TUG) test after the training period.		

Figure A.4: Full example comparing QAs generated by the NLI pipeline, humans and GPT-4. Colors indicate semantic alignment between questions. We observe different tendencies for question compositionality: while humans aggregate several missing information into higher-level questions, the NLI pipeline has a tendency to generate more restricted questions, even though in aggregate they cover a similar content.

Please simplify the following technical abstract into plain language that an average adult would understand. If the abstract has sections, keep them.

{{ abstract }}

Figure A.5: GPT-4 prompt to simplify the RCT abstracts. Colors indicate chat roles: **system** and **user**.

You are a helpful assistant to classify text into categories.

Instructions

You are asked to classify questions according to an ontology of question types. The question type reflects the nature of the question. It is NOT determined by the interrogative word of the question. There are 10 question types in total. The definition for each type is shown below. Please select the question type which is most likely for a given question. Only output the category title, not the description.

Question Types

1. VERIFICATION: Asking for the truthfulness of an event or a concept.

- Was the study double-blinded?
- Was there a trend towards smaller increases in macular pigment for subjects with high baseline values?

2. DISJUNCTION: Asking for the true one given multiple events or concepts, where comparison among options is not needed.

- no example available, match by the description

3. CONCEPT: Asking for a definition of an event or a concept.

- What kind of result is being measured in this study?
- What were the main inclusion criteria for this study?
- What does the WCJ-III test specifically measure?
- Which areas of the brain were analyzed?

4. EXTENT: Asking for the extent or quantity of an event or a concept.

- How reliable are these results?
- How long were the participants observed?
- How much ibuprofen was in the small dose?
- How many young and old people participated in the study?

5. EXAMPLE: Asking for example(s) or instance(s) of an event or a concept.

- What kind of conditions cause corneal neovascularization?

6. COMPARISON: Asking for comparison among multiple events or concepts.

- On what results did the control group do better than the intervention group?
- How did headache of participants receiving ibuprofen compare to those participants that received a placebo?
- How more effective was the arm cranking exercise with and without electrical muscle stimulation?

7. CAUSE: Asking for the cause or reason for an event or a concept.

- What motivates this study?
- Why is EMS being investigated?

8. CONSEQUENCE: Asking for the consequences or results of an event.

- What was the effect of using ibuprofen to treat headaches?
- What were the main findings of the study?
- What does this study tell us about arm cranking with electrical muscle stimulation?

9. PROCEDURAL: Asking for the procedures, tools, or methods by which a certain outcome is achieved.

- What questionnaire was used for this study?
- What kind of lab tests were done?
- How were the patients assigned to a group?
- How were the different drugs administered to participants?

10. JUDGMENTAL: Asking for the opinions of the answerer's own.

- no example available, match by the description

Instances to classify

```
{% for i, question in enumerate(questions) %}  
  {{ i+1 }}. {{ question }}  
{% endfor %}
```

Figure A.6: Few-shot prompt for GPT-4 to classify questions according to the taxonomy of Cao and Wang (2021). Colors indicate chat roles: **system** and **user**.

Statement: {{ sent }}

List all the facts we explicitly know from the statement. Make each fact as atomic as possible.

Figure A.7: GPT-4 prompt to decompose a sentence into a list of atomic facts.

```
## Original
{{ original }}

## Simplification
{{ simplification }}

## Missing facts
{{ facts }}

The above facts are missing from the simplified text ("Simplification"). For each fact, please write a question–answer pair that would elicit the missing information from the original text ("Original"). Phrase the question in such a way that a reader can understand it without having seen the original text. It should only contain concepts (entities, events, or states) that were mentioned in the simple text, or concepts that have not been directly mentioned but are generally known or inferable from mentioned ones. The answer should be understandable by an average adult, so please explain technical jargon if necessary. Make each question–answer pair as specific as possible and make sure that they are independent of each other. Ask only about one information unit at a time. Do this for all facts, and format your output as follows:

– Fact:
– Question:
– Answer:
```

Figure A.8: GPT-4 prompt to generate an INFOLOSSQA pair given original, simplification and a missing fact.

You are an expert annotator for outputs of text simplification systems. This annotation task is to identify pieces of information that were lost in the simplification process. You will be given two texts: the original and a simplification. Assume that a lay reader only sees the simplification. Identify all things which the reader can NOT learn from the simplification but that they could have learnt from the original.

Please classify each unit of information into one of the following two categories:

1. Omissions: Pieces of information which were not included in the simplification.
2. Imprecise/fuzzy concepts: Pieces of information which are included in the simplification, but that have been simplified to the extent that they became imprecise or completely lost their original meaning.

Afterwards, please write a question–answer pair that would elicit the missing information from the original text. Phrase the question in such a way that a reader can understand it without having seen the original text. It should only contain concepts (entities, events, or states) that were mentioned in the simple text, or concepts that have not been directly mentioned but are generally known or inferable from mentioned ones. The answer should be understandable by an average adult, so please explain technical jargon if necessary. Make each question–answer pair as specific as possible and make sure that they are independent of each other. Ask only about one information unit at a time.

A useful heuristic to decide between omissions and imprecise/fuzzy concepts is to see if the question–answer pair clarifies/expands some topic which is discussed in the simplification. If so, classify it as imprecise/fuzzy concepts, otherwise omission.

Adhere to this output format:

- Original: ""
- Rationale: <a short rationale that describes what makes this an information loss>
- Question: <the question that elicits missing information>
- Answer: <the answer that provides the missing information from the original span>

For fuzzy/imprecise concepts, please also indicate the corresponding span in the simplified text.

Here is an example.

Original

This study evaluates the effects of vitamin D3 on disease activity and quality of life in ulcerative colitis (UC) patients with hypovitaminosis D. The study was a prospective double–blinded, randomized trial conducted at Community Regional Medical Center, Fresno, CA from 2012–2013. Patients with UC and a serum 25(OH)D level <30 ng/ml were eligible for the study. Enrolled subjects were randomized to receive either 2,000 IU or 4,000 IU of oral vitamin D3 daily for a total of 90 days. The Short IBD Questionnaire (SIBDQ) for quality of life, the Partial Mayo Score for UC disease activity and serum lab tests were compared between the two treatment groups.

Simplification

This study looks at whether taking vitamin D3 can affect a particular form of bowel disease (ulcerative colitis) and improve the lives of patients with low levels of vitamin D. The study was carried out at a medical center in Fresno, California, between 2012 and 2013. Patients with this disease and low vitamin D levels were included. Participants were randomly given either 2,000 IU or 4,000 IU of oral vitamin D3 daily for 90 days. Researchers used a survey about participant's quality of life and conducted lab tests between the two groups.

Omissions

- Original: "Partial Mayo Score for UC disease activity"
- Rationale: the simplification does not mention that the Partial Mayo Score was used.
- Question: Besides quality of life survey and the serum tests, what other measures did the researchers use?
- Answer: The researchers also tracked the partial mayo for UC diseases. UC stands for ulcerative colitis and is an inflammatory bowel disease. The mayo score indicates how active or severe the disease is.

Imprecise/fuzzy concepts

- Original: "The study was a prospective double–blinded, randomized trial"
- Simplification: "The study"
- Rationale: The simplification does not explain the design of the study, it only mentions that it was a "study".
- Question: How did the study control for bias?
- Answer: The study was double–blinded, so that neither the researcher nor the participants knew which treatment each participant received, and it was randomized, meaning the participants were randomly assigned one of the treatments.
- Original: "a serum 25(OH)D level <30 ng/ml"
- Simplification: "low vitamin D levels"
- Rationale: the simplification does not explain the inclusion criteria, namely how low the vitamin D levels of eligible patients were.
- Question: How low were the vitamin D levels in patients that were included in the study?
- Answer: Participants in the study all had less than 30ng/ml of vitamin D, which is below the minimum recommendation of vitamin D levels in the body.
- Original: "Short IBD Questionnaire (SIBDQ) for quality of life"
- Simplification: "a survey about participant's quality of life"
- Rationale: the specific type of survey (SIBDQ) is not mentioned in the simplification
- Question: What survey was used to measure participants' quality of life?
- Answer: The Short Inflammatory Bowel Disease Questionnaire (SIBDQ) was used which gives insights about the physical, social, and emotional status of patients with bowel diseases.

Original
{{ original }}

Simplification
{{ simplification }}

Figure A.9: End-to-end system prompt for GPT-4. Colors indicate chat roles: **system** and **user**.

You are an expert annotator for outputs of text simplification systems. This annotation task is to identify pieces of information that were lost in the simplification process. You will be given two texts: the original and a simplification. Assume that a lay reader only sees the simplification. Identify all things which the reader can NOT learn from the simplification but that they could have learnt from the original.

Please classify each unit of information into one of the following two categories:

1. Omissions: Pieces of information which were not included in the simplification.
2. Imprecise/fuzzy concepts: Pieces of information which are included in the simplification, but that have been simplified to the extent that they became imprecise or completely lost their original meaning.

Afterwards, please write a question–answer pair that would elicit the missing information from the original text. Phrase the question in such a way that a reader can understand it without having seen the original text. It should only contain concepts (entities, events, or states) that were mentioned in the simple text, or concepts that have not been directly mentioned but are generally known or inferable from mentioned ones. The answer should be understandable by an average adult, so please explain technical jargon if necessary. Make each question–answer pair as specific as possible. Ask only about one information unit at a time.

A useful heuristic to decide between omissions and imprecise/fuzzy concepts is to see if the question–answer pair clarifies/expands some topic which is discussed in the simplification. If so, classify it as imprecise/fuzzy concepts, otherwise omission.

Adhere to this output format:

- Original Fragment: <exact span in original>
- Simplification Fragment: null for Omissions OR <exact span in simplification> for Imprecise/fuzzy concepts
- Rationale: <a short rationale that describes what makes this an information loss>
- Question: <the question that elicits missing information>
- Answer: <the answer that provides the missing information from the original span>

Strictly include the above elements. There can be many omissions and imprecise concepts, so be exhaustive. Use simple language in the answer. Cite the original/simplification span EXACTLY so that span == text[text.index(span):text.index(span)+len(span)].

Original

This study evaluates the effects of vitamin D3 on disease activity and quality of life in ulcerative colitis (UC) patients with hypovitaminosis D. The study was a prospective double–blinded, randomized trial conducted at Community Regional Medical Center, Fresno, CA from 2012–2013. Patients with UC and a serum 25(OH)D level <30 ng/ml were eligible for the study. Enrolled subjects were randomized to receive either 2,000 IU or 4,000 IU of oral vitamin D3 daily for a total of 90 days. The Short IBD Questionnaire (SIBDQ) for quality of life, the Partial Mayo Score for UC disease activity and serum lab tests were compared between the two treatment groups.

Simplification

This study looks at whether taking vitamin D3 can affect a particular form of bowel disease (ulcerative colitis) and improve the lives of patients with low levels of vitamin D. The study was carried out at a medical center in Fresno, California, between 2012 and 2013. Patients with this disease and low vitamin D levels were included. Participants were randomly given either 2,000 IU or 4,000 IU of oral vitamin D3 daily for 90 days. Researchers used a survey about participant's quality of life and conducted lab tests between the two groups.

Omissions

- Original Fragment: "Partial Mayo Score for UC disease activity"
- Simplification Fragment: null
- Rationale: the simplification does not mention that the Partial Mayo Score was used.
- Question: Besides quality of life survey and the serum tests, what other measures did the researchers use?
- Answer: The researchers also tracked the partial mayo for UC diseases. UC stands for ulcerative colitis and is an inflammatory bowel disease. The mayo score indicates how active or severe the disease is.

Imprecise/fuzzy concepts

- Original Fragment: "The study was a prospective double–blinded, randomized trial"
- Simplification Fragment: "The study"
- Rationale: The simplification does not explain the design of the study, it only mentions that it was a "study".
- Question: How did the study control for bias?
- Answer: The study was double–blinded, so that neither the researcher nor the participants knew which treatment each participant received, and it was randomized, meaning the participants were randomly assigned one of the treatments.
- Original Fragment: "a serum 25(OH)D level <30 ng/ml"
- Simplification Fragment: "low vitamin D levels"
- Rationale: the simplification does not explain the inclusion criteria, namely how low the vitamin D levels of eligible patients were.
- Question: How low were the vitamin D levels in patients that were included in the study?
- Answer: Participants in the study all had less than 30ng/ml of vitamin D, which is below the minimum recommendation of vitamin D levels in the body.
- Original Fragment: "Short IBD Questionnaire (SIBDQ) for quality of life"
- Simplification Fragment: "a survey about participant's quality of life"
- Rationale: the specific type of survey (SIBDQ) is not mentioned in the simplification
- Question: What survey was used to measure participants' quality of life?
- Answer: The Short Inflammatory Bowel Disease Questionnaire (SIBDQ) was used which gives insights about the physical, social, and emotional status of patients with bowel diseases.

Original

{{ original }}

Simplification

{{ simplification }}

Figure A.10: End-to-end system prompt for Mistral. Colors indicate chat roles: **user** and **assistant**.

You are an expert annotator for outputs of text simplification systems. You are given an example, a sample to annotate and detailed instructions.

Original

This study evaluates the effects of vitamin D3 on disease activity and quality of life in ulcerative colitis (UC) patients with hypovitaminosis D. The study was a prospective double-blinded, randomized trial conducted at Community Regional Medical Center, Fresno, CA from 2012–2013. Patients with UC and a serum 25(OH)D level <30 ng/ml were eligible for the study. Enrolled subjects were randomized to receive either 2,000 IU or 4,000 IU of oral vitamin D3 daily for a total of 90 days. The Short IBD Questionnaire (SIBDQ) for quality of life, the Partial Mayo Score for UC disease activity and serum lab tests were compared between the two treatment groups.

Simplification

This study looks at whether taking vitamin D3 can affect a particular form of bowel disease (ulcerative colitis) and improve the lives of patients with low levels of vitamin D. The study was carried out at a medical center in Fresno, California, between 2012 and 2013. Patients with this disease and low vitamin D levels were included. Participants were randomly given either 2,000 IU or 4,000 IU of oral vitamin D3 daily for 90 days. Researchers used a survey about participant's quality of life and conducted lab tests between the two groups.

Missing information in simplification

```
[
  {
    "category": "imprecise/fuzzy concepts",
    "original": "The study was a prospective double-blinded, randomized trial",
    "simplification": "The study",
    "rationale": "The simplification does not explain the design of the study, it only mentions that it was a 'study'.",
    "question": "How did the study control for bias?",
    "answer": "The study was double-blinded, so that neither the researcher nor the participants knew which treatment each participant received, and it was randomized, meaning the participants were randomly assigned one of the treatments."
  },
  {
    "category": "imprecise/fuzzy concepts",
    "original": "a serum 25(OH)D level <30 ng/ml",
    "simplification": "low vitamin D levels",
    "rationale": "The simplification does not explain the inclusion criteria, namely how low the vitamin D levels of eligible patients were.",
    "question": "How low were the vitamin D levels in patients that were included in the study?",
    "answer": "Participants in the study all had less than 30ng/ml of vitamin D, which is below the minimum recommendation of vitamin D levels in the body."
  },
  {
    "category": "imprecise/fuzzy concepts",
    "original": "Short IBD Questionnaire (SIBDQ) for quality of life",
    "simplification": "a survey about participant's quality of life",
    "rationale": "The specific type of survey (SIBDQ) is not mentioned in the simplification",
    "question": "What survey was used to measure participants' quality of life?",
    "answer": "The Short Inflammatory Bowel Disease Questionnaire (SIBDQ) was used which gives insights about the physical, social, and emotional status of patients with bowel diseases."
  },
  {
    "category": "omission",
    "original": "Partial Mayo Score for UC disease activity",
    "simplification": null,
    "rationale": "The simplification does not mention that the Partial Mayo Score was used.",
    "question": "Besides quality of life survey and the serum tests, what other measures did the researchers use?",
    "answer": "The researchers also tracked the partial mayo for UC diseases. UC stands for ulcerative colitis and is an inflammatory bowel disease. The mayo score indicates how active or severe the disease is."
  }
]
```

Original

{{ original }}

Simplification

{{ simplification }}

Instructions

This annotation task is to identify pieces of information that were lost in the simplification process. Annotate the sample above. Assume that a lay reader only sees the simplification. Identify all things which the reader can NOT learn from the simplification but that they could have learnt from the original.

Please classify each unit of information into one of the following two categories:

1. Omissions: Pieces of information which were not included in the simplification.
2. Imprecise/fuzzy concepts: Pieces of information which are included in the simplification, but that have been simplified to the extent that they became imprecise or completely lost their original meaning.

Afterwards, please write a question–answer pair that would elicit the missing information from the original text. Phrase the question in such a way that a reader can understand it without having seen the original text. It should only contain concepts (entities, events, or states) that were mentioned in the simple text, or concepts that have not been directly mentioned but are generally known or inferable from mentioned ones. The answer should be understandable by an average adult, so please explain technical jargon if necessary. Make each question–answer pair as specific as possible and make sure that they are independent of each other. Ask only about one information unit at a time.

A useful heuristic to decide between omissions and imprecise/fuzzy concepts is to see if the question–answer pair clarifies/expands some topic which is discussed in the simplification. If so, classify it as imprecise/fuzzy concepts, otherwise omission.

Include following elements in your annotation:

- Original: <exact span in original>
- Simplification: <exact span in simplification> or null for omissions
- Rationale: <a short rationale that describes what makes this an information loss>
- Question: <the question that elicits missing information>
- Answer: <the answer that provides the missing information from the original span>

Strictly follow the above json format. There can be many omissions and imprecise concepts, so be exhaustive. Use simple language in the answer. Cite the original/simplification span EXACTLY so that span == text[text.index(span):text.index(span)+len(span)]. Output ONLY the json!

Missing information in simplification

Figure A.11: End-to-end system prompt for Llama2. Colors indicate chat roles: **system**, **user** and **assistant**.

B Annotation Guidelines

Introduction. Text simplification aims to rewrite a complex text into a simpler version that can be understood by a lay audience. When simplifying, editors decide what and how to simplify, often omitting content which is deemed less important or too technical. However, this can deny readers access to potentially useful information and the opportunity to learn new terms and concepts.

Goal. We aim to identify instances where information is lost and to recover it through Question-Answer (QA) pairs. We hypothesize that readers can better understand the text by referring to the simplified version and the associated QA pairs.

The data. We work with abstracts of randomized controlled trials (RCTs) along with their automatically generated simplifications. RCTs are scientific experiments testing the efficacy of clinical interventions like new drugs, treatments, or diagnostic methods. They typically involve recruiting patients and dividing them into an experimental group (receiving the intervention) and a control group (not receiving it).

Annotation task. Your task involves the following steps:

1. Read both the original and simplified text
2. Compare the two texts and highlight
 - (a) **Deletions** from the original, and
 - (b) **Oversimplifications** in the simplification
3. For each highlight, write a QA pair that re-introduces the information in lay language.

You can find two examples below.

Deletions. Pieces of information which were not included in the simplification. The questions should allow readers to reveal the omitted content.

Original: First, nine healthy young men performed two ① **20-min** arm-cranking trials ② **at 50% VO₂-max** with and without EMS applied to the lower limbs.

Simplified: First, they had nine healthy young men do the arm-cranking exercise with and without the EMS added to the lower body.

Instance 1:

Q: For how long is the exercise applied?

A: Participants do the exercise twice for 20 minutes.

Instance 2:

Q: At what intensity is the exercise applied?

A: The exercise is applied at 50% VO₂-max.

Oversimplifications. Pieces of information that were simplified to the extent that they became too imprecise or completely lost their original meaning. The questions should clarify those concepts and restore their original meaning. For this category we highlight both the simplification and the original text that elicited the simplification.

Original: The **flow-mediated vasodilation (FMD)** of the right brachial artery was measured using a high-resolution ultrasound device.

Simplified: They used a high-resolution ultrasound to measure **how well the main artery in the arm was working.**

Q: How do researchers measure how well the intervention (EMS) works?

A: The researchers measure to what extent the main artery of the arm widens, which is called flow-mediated vasodilation (FMD).

Scenario grounding: deciding what is important enough.

- Do role play. Imagine you are someone who could benefit from understanding the RCT. For example, a patient who has a condition that this RCT addresses.
- When in doubt, try to take an inclusive perspective. People can always decide a question is not relevant to them.
- For numerical results like p-values or Z-scores, make a judgment if they are necessary for a correct understanding of the RCT. Do not interpret these results, rather rely on the authors interpretation of the values.

Tips and other notes.

- When is something a deletion vs. oversimplification? A useful heuristic is to see if you could “attach” the omitted information somewhere to the simplification. If so, it is likely an oversimplification.
- You are free to use a search engine for writing the answers. Please only use high quality sources.
- The text may refer to the same concept multiple times. Please highlight all occurrences, and copy-paste the QA. Add any comments, questions or concerns in the comment field.
- Ignore other artifacts in the simplifications that are not about an information loss, including factual errors, under-simplifications (i.e., when something is still too complicated) and unnecessary information (e.g., the registration number of a trial).

Checklist for a good QA pair.

- Address an information gap between the original and the simplification.
- A question should be self-contained. Readers should get a sense of “why” it is an important question to ask and “what” they will learn if they look at it.
 - Deletions: keep in mind that readers only see the simple text. So you may need to add a bit more context into the question to make it apparent to readers “why” this is important.
 - Oversimplifications: explicitly connect to the concept that is being clarified. This can often be done by somehow including the concept in the question.
- Question scoping: try to phrase the question such that the highlighted piece is the most likely answer, ideally there should be a singular answer.
- The answer must draw on information from the original.
- The answer must use plain language that an average adult would understand. You may have to explain difficult terms and concepts.
- For each QA pair, please check that the answer cannot be found at a later stage in the simplification. The reason is that a simplification may reorder content or drop redundant information to reduce the length. We define information loss globally, so be aware of these cases. Basically, this is a small sanity check that you can do while drafting the QA pair.
- **Avoid:** Yes/no questions (rather, reformulate it using a question word: what, how, why, etc.).
- **Avoid:** Questions that ask for further simplification or external information (these are not about an information gap).

The screenshot displays the INFOLOSSQA annotation interface. On the left, a text editor shows a comparison between the 'Original' and 'Simplification' of a paragraph about a clinical trial. The original text is on the left, and the simplified text is on the right. A red box highlights a portion of the original text, and a blue box highlights the corresponding portion in the simplified text. On the right side of the interface, there are three buttons: 'Deletion' (with a red minus sign), 'Oversimplification' (with a blue plus sign), and 'Comment'. Below these buttons, there are two sections for selecting text spans: 'Select the text span from the Original.' and 'Select the text span from the Simplification.' Each section has a 'Selected:' field with a dropdown menu and a text input field. Below these sections, there is a question field: 'A question that would make the concept more concrete.' with a text input field. At the bottom right, there is a 'Comment (optional)' field with a text input field and 'CANCEL' and 'SAVE' buttons. The interface also shows 'EDIT ANNOTATIONS (0/0)' and an 'Add Edit' button.

Figure B.1: Annotation interface for INFOLOSSQA built with Thresh (Heineman et al., 2023).

C Evaluation Guidelines

We seek to evaluate models that identify information loss caused by simplifying text. These models provide two outputs: (i) a localization of what information was lost, and (ii) a QA pair that elicits the missing information.

C.1 Model Recall of Human-written QA

We aim to determine how many of the human-written QAs are generated by the models. This is a pairwise comparison. You will be given a reference QA and a model QA. Please assign one of the following categories:

- **Fully recalled:** the model QA gives the same information as the reference QA
- **Partially recalled:** the model QA partially answers the reference QA
- **Not recalled:** there is no overlap in the presented content

As our goal is to identify if models identify the same information loss, we should abstract away from surface level dissimilarities and framing of questions. Here are some guidelines to help with this judgment:

- First, identify the specific unit of information that the reference is asking about by looking both at the reference question, answer and (if needed) the localization. Then check if this information is conveyed by the model answer. The full texts and answer localizations are given for your reference. Note, however, that the model localization may not reflect the generated QA.
- A useful heuristic is to check how well the reference answer could be replaced by the model-generated answer, even if this may lead to a slightly weird/unintuitive phrasing.
- Disregard background explanations. When a reference answer explains a technical concept but the model answer does not include this explanation, it can still be fully recalled. We evaluate simplicity separately.
- Both the reference or the model-generated QA can include more information that cannot be matched to the other QA. For example, the model QA could ask a broader question which entails multiple reference QAs.
- In examples where both QAs are asking for the same information, but the model answer misinterprets the information, we mark as not recalled.

See the table below for example annotations.

#	Reference QA	Model QA	Rating	Comment
1	How were the carbon nanoparticles administered to the participants? This study was a dose-escalation study, meaning that the carbon nanoparticles were administered in increasing doses sequentially to different groups of participants.	What type of study was conducted to test the effects of carbon nanoparticles? An investigator-initiated, randomized controlled, dose-escalation study was conducted.	Full	Disregard background explanation in the original.
2	What condition did the babies in the study have? The babies in the study were diagnosed with moderate-to-severe bronchiolitis.	How was the study conducted? The study was a double-blind, randomized controlled trial on infants (1 to 12 months) who were diagnosed in the emergency department with moderate-to-severe bronchiolitis.	Full	Fully recalled even if phrased differently
3	What statistical tests were used to analyze the results? The results were analyzed using a repeated measures (the procedures were conducted multiple times), mixed design ANOVA test (compares the results from three or more study groups).	What type of statistical analysis was used to analyze the results? A repeated measures, mixed design ANOVA was used to analyze the results.	Full	Disregard background explanation

continued ...

... continued

#	Reference QA	Model QA	Rating	Comment
4	What was done to prevent bias in the trial? The trial was designed as a double-blind trial, meaning neither the patients nor the researchers knew which group of participants had been assigned which combination of sedative drugs.	How was this study conducted? This study was conducted as a clinical trial at a single center. It was double-blinded, meaning that neither the researcher nor the participants knew which treatment each participant received, and it was randomized, meaning the participants were randomly assigned one of the treatments.	Full	Model QA is quite broad but includes the missing bit of information (double-blind).
5	What was the major criterion for patients to participate in the study? The 155 patients who participated in this study had a glaucoma-related diagnosis, meaning they had been diagnosed with an eye condition that causes vision loss.	What kind of health condition did all enrolled patients have? All enrolled participants were diagnosed with a glaucoma-related condition.	Full	Main subject is "major criterion" which the model QA identifies. 155 patients is not the subject of the question, and background explanation "vision loss" can be ignored.
6	How were the carbon nanoparticles administered to the participants? This study was a dose-escalation study, meaning that the carbon nanoparticles were administered in increasing doses sequentially to different groups of participants.	What were the dosage amounts of the carbon nanoparticles used in the study? Carbon nanoparticles were given in doses of 10 µg, 50 µg, or 100 µg.	Partial	The question asks for dose-escalation. The answer shows increasing doses, but it misses calling out "dose escalation."
7	What specific parts of the digestive system were observed? The study looked at the antrum (part of the upper stomach) and the second duodenal portion (lower part of the stomach).	In this study, where were the muscle contractions assessed? The muscle contractions were assessed in the antrum, which is a part of the stomach.	Partial	
8	What other analyses were done of the rsFC strength? Correlation analyses were done to see if the rsFC strength under heroin exposure was connected to the rewarding effect and plasma (in the blood) concentrations of heroin.	Did the researchers also look for a relationship between the strength of the brain connections and the level of morphine in the patients' blood? Yes, the researchers studied whether there was a correlation between the strength of resting state functional connectivity (rsFC) and the plasma concentration of morphine, a major metabolite of heroin.	Partial	Rewarding effect is missing.
9	What kind of tests were used to assess patients in this study? Participants were assessed with laboratory tests, the United Kingdom screening test [...], the Michigan neuropathy screening score [...], and the Michigan diabetic neuropathy score [...]. In addition, [...]	What method was used to assess the level of neuropathy in the patients? The Michigan neuropathy screening score was used to assess the level of neuropathy in all patients.	Partial	One model generates relatively narrow questions which can touch upon a small part of a larger reference question.
10	What kind of factors is the study analyzing to determine how effective the fast track area (FTA) method is? The study is analyzing the impact of a fast track area (FTA) using both effectiveness measures, such as waiting times and length of stay in the waiting room, and quality measures, such as mortality rates and LWBS rates (patients who leave without being seen by a medical professional).	What is the purpose of the fast track area? The purpose of the fast track area is to reduce waiting times, patient dissatisfaction, and morbidity by providing a separate area for less urgent cases, allowing for faster treatment and better patient flow.	No	Superficially a lot of overlap, but the answer here needs to make clear <i>what is measured</i> not what the fast track area can hypothetically do.

continued ...

... continued

#	Reference QA	Model QA	Rating	Comment
11	What aspects of heart function were increased using the combination of LS and SNP drugs? Left ventricular systolic (the heart's ability to pump blood) and diastolic (heart relaxation and how long it takes for the ventricles, two chambers of the heart, to fill with blood) functions including LVEF (how much blood is pumped out of the heart each time it contracts (beats)) improved using the combination of LS and SNP drugs.	What was the main conclusion of the study? The study shows that the LS and SNP combination really improved heart function during and after surgery. Using these drugs led to less need for other supporting drugs or equipment. Thus, we suggest using this drug combination before and during surgery.	No	Does not give an answer about the actual heart function.
12	How much did the special questionnaire overestimate calcium intake compared to the 24-hour recall? The questionnaire overestimated the average total calcium intake by 221 mg/d (milligrams per day) compared to the 24-hour recall across racial groups.	What was the variability in daily calcium intake estimated by the 24-hour recall? The variability or standard deviation in daily calcium intake estimated based on the 24-hour dietary recall was 433 milligrams per day.	No	
13	How did the other treatments perform in comparison to each other? The results showed that Group 2 that received only the laser treatment showed better results than Group 1 who received only a low-dose of potassium nitrate.	What were the specifics of the treatment groups in the study? The treatment groups in the study were: Group 1 - 5% potassium nitrate, Group 2 - gallium-aluminum-arsenide diode laser with a wavelength of 980 nm, noncontact pulse mode, and power wattage of 0.5 W, Group 3 - combined 5% KNO ₃ and the diode laser, and Group 4 - placebo (control).	No	

What kind of test were the workers being chosen for?
A randomized trial, which is a study design that randomly assigns participants into an experimental group or a control group.

What kind of experiment was conducted after the workers were chosen for the program?
A randomized trial was conducted, meaning the workers were randomly placed into different groups to see how effectively the program worked.

Fully recalled^[1]
 Partially recalled^[2]
 No overlap^[3]

Reference Highlight

BACKGROUND.
It is unclear if objective selection of employees, for an intervention to prevent sickness absence, is more effective than subjective 'personal enlistment'. We hypothesize that objectively selected employees are 'at risk' for sickness absence and eligible to participate in the intervention program.

METHODS.
The dispatch of 8603 screening instruments forms the starting point of the objective selection process. Different stages of this process, throughout which employees either dropped out or were excluded, were described and compared with the subjective selection process. Characteristics of ineligible and ultimately selected employees, for a randomized trial, were described and quantified using sickness absence data.

RESULTS.
Overall response rate on the screening instrument was 42.0%. Response bias was found for the parameters sex and age, but not for sickness absence. Sickness absence was higher in the 'at risk' (N = 212) group (42%) compared to the 'not at risk' (N = 2503) group (25%) (OR 2.17 CI 1.63–2.89; p = 0.000). The selection process ended with the successful inclusion of 151 eligible, i.e. 2% of the approached employees in the trial.

CONCLUSION.
The study shows that objective selection of employees for early intervention is effective. Despite methodological and practical problems, selected employees are actually those at risk for sickness absence, who will probably benefit more from the intervention program than others.

Predicted Highlight

BACKGROUND.
It is unclear if objective selection of employees, for an intervention to prevent sickness absence, is more effective than subjective 'personal enlistment'. We hypothesize that objectively selected employees are 'at risk' for sickness absence and eligible to participate in the intervention program.

METHODS.
The dispatch of 8603 screening instruments forms the starting point of the objective selection process. Different stages of this process, throughout which employees either dropped out or were excluded, were described and compared with the subjective selection process. **Characteristics of ineligible and ultimately selected employees, for a randomized trial, were described and quantified using sickness absence data.**

RESULTS.
Overall response rate on the screening instrument was 42.0%. Response bias was found for the parameters sex and age, but not for sickness absence. Sickness absence was higher in the 'at risk' (N = 212) group (42%) compared to the 'not at risk' (N = 2503) group (25%) (OR 2.17 CI 1.63–2.89; p = 0.000). The selection process ended with the successful inclusion of 151 eligible, i.e. 2% of the approached employees in the trial.

CONCLUSION.
The study shows that objective selection of employees for early intervention is effective. Despite methodological and practical problems, selected employees are actually those at risk for sickness absence, who will probably benefit more from the intervention program than others.

Comment (optional)

Figure C.1: Interface for pairwise evaluation of QA recall. Built with LabelStud.io.

C.2 Quality of Generated QA

Motivation/rationale (all questions). For each rating, please provide a brief explanation that motivates your choice. For positive ratings, the rationale can be omitted. For negative ratings, explain how the QA pair could be improved to get to a positive rating. This helps us to analyze the answers and to improve the annotation guidelines.

Criterion 1: question givenness. A question should be interpretable for a reader. It should only contain concepts (entities, events, or states) that were mentioned in the question context or concepts that are generally known or inferable from mentioned ones. We define question context as follows: For deletions, the context is the entire simple text. For oversimplifications, the context is everything in the simple text up to and including the question localization. For this evaluation, please pretend that you only see the simplified text.

How is the question phrased?

- Good (reader focused, no new concepts)
- Bad (e.g., question introduces new concepts, answer leakage, hallucinations)

Criterion 2: question localization. When the question seeks more information/clarification about an oversimplified concept, the corresponding span in the simplified text should be highlighted. For “Missing:” highlight the corresponding text in the simple text which discusses the topic in an oversimplified way.

To what extent does the highlight relate to the topic under discussion?

- Good: the highlight corresponds to the topic that the question discusses
- Unrelated: the highlighted text does not relate to what the question is asking
- Missing: there should be a highlight, but there is not (please add highlight...)
- n/a: the topic under discussion is not part of the simple text (= deletion)

Criterion 3: answer simplicity. The answer should be easy to understand. Please focus on the readability and simplicity of the answer. This is different from accuracy which we will evaluate later.

Does the answer contain jargon?

- The answer is jargon-free
- The answer contains jargon but it is adequately explained in the answer
- The answer contains jargon but it is adequately explained in the simplified text
- The answer contains unexplained jargon

Is the answer standalone?

- Yes, the answer can be understood without looking at the original
- No, the answer contains confusing aspects (e.g., unresolved coreferences, abbreviations/acronyms)

Criterion 4: answerability/question Relevance. The question should be about an information loss between the original and simplified texts. We evaluate this in two steps: answerability on the original text, and answerability on the simplified text. A question is about an information loss if it is answerable on the original, but unanswerable/only vaguely answerable on the simplified text.

When making your assessment, you may use the answer snippet to help with this evaluation. However, be aware that the snippet may not correctly answer the question. In those cases, disregard the snippet and look for other places which answer the question.

Is the question answerable with the original text?

- Yes, and there is a single obvious answer
- Yes, but there could be multiple valid answers
- No

To what extent is the question answerable with the simplified text (i.e., degree of information loss)? The benchmark for this is the answer on the original text.

- Fully answerable. Asking it on the simplified text would give the same answer or a closely paraphrased answer as on the original.

- Partly or vaguely answerable. The simplified text gives some relevant information, but is less specific or exhaustive than the original.
- Unanswerable.

Criterion 5: answer accuracy. The question should be correctly answered. This criterion only applies to questions that can be answered with the original text. Unanswerable questions are out of scope of this investigation. Ideally a question should be specific enough so that there is a singular answer (see Criterion 4: answerability). If the question is ambiguous/vague (i.e., there are multiple valid answers), its answer has a high chance of being incomplete or only partially answering the question.

Does the provided answer correctly answer the question?

- Yes
- Partially, the answer is related but misses information
- No

Does the provided answer have any hallucinations? Hallucinations are information or claims that cannot be traced back to the original. Disregard general background explanations and elaborations.

- Good: there are no hallucinations
- Bad: the answer contains hallucinations

Does the answer snippet correctly answer the question? Imagine it was rephrased into a fluent answer.

- Yes
- Partially, the answer snippet is related but misses information
- No

Figure C.2: Interface for quality assessment of QA (criteria continued in scrollbar). Built with LabelStud.io.

D Data Release and License

We reused RCT abstracts from the *Evidence Inference V2.0* dataset (DeYoung et al., 2020b; evidence-inference.ebm-nlp.com, accessed 2024-05-28). After personal communication with the authors, it was confirmed that all articles in this dataset were from the PubMed Open Access subset which only includes license terms that allow reuse (ncbi.nlm.nih.gov/pmc/tools/openftlist, accessed 2024-05-28). After discussion with our institutions' librarian on fair use, we release the annotations in INFOLOSSQA under CC-BY-4.0.