Scalable and Domain-General Abstractive Proposition Segmentation

Mohammad Javad Hosseini[1] **Yang Gao**[1] **Tim Baumgärtner**[2*]
Alex Fabrikant[1] **Reinald Kim Amplayo**[1]
[1]Google DeepMind
[2]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
{javadh,gaostayyang,fabrikant,reinald}@google.com
tim.baumgaertner@tu-darmstadt.de

Abstract

Segmenting text into fine-grained units of meaning is important to a wide range of NLP applications. The default approach of segmenting text into sentences is often insufficient, especially since sentences are usually complex enough to include multiple units of meaning that merit separate treatment in the downstream task. We focus on the task of *abstractive proposition segmentation* (APS): transforming text into simple, self-contained, well-formed sentences. Several recent works have demonstrated the utility of proposition segmentation with few-shot prompted LLMs for downstream tasks such as retrieval-augmented grounding and fact verification. However, this approach does not scale to large amounts of text and may not always extract all the facts from the input text. In this paper, we first introduce evaluation metrics for the task to measure several dimensions of quality. We then propose a scalable, yet accurate, proposition segmentation model. We model proposition segmentation as a supervised task by training LLMs on existing annotated datasets and show that training yields significantly improved results. We further show that by using the fine-tuned LLMs (Gemini Pro and Gemini Ultra) as teachers for annotating large amounts of multi-domain synthetic distillation data, we can train smaller student models (Gemma 1 2B and 7B) with results similar to the teacher LLMs. We then demonstrate that our technique leads to effective domain generalization, by annotating data in two domains outside the original training data and evaluating on them. Finally, as a key contribution of the paper, we share an easy-to-use API[1] for NLP practitioners to use.

1 Introduction

From retrieval systems that build indices over passages rather than documents (Tiedemann and Mur, 2008), to automatic evaluation metrics for generative tasks that evaluate sentence-level similarity to references (e.g. Amplayo et al. (2023)), to structured event representations used for cross-document summarization (Zhang et al., 2023), segmenting a document into significantly finer units that retain relevant meaning is a major component of many NLP systems.

In "well-formed" prose, an easy and frequently used choice for segmenting documents is sentence segmentation. But for most applications, sentences are an imperfect fit: they are often still too complex, containing multiple units of underlying information (Chen et al., 2023b; Min et al., 2023); they typically require context from elsewhere in the document to understand the meaning (Choi et al., 2021). Furthermore, well-formed sentences are not always available in situations ranging from casual speech (Stainton, 2005) to non-prose formats (Fang et al., 2024; Maheshwari et al., 2024), where "sentences" are not even a natural unit of discourse.

To provide useful fine-grained segmentation, several recent works have taken the approach of *proposition segmentation*[2] (Chen et al., 2023b; Min et al., 2023; Wanner et al., 2024), seeking to break text into fine-grained, minimal units of meaning that together convey all the information in the source text. Similarly to the extractive-vs-abstractive contrast in the summarization literature, the two strands of proposition segmentation work so far have considered either (a) an *extractive* approach, representing propositions as one or more spans in the source text (Chen et al. (2023b); Gunel et al. (2023); etc.); or (b) few-shot LLM prompts for *abstractive* proposition segmentation, generatively writing out each unit as a well-formed sentence (Kamoi et al., 2023; Wanner et al., 2024; Scirè et al., 2024).

[*] Work done as an intern at Google.

[1] Our Gemma-APS API (Gemma 1 2B and 7B) can be found on Hugging Face.

[2] Others in the literature have also used terms such as "claim decomposition", "claim extraction", and "atomic fact extraction" for the same concept. We follow the naming in (Chen et al., 2023b).

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8856–8872
November 12-16, 2024 ©2024 Association for Computational Linguistics

More formally, abstractive proposition segmentation (APS), the focus of this paper, is to transform a given document into a collection of *propositions* represented as natural-language sentences which: 1. are atomic and minimal semantic unit that cannot be further decomposed into meaningful units (Liu et al., 2023); 2. are fully decontextualized (Choi et al., 2021) — i.e. they can be understood just as well with no access to the rest of the document; 3. present information explicitly given in the document; and 4. when taken together, cover all of the information in the document.

APS has already found applications in grounding (Gao et al., 2023a), summarization evaluation (Liu et al., 2023; Scirè et al., 2024), and fact checking (Min et al., 2023). In this paper, we focus on *making abstractive proposition segmentation practical*. The few-shot prompting approaches are typically too costly to run at large scales, and, furthermore, we show that they tend to under-extract compared to our proposed solutions. Our core contributions are:

1. **A suite of automatic evaluation metrics** to measure the quality of APS methods along several relevant dimensions, allowing informed comparisons between methods
2. **Supervision by existing datasets (Liu et al., 2023)**, which empirically shows improvement on APS over few-shot prompting baselines.
3. **Scalable, domain-general student models (Gemma 1 2B and 7B, Mesnard et al. (2024))** for APS distilled from the supervised models over synthetic multi-domain data (Hosseini et al., 2024), yielding performance comparable to the teacher models even on domains not seen in the human-annotated training data.
4. **An APS API** for NLP practitioners to use.

2 Related Work

Linguistic compositionality, the idea that sentences are comprised by smaller units of meaning, has been debated since the early 1800s (Janssen, 2012), and understood surely long beforehand. In the context of modern NLP, the value of proposition segmentation for standard tasks can be seen from the empirical measurements in, e.g., (Chen et al., 2023b), which shows for document-level NLI that 72% of sentences partially aligned between two highly related documents don't fully entail each other, and in (Min et al., 2023), which shows that 40% of ChatGPT sentences at that time contained a mix of supported and unsupported propositions.

Indeed, several previous results have shown APS by few-shot prompted LLMs benefits retrieval-augmented fact verification and grounding (Kamoi et al., 2023; Min et al., 2023). A concurrent result in (Wanner et al., 2024) looks more specifically at APS itself with few-shot prompting. Scirè et al. (2024) also perform few-shot prompting followed by distillation.

Other formats of proposition segmentation have also been explored. Extractive proposition segmentation is shown in (Chen et al., 2023b,c) to benefit document-level NLI and retrieval. Several open-book QA and grounding works have generated *fine-grained questions* corresponding implicitly to the fine-grained claims in the text (Gao et al., 2023a; Chen et al., 2022, 2023a). In the summarization evaluation literature, "Summary Content Units", initially human-annotated (Nenkova and Passonneau, 2004), later generated from syntactic signals (Gao et al., 2019) have long been used for summary evaluation. Before the LLM era, decomposing text into semantic triples, known as Open Information Extraction (Etzioni et al., 2008), drove a variety of downstream applications.

Our desiderata for proposition segmentation include context-independence, earlier studied at the sentence level by Choi et al. (2021). Deng et al. (2024) perform document-level claim extraction for fact checking. They specifically extract claims that are check-worthy, where these claims are decontextualized, but not necessarily atomic.

In our work, we propose a suite of automatic evaluation metrics. Previous efforts have not focused much on metric definition. The exceptions are two concurrent works: A) Wanner et al. (2024) propose a specific single metric for APS, DecompScore, that combines our reference-free precision metric (§3.2) with the count of claims generated. B) Scirè et al. (2024) propose metrics based on ROUGE (Lin, 2004) and similar to our reference-based precision and recall. Our metrics are based on NLI that is more suitable for checking semantic equivalence of predicted and gold propositions.

3 Abstractive Proposition Segmentation

In this section, we formally define the task (§3.1) and propose metrics (§3.2).

3.1 Task Definition

We are given an input text t, which comprises a naturally-occurring sequence of English words, possibly split into multiple sentences, i.e., $t = \{s_1, ..., s_n\}$. In text t, there are k latent gold propositions (claims) $\{p_1, ..., p_k\}$. The task is then to segment t into a list of propositions $\{q_1, ..., q_k\}$ with the following conditions:

1. **Well-formed**: Proposition q_i should be grammatically correct and conform to the rules of the English language.
2. **Atomic**: Proposition q_i should contain a single atomic fact.
3. **Self-contained**: Proposition q_i should not need additional context to be understood.
4. **Supported**: Proposition q_i should be found in the given text t.
5. **Comprehensive**: The list of propositions $\{q_1, ..., q_k\}$ should cover all the latent gold propositions (claims) in text t.

3.2 Evaluation Metrics

To evaluate systems that produce propositions following the conditions above, we propose two sets of metrics that make use of an entailment model. We employ Natural Language Inference (NLI) as backbone to our metrics because by definition (Dagan et al., 2013), it can be used to check factual support (i.e., one entails another) and semantic equivalence (i.e., both entail each other). In addition, NLI has been successfully used for evaluating factual consistency (Anil et al., 2023a; Gao et al., 2023b; Fierro et al., 2024; Honovich et al., 2022).

We use a T5-11B model (Raffel et al., 2020) fine-tuned on the ANLI dataset (Nie et al., 2020) as our entailment model $\texttt{NLI}(\texttt{premise}, \texttt{claim})$ that returns an entailment score between 0 and 1. This model is shown to obtain the highest factual consistency results on the TRUE benchmark in Honovich et al. (2022)'s experiments.

Reference-free (RF) The first set of metrics compare the system-generated propositions $Q = \{q_1, ..., q_{k'}\}$ with input text $t = \{s_1, ..., s_n\}$, which helps us evaluate whether the propositions are supported and comprehensive. Specifically, we calculate precision RF_p and recall RF_r as follows:

$$RF_p = \frac{\sum_{q_i \in Q} \texttt{NLI}(\texttt{premise} = t, \texttt{claim} = q_i)}{k} \quad (1)$$

$$RF_r = \frac{\sum_{s_j \in t} \texttt{NLI}(\texttt{premise} = \bar{Q}, \texttt{claim} = s_j)}{n} \quad (2)$$

where $\bar{Q}$ is the space-concatenated version of Q to create a single text. Here, precision essentially evaluates whether each proposition q_i in Q is supported in text t, while recall evaluates whether each latent gold proposition mentioned in each sentence s_j is covered in Q. We can then combine both precision and recall by calculating the f1-score RF_{f1}.

Reference-based (RB) The second set of metrics rely on a gold-standard set of propositions $P = \{p_1, ..., p_k\}$ and check whether each proposition in P is semantically equivalent to a predicted proposition (and vice versa). To this end, we use a bidirectional version of NLI where premise and claim need to entail each other, i.e.:

$$\texttt{BiNLI}(p_i, q_j) = \texttt{min}\big(\texttt{NLI}(p_i, q_j), \texttt{NLI}(q_j, p_i)\big) \quad (3)$$

The first NLI call (i.e., does gold entail predicted?) ensures atomicity: If the predicted proposition q_j is not as atomic as a gold proposition p_i, then p_i will not entail q_j (since q_j has more information that p_i). The second NLI call (i.e., does predicted entail gold?) ensures self-containedness: If the predicted proposition q_j is not as self-contained as a gold proposition p_i, then q_j will not entail p_i (since p_i has more information than q_j).

q_j should not need further context (otherwise, the entailment does not hold). We calculate precision RB_p and recall RB_r as follows:

$$RB_p = \frac{\sum_{q_j \in Q} \texttt{argmax}_{p_i \in P} \texttt{BiNLI}(p_i, q_j)}{k'} \quad (4)$$

$$RB_r = \frac{\sum_{p_i \in P} \texttt{argmax}_{q_j \in Q} \texttt{BiNLI}(p_i, q_j)}{k} \quad (5)$$

In RB_p metric, for each predicted q_j, we find the most equivalent p_i based on $\texttt{BiNLI}(p_i, q_j)$, and then average over all predicted propositions. RB_r is calculated similarly in the other direction. Finally, we can combine both precision and recall by calculating the f1-score RB_{f1}. We note that our reference-based scores are equivalent to SMART metrics proposed by Amplayo et al. (2023) as an evaluation metric for text generation. They treat sentences as basic units of information, and compare the set of gold and predicted sentences. We compare propositions as basic units of information rather than sentences.

Note that we do not measure well-formedness since we assume such property for system predictions, given the advancements of pretrained LMs.

In order to validate the effectiveness of our metrics, we perform human correlation studies and show positive results (§6). We show examples of how metrics are calculated in Appendix A.

4 Domain-General APS

Given an input text (passage) t, our goal is to generate a list of propositions $\{p_1, \ldots, p_k\}$, where propositions should be well-formed, atomic, self-contained, supported, and comprehensive.

In this section, we discuss our proposed method to distill a relatively small, yet domain general proposition segmentation model: A) We train a teacher LLM on an existing proposition segmentation dataset (§4.1). B) We generate a large set of multi-domain synthetic data with different lengths (§4.2). C) We generate a large synthetic dataset with pairs of (text, propositions list) and train a student model on it (§4.3).

4.1 Training an APS Model

We train a teacher APS model based on an LLM. In particular, we train a model by using examples in the ROSE dataset (Liu et al., 2023). Each example contains an input text t, and a list of propositions $\{p_1, \ldots, p_k\}$. We trained using two approaches: ungrouped propositions and grouped propositions.

In the ungrouped propositions version, the input contains an instruction and a passage (Figure 1 top). We add an instruction since we use instruction-tuned LLMs for training. The output contains the list of propositions each prepended by "-" and separated by a newline character (Figure 1 bottom).

In the grouped propositions version, we leverage the existing sentence structure from the passage. We split the passage into sentences before feeding it into the proposition segmentation model. We specify the sentence boundaries with special start of sentence (<s>) and end of sentence (</s>) tokens. In addition, we group the propositions of each sentence together and place them inside start and end of sentence tokens. Figure 2 shows an example.

The grouped propositions approach has two benefits: A) The trained model could use the sentence boundaries to obtain improved performance, since it can learn how to generate propositions per sentence rather than generating a longer list of propositions for the full passage. B) During inference, we can automatically attribute each proposition to its corresponding sentence. This is useful for downstream applications. For example, in grounding applications, we can spot which sentences have propositions that are supported or contradicted by an arbitrary source.

We fine-tuned two different LLMs as our teachers: Gemini Pro and Gemini Ultra (Anil et al., 2023a).[3]

4.2 Generating Multi-Domain Synthetic Data

In order to generate a synthetic dataset for distillation, we require a large set of passages so that we can apply the teacher LLM to them and produce (text, propositions) pairs. The ROSE dataset contains examples only in the news domain. To have maximum generalization to new domains, the passages should cover as many domains as possible. In addition, the passage should have different lengths so that the model works well with new texts of different lengths.

We follow Hosseini et al. (2024) that take a practical approach and consider various text properties as contributing factors to domains: text genre, topic, and even the platform or venue that the text comes from. They design a prompt with 18 few-shot examples, where each example is a triple of (*length*, *domain*, *text*). The length can take either the value *short* (just one or a few sentences) or *paragraph*. Appendix B shows an example.[4] The set of 18 few-shot examples cover 8 seed domains such as *shopping reviews*, *twitter* and *reddit post*. However, to have a wide range of domains, they first prompt FLAN-PaLM2 L (Unicorn) model (Anil et al., 2023b) to generate new domains. Then, they manually select a number of non-repetitive domains. Finally, they prompt the LLM to generate text in those domains with the two lengths.

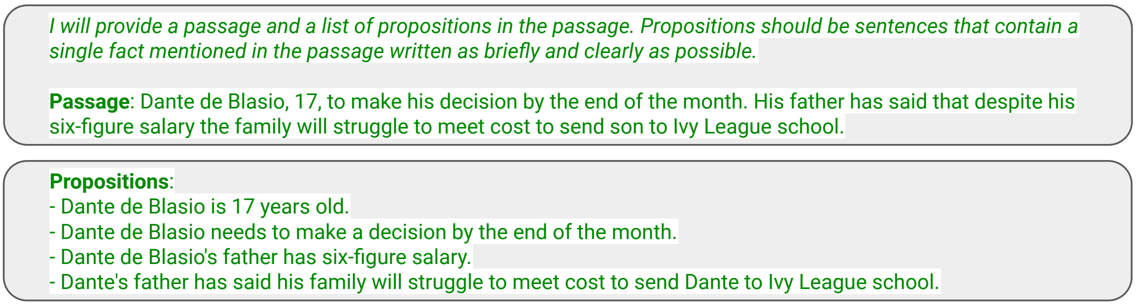
We replicated their approach using Gemini Ultra (Anil et al., 2023a). We first prompted Gemini Ultra $4,000$ times to generate new domains.[5] We obtained a set of 105 domains, from which we manually selected 75. We then prompted the LLM and generated 226K examples with the selected domains and the two lengths.[6]

[3] Available from https://cloud.google.com/apis

[4] The full list can be found in Hosseini et al. (2024).

[5] Many of the calls generate one of the existing domains from the few-shot examples. Therefore, in order to obtain many unseen domains, we prompted the LLM $4,000$ times.

[6] We first generated 228K examples, but filtered examples with $n \geq 4$-gram overlap with any of the seed examples (≈ 2K examples).

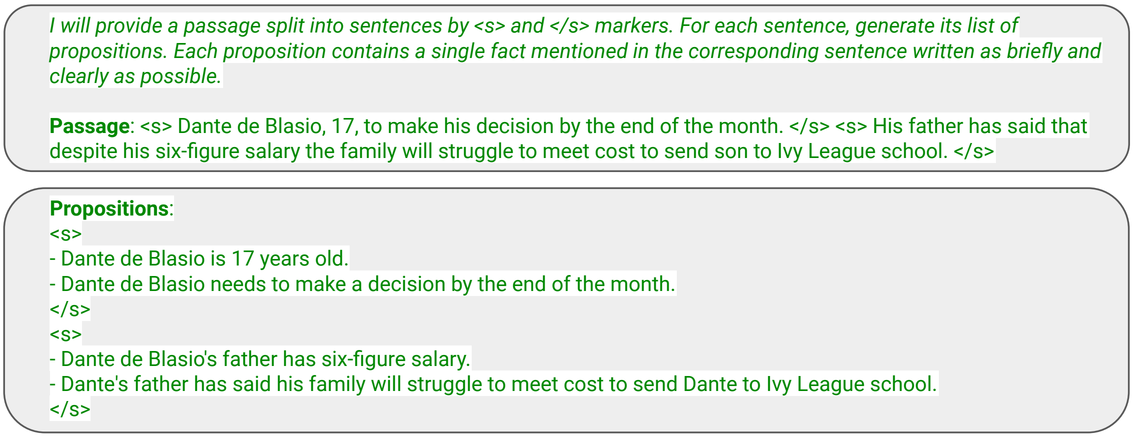


I will provide a passage and a list of propositions in the passage. Propositions should be sentences that contain a single fact mentioned in the passage written as briefly and clearly as possible.

Passage: Dante de Blasio, 17, to make his decision by the end of the month. His father has said that despite his six-figure salary the family will struggle to meet cost to send son to Ivy League school.

Propositions:
- Dante de Blasio is 17 years old.
- Dante de Blasio needs to make a decision by the end of the month.
- Dante de Blasio's father has six-figure salary.
- Dante's father has said his family will struggle to meet cost to send Dante to Ivy League school.

Figure 1: The input (top) and output (bottom) for training an APS model with ungrouped propositions. The input contains an instruction and a passage. The output contains the list of propositions.



I will provide a passage split into sentences by <s> and </s> markers. For each sentence, generate its list of propositions. Each proposition contains a single fact mentioned in the corresponding sentence written as briefly and clearly as possible.

Passage: <s> Dante de Blasio, 17, to make his decision by the end of the month. </s> <s> His father has said that despite his six-figure salary the family will struggle to meet cost to send son to Ivy League school. </s>

Propositions:
<s>
- Dante de Blasio is 17 years old.
- Dante de Blasio needs to make a decision by the end of the month.

</s>
<s>
- Dante de Blasio's father has six-figure salary.
- Dante's father has said his family will struggle to meet cost to send Dante to Ivy League school.

</s>

Figure 2: The input (top) and output (bottom) for training an APS model with grouped propositions. The input contains an instruction and a passage. The output contains the list of propositions. The input passage is separated by special start and end of sentence tokens. Similarly, the output propositions of each sentence are grouped together using special tokens.

4.3 Distillation

The teacher proposition segmentation LLMs learn the task well since they have a large number of parameters and are supervised trained on the ROSE dataset (§4.1). However, they are too costly for direct use in practical applications. Therefore, we distill them into student models.

In our preliminary experiments, we observed better results from the grouped propositions version (§5.3), so we trained the student model based on this type of teacher. We apply the teacher LLMs to the synthetic multi-domain set of texts (§4.2) and produce 226K (text, propositions) pairs. We then train a model with the same input and output format as the teacher models with grouped proposition. We used Gemma 1[7] (2B and 7B) (Mesnard et al., 2024), a lightweight state-of-the-art LM[8], as our student model.

5 Experiments

We explore the effectiveness of our distillation approach for training a *scalable* and *domain-general* proposition segmentation approach. We describe the datasets we have used for training and evaluation (§5.1). We then introduce our baselines (§5.2). We first compare our proposed method with multiple baselines on the ROSE dataset (§5.3). We then show that our method is effective on two datasets from new and unseen domains (§5.4).

5.1 Datasets

We use the annotated ROSE dataset for supervised training. The ROSE dataset has examples from the *news* domain. We manually annotate two out-of-domain datasets, ensuring that the propositions have the desired properties (§3). We use these datasets for assessing the domain generalization capabilities of our models.

ROSE dataset. This dataset is built by manually splitting news summaries into Atomic Content Units (ACUs) for the purpose of evaluating such summaries (Liu et al., 2023). The ACUs are anno-

[7] We used Gemma 1 in all our experiments, but we refer to the language model as Gemma for simplicity throughout most of the following text.

[8] https://ai.google.dev/gemma

tated based on a set of well-defined rules to extract atomic facts, i.e., elementary information units in the input text which no longer need to be further split for the purpose of ambiguity reduction for human evaluation (Liu et al., 2023).

The ACU definition in the ROSE dataset are very close to our propositions definition, therefore we used them for training. We observed some cases in the dataset where the propositions are either not supported or not comprehensive, but we filtered those examples automatically. The dataset contains $2,500$ passages ($21,797$ propositions). We randomly split the dataset into a training and development set (for hyper-parameter tuning[9]). The training set contains $2,089$ passages ($18,994$ propositions), and the development set contains 411 passages ($2,803$ propositions).

The original dataset contains the full set of propositions for each passage. However, for training the grouped propositions version (§4.1), we need to align each proposition to a sentence. We preprocess the dataset to obtain such alignment. We use the NLI score (from T5 11B trained on ANLI) between sentences (premise) and propositions (hypothesis) to obtain the alignment. In particular, for each proposition, we find the sentence with the maximum NLI score. If the NLI score from that sentence $\geq \tau = 0.9$, we align the proposition to the sentence. Otherwise, we discard the example (unsupported proposition). After aligning all the propositions, if a sentence is not aligned with any proposition, we again discard the example (a special case of non-comprehensive propositions). We provide more details about the alignment, filtering, and pre-processing in Appendix D.

The final dataset has high RL_p (supported) and RL_r (comprehensive) scores (§5.3). We manually evaluated the alignment on $\approx$ 200 propositions from the ROSE development set, and the error rate of this approach was $\approx 2\%$. The final training and development sets contain $1,923$ examples ($15,092$ propositions) and 383 examples ($2,237$ propositions), respectively.

Since the examples in the ROSE dataset are based on news summaries, they are quite general and cover many linguistic forms such as presuppositions, attribution to the speaker, modals, and sentence connectors (e.g., *because* and *however*).

Reddit. The Reddit dataset contains 20 randomly sampled human-written answer passages from WebGPT (Nakano et al., 2021), which is a subset of ELI5 dataset, originally used for long-form question answering (Fan et al., 2019). We sampled from one paragraph long answers. We manually annotated the passages with propositions.

Amazon Review. The Amazon Review dataset contains 20 randomly sampled reviews with 3 to 7 sentences from the 2018 version[10] of the Amazon Review Data (Ni et al., 2019). We specifically sampled from the 5-core subset. Finally, we manually annotated each review with propositions.

The manual annotations for the Reddit and Amazon Review datasets were done by two of the authors (each annotated one dataset). The instructions were based on the task definition (§3.1). The authors looked at examples from the ROSE dataset to be mostly aligned with those examples as well.

5.2 Baselines

We compare the following set of models:

Gold has the human annotated propositions.

Sentence is a trivial baseline where we consider each sentence as a proposition.

Few Shot extracts propositions by few-shot prompting an LLM. For each test example, we selected the most similar $K = 10$ examples from the training set based on ROUGE-1 score (Lin, 2004). We report the results for two LLMs, Gemini Pro and Gemini Ultra. We also tried two additional few-shot prompting approaches: A) Randomly sampling few-shot examples (average of 5 runs), B) The few-shot examples from (Wanner et al., 2024). In both cases, the results were overall worse than the dynamic approach and had a similar pattern compared to our models (Appendeix E).

Trained on ROSE are cases where we supervised trained a LM. We trained two versions for each language model (§4.1): ungrouped propositions (UG) and grouped propositions (G). We trained Gemma 7B, Gemini Pro and Gemini Ultra. We also tried Gemma 2B and obtained consistent results with 7B (Appendix E). Moreover, we did preliminary experiments with T5 and obtained consistent results, although lower than Gemma.

Gemma 7B Distilled Models are our final models (§4.3). We fine-tuned Gemma 7B as the student model on distillation data from Gemini Pro and Ultra teacher models (grouped propositions).

[9] See details of hyper-parameters in Appendix C.

[10] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

	REFERENCE-LESS METRICS			REFERENCE-BASED METRICS			
	Precision	Recall	F1	Precision	Recall	F1	# Props
Gold	99.71	96.54	97.53	100.00	100.00	100.00	5.84
Sentence	100.00	100.00	100.00	24.71	18.24	20.42	2.52
FEW SHOT							
Gemini Pro Dyn	99.21	93.26	94.31	47.10	41.41	43.20	4.22
Gemini Ultra Dyn	99.37	89.75	91.88	49.49	47.49	47.74	5.29
TRAINED ON ROSE							
Gemma 7B UG	98.09	96.57	96.54	52.16	50.93	51.02	5.57
Gemma 7B G	98.57	97.48	97.48	53.70	51.43	51.93	5.61
Gemini Pro UG	99.51	97.84	98.20	54.76	52.48	53.02	5.54
Gemini Pro G	99.31	96.66	97.23	55.96	54.87	54.83	5.66
Gemini Ultra UG	99.46	98.05	98.33	**57.69**	56.32	56.45	5.72
Gemini Ultra G	**99.53**	**98.16**	**98.50**	57.62	**56.50**	**56.49**	**5.77**
GEMMA 7B DISTILLED MODELS							
Gemini Pro Data	98.98	97.91	98.02	55.14	53.02	53.50	5.53
Gemini Ultra Data	98.93	98.08	98.23	56.82	55.18	55.41	5.65

Table 1: Results on the ROSE dataset. The methods are split into 4 blocks. The first block has gold and sentence baselines. The second one has few-shot baselines with dynamically (Dyn) selected examples. The third block has baselines directly trained on ROSE with ungrouped (UG) and grouped (G) propositions. The fourth block contains the distilled models results. The best result for each metric (excluding gold and sentence baselines) is boldfaced.

	REFERENCE-LESS METRICS			REFERENCE-BASED METRICS			
	Precision	Recall	F1	Precision	Recall	F1	# Props
Gold	98.72	99.22	98.86	100.00	100.00	100.00	10.70
Sentence	100.00	100.00	100.00	32.99	17.80	22.43	4.40
FEW SHOT							
Gemini Pro Dyn	**100.00**	83.40	89.31	**56.98**	42.03	47.74	7.30
Gemini Ultra Dyn	98.99	72.61	80.44	54.59	44.73	**48.53**	8.15
TRAINED ON ROSE							
Gemma 7B G	98.25	98.73	98.35	35.49	38.30	36.57	10.25
Gemini Pro G	98.80	94.41	96.08	41.80	43.44	42.06	**10.65**
Gemini Ultra G	99.68	96.66	97.83	40.82	44.69	42.39	11.20
GEMMA 7B DISTILLED MODELS							
Gemini Pro Data	99.47	97.08	98.00	45.22	**48.08**	46.20	10.90
Gemini Ultra Data	98.88	**99.96**	**99.40**	40.43	43.21	41.46	11.00

Table 2: Results on the REDDIT dataset. See Table 1's caption for details.

5.3 In-Domain Results

We first compare our method with all the baselines on ROSE development set. Table 1 shows the results. We split the metrics (columns) into two main blocks: reference-less and reference-based (§3.2). In addition, we report the average number of propositions per baseline. In the ideal scenario, the average number of predicted propositions should be as close as possible to gold propositions.

Gold and sentence baselines. The gold propositions have very high RL_p (99.71%) and RL_r (96.54%), which shows that the pre-processed dataset (§5.1) has high quality and satisfy the supported and comprehensive conditions. The RB metrics, on the other hand, are 100% by definition. The average number of propositions is 5.84%. The sentence baseline has perfect RL metrics by definition. However, the RB metrics are very low.

Few-shot models. These baselines (Gemini Pro Dyn and Gemini Ultra Dyn) have very high RL_p (99.21% and 99.37%), but their RL_r (93.26% and 89.75%) is relatively low compared to supervised baselines. The RB metrics are considerably lower than trained models.

Grouped vs Ungrouped versions. Among the trained models, we observe that the grouped ones outperform the ungrouped ones (with only a few exceptions). For examples, Gemma 7B G has 97.48% RL_{f1} (51.93% RB_{f1}), while the UG version has 96.54% RL_{f1} (51.02% RB_{f1}). Therefore, we performed distillation data generation (§4.3) using the grouped propositions version. We also note that the grouped propositions trained models always output the correct format in our experiments, i.e., they output an equal number of start and end tokens, and the same number of groups as the sentences.

Size of trained LMs. Larger LMs get better results than smaller ones when trained on ROSE (with only a few exceptions): Gemini Ultra gets

	REFERENCE-LESS METRICS			REFERENCE-BASED METRICS			
	Precision	Recall	F1	Precision	Recall	F1	# Props
Gold	100.00	99.98	99.99	100.00	100.00	100.00	6.55
Sentence	100.00	100.00	100.00	37.94	24.50	29.21	3.55
FEW SHOT							
Gemini Pro Dyn	99.54	62.97	71.82	50.02	46.74	47.89	6.10
Gemini Ultra Dyn	85.09	53.10	57.17	44.41	44.38	42.06	11.60
TRAINED ON ROSE							
Gemma 7B G	**99.98**	**99.99**	**99.99**	51.90	48.88	49.83	6.25
Gemini Pro G	99.53	97.98	98.50	55.62	54.09	54.52	6.60
Gemini Ultra G	99.83	96.98	98.09	**56.75**	57.08	**56.65**	6.80
GEMMA 7B DISTILLED MODELS							
Gemini Pro Data	98.30	96.72	97.30	56.00	**57.09**	56.25	7.05
Gemini Ultra Data	99.27	99.16	99.21	53.43	53.03	53.09	**6.55**

Table 3: Results on the AMAZON REVIEW dataset. See Table 1's caption for details.

better results compared to Gemini Pro, which itself gets better results than Gemma 7B.

Student models. We trained two different Gemma 7B student models, one trained on distillation data from Gemini Pro teacher model, and one from Gemini Ultra teacher model. The Gemma 7B student models outperform Gemma 7B trained directly on ROSE (i.e., with no distillation) on all metrics. In addition, Gemma 7B student models (the last two rows) perform close to their corresponding teacher models (the last two rows of the trained on ROSE block).

Number of predicted propositions. The number of predicted propositions correlate well with RB metrics and follow similar patterns.

5.4 Out-of-Domain Results

Table 2 and Table 3 show the results of different models on the Reddit and Amazon Review datasets.

Gold and sentence baselines. The gold data has high RL metrics confirming that our annotations satisfy supported and comprehensive conditions. The sentence baselines have perfect RL metrics by definition, but very low RB metrics.

Student models vs teacher models and training directly on ROSE. In both datasets, all the trained and distilled models have very high RL metrics ($RL_{f1} \geq 96\%$). However, the student models perform significantly better than Gemma 7B trained directly on ROSE (i.e., with no distillation) in RB metrics. This confirms that our distillation approach using synthetic multi-domain data leads to successful domain adaptation. In addition, the student models get results on par with teacher models (and sometimes even better) on out-of-domain datasets.

Few-shot models compared to student and teacher. The few-shot models have very low RL_r (53% to 83%) compared to the student models ($\geq 97\%$). This makes the few-shot models unreliable for downstream applications such as fact verification that require to have access to all the claims in the input passage. Table 13 in Appendix F shows examples. The RB metrics for few-shot models is slightly better than the student models on Reddit, but much worse on Amazon Review.

Note on RB metrics. The RB metrics are very strict when comparing gold and predicted propositions, and some minor changes from the gold propositions could lead to low RB metrics. In particular, when computing RB_p, if a predicted proposition is not a paraphrase of any gold proposition, then it will have a score $= 0$ (§3.2).

In many cases, it is challenging and subjective to decide on the right level of atomicity and decontextualization, which directly affects RB metrics (§3.2). Table 4 shows a number of examples where our annotated and predicted propositions (Gemma 7B distilled from Gemini Pro data) are different, although the predicted ones are not necessarily wrong especially when it comes to the atomicity level. For example, the sentence "There are lots of things that feel good that carry some kind of risk" has the right level of atomicity to be considered as a proposition if we want to emphasize on the two points jointly ("feeling good" and "carrying some kind of risk"). Otherwise, the sentence could be split into two propositions. In our work, since we trained the teacher LLMs on the ROSE dataset, we expect the trained models to mirror the atomicity levels in the dataset.

6 Human Correlation Studies for APS Metrics

We measure how well our defined metrics (§3.2) align with human judgements in order to vali-

Input text	Gold Propositions	Predicted Propositions	Category
There are lots of things that feel good that carry some kind of risk.	**There are lots of things that feel good that carry some kind of risk.**	- **There are lots of things that feel good.** - **There are lots of things that carry some kind of risk.**	Atomicity
Fits well and is stylish! Light weight and great options such as the stand. Cant beat this one for the money.	- It fits well. - It is stylish! - It is light weight. - **It has great options.** - **One great option of it is the stand.** - Cant beat it for the money.	- Fits well. - Is stylish. - Light weight. **Great options such as the stand.** - Cant beat this one for the money.	Atomicity
I've always used this type of pick for playing bass. I like the material and the thickness is just right.	- I've always used this type of pick for playing bass. - **I like the material of this type of pick.** - **The thickness of this type of pick is just right.**	- I've always used this type of pick for playing bass. - **I like the material.** - **the thickness is just right.**	Decont
But fish near reefs (or in small streams) have other options. They have shelter to hide behind if they spot a predator (meaning camouflage isn't as important).	- Fish near reefs have other options. - Fish in small streams have other options. - **Fish near reefs have shelter to hide behind if they spot a predator.** - **Fish in small streams have shelter to hide behind if they spot a predator.** - **Camouflage isn't as important as hiding behind shelters.**	Fish near reefs have other options. - Fish in small streams have other options. - **Fish have shelter to hide behind.** - **Fish can hide behind if they spot a predator.** - **Camouflage isn't as important.**	Decont

Table 4: Examples where gold propositions and predicted propositions are not paraphrase because they do not have the exact atomicity or decontextualization level (boldfaced propositions). However, the predicted propositions are not necessarily wrong especially when it comes to the atomicity level. In these cases, the RB scores will be 0.

date the metrics' effectiveness. We performed a study on 40 passages (142 sentences, 263 predicted propositions, and 262 gold propositions) from the Amazon Review dataset. We used the predictions from two models: Gemini Pro few-shot and Gemma 7B distilled from fine-tuned Gemini Pro. The annotations were done by two of the authors (each example were annotated by one author). For each input sentence, we annotated whether the predicted propositions cover all the claims in the sentence (used for measuring reference-less recall). For each predicted proposition, we annotated whether it is supported by the input passage (reference-less precision) and whether it is equivalent to any of the gold propositions (reference-based precision). For each gold proposition, we annotated whether it is equivalent to any of the predicted propositions (reference-based recall).

The Pearson correlation coefficients of example-level metrics and human annotations were generally high (Table 5), confirming that our proposed metrics do correlate well with human judgements (p-value < 0.01).[11]

Metric	Pearson Correlation Coefficient
Reference-based Pr	0.718
Reference-based Rec	0.731
Reference-less Pr	0.476
Reference-less Rec	0.647

Table 5: Pearson correlation coefficients of metrics and human judgements (p-value < 0.01).

7 The `propositions` API

We showed that our student models resolve two issues with the commonly used few-shot prompting approach: under-extraction (low RL_r) and cost. As part of this paper, we release the Gemma-APS API on Hugging Face based on Gemma 1 2B G and Gemma 1 7B G student models trained from Gemini Pro data (grouped propositions version). We invite researchers that require proposition segmentation on input text to try out our models instead of few-shot prompting LLMs.

8 Conclusion

We define the abstractive proposition segmentation task more formally by specifying the desired properties of propositions and present a suite of automatic evaluation metrics that allow us to measure different dimensions of quality. While previous work often uses few-shot prompting, we show that supervision from existing datasets yields significant quality improvement. We then propose a distillation approach for training scalable and domain-general models that get on-par results with the teachers (and sometimes even better). We release an API based on Gemma 7B student models and invite researchers to use that instead of few-shot prompting LLMs.

[11] The reference-less precision metric is almost always equal to 1 except for a few examples. The NLI accuracy compared to human annotation is 0.985. In addition, both the automatic metric and the human evaluated metric are $>= 0.98$.

9 Limitations

In our analysis we showed that reference-based metrics depend on the atomicity and decontextualization level of propositions. On the other hand, the right level of atomicity and decontextualization depends on the downstream applications and how propositions will be used. In addition, our models outputs mirror the atomicity and decontextualization levels of the ROSE dataset examples. Future models and metrics could be flexible in these two levels and let the user decide on the actual style needed for their downstream application.

We used NLI as the backbone to our metrics. We note that NLI as a task is not fully solved, and there are some levels of disagreement in human annotation (Pavlick and Kwiatkowski, 2019; Weber-Genzel et al., 2024). However, we showed strong correlations between human judgements and the defined metrics. In addition, NLI computation is done using a fine-tuned language model, so it is not very lightweight. However, the metric computation usually needs to be done on a small dataset.

We performed our experiments only on English; however, our abstractive proposition segmentation definition and proposed metrics are language independent. In addition, we observed multilingual capabilities with the teacher models when tried on examples from multiple languages. This capability could be used for training multilingual student models in the future.

We note that although our proposition segmentation model is quite accurate and outperforms existing approaches, it is still possible for it to generate wrong and hallucinated outputs, as with all other baselines. Downstream applications should be attuned to the possibility of APS outputs that are occasionally not supported by the original documents.

References

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2023. SMART: sentences as basic units for text evaluation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023a. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023b. Palm 2 technical report. *CoRR*, arXiv:2305.10403.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023a. Complex claim verification with evidence retrieved in the wild. *CoRR*, abs/2305.11859.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2023b. PropSegmEnt: A large-scale corpus for proposition-level segmentation and entailment recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8874–8893, Toronto, Canada. Association for Computational Linguistics.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023c. Dense X retrieval: What retrieval granularity should we use? *CoRR*, abs/2312.06648.

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.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand. Association for Computational Linguistics.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models(llms) on tabular data: Prediction, generation, and understanding - A survey. *CoRR*, arXiv:2402.17944.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

Beliz Gunel, Sandeep Tata, and Marc Najork. 2023. STRUM: extractive aspect-based contrastive summarization. In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 28–31. ACM.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. A synthetic data approach for domain generalization of NLI models. *CoRR*, abs/2402.12368.

Theo Janssen. 2012. 19 Compositionality: Its Historic Context. In *The Oxford Handbook of Compositionality*. Oxford University Press.

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

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

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella, and Anandhavelu Natarajan. 2024. Presentations are not always linear! GNN meets LLM for document-to-presentation transformation with attribution. *CoRR*, arXiv:2405.13095.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Bertrand Russell. 2014. The Philosophy of Logical Atomism. *The Monist*, 28(4):495–527.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14148–14161, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Robert J. Stainton. 2005. 383In Defense of Non-Sentential Assertion. In *Semantics versus Pragmatics*. Oxford University Press.

Jörg Tiedemann and Jori Mur. 2008. Simple is best: Experiments with different document segmentation strategies for passage retrieval. In *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pages 17–25, Manchester, UK. Coling 2008 Organizing Committee.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. *CoRR*, abs/2403.11903.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Zixuan Zhang, Heba Elfardy, Markus Dreyer, Kevin Small, Heng Ji, and Mohit Bansal. 2023. Enhancing multi-document summarization with cross-document graph-based information extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1696–1707, Dubrovnik, Croatia. Association for Computational Linguistics.

A Metric Calculation Examples

We show examples of how metrics are calculated in Table 6. In each row, we show an example and the expected and calculated metrics for it. We also mention which propositions property is mainly measured by the metric. Finally, we provide an explanation about how the property affects the score.

B Few-shot Prompting Example for Synthetic Multi-Domain Text Generation

Table 7 shows one of the 18 few-shot examples used to generate synthetic multi-domain data (Section 4.2). The full list can be found in Hosseini et al. (2024).

C Hyper-parameters

For training Gemma modals, we used a batch size of 8, and an initial learning rate of $5e-5$ and minimum learning rate of $5e-7$ with linear warmup cosine annealing (warmup step of 100 and cosine decay exp 1.0). We trained for 1 epoch.

For all few-shot models, we used a temperature of 0. We tried higher temperatures, but the results were worse.

We trained Gemini Pro with two different learning rates, $1e-4$ and $1e-5$, and selected the first one since it gave better results on ROSE development set. We trained the model with a batch size of 32 for around 4 epochs. We saved checkpoints every 50 steps and selected the one with lowest loss on the development set.

D Pre-processing the ROSE Dataset, Aligning Propositions with Sentences, and Filtering Problematic Examples

We pre-process the dataset to improve its quality based on the following steps:

A) In some cases, ACUs end with a space before the period. We remove the extra space. Additionally, some ACUs do not end with a period (and do not end with "..." either). In these cases, we add a period to the ACU.

B) For each sentence, the ROSE dataset annotators first write an ACU consisting the main information from the subject of the main clause. Then, they add one ACU for each additional information in the sentence by adding minimal necessary information to the original ACU (Liu et al., 2023). In some cases, the original ACU is exactly repeated in other ACUs. In these cases, we removed the first ACU as they are often very short and not very informative. For example, the ACU "Many seals are shot" is removed because we also have another ACU "Many seals are shot to death for their fur".

C) As explained in §5.1, we align propositions in the ROSE dataset with their corresponding sentences, and filter problematic examples. In particular, we follow these steps:

1. For each proposition j:
   (a) compute NLI (sentence i, proposition j) for all sentences. If the sentence with maximum NLI score to proposition j has a score $\geq \tau$=0.9, then we use that sentence as the alignment. Otherwise:
   (b) Compute NLI (prefix $(i-1)+$ sentence i) for all sentences, where prefix $(i-1)$ means the sentences up to sentence i, and $+$ means space concatenation. Find the first sentence which yields entailment score $\geq \tau$=0.9 (if any). If such a sentence exists, we use that as the alignment. Otherwise, we discard the example.
2. If a sentence is not aligned with any proposition, we discard the whole example.

The reason that in step 1 (b), we add the prefix of the sentences before computing the NLI score is that sometimes the full context is necessary to obtain a high NLI score, e.g., cases where the sentence contains a pronoun, but the proposition has the full name. After aligning all the propositions with the above approach, we autoamtically remove examples that have unsupported propositions, and cases where a sentence might not have any propositions, a special case of non-comprehensive propositions.

Table 8 and Table 9 show filtered examples with unsupported propositions and non-comprehensive propositions list, respectively.

E Full Results on All Datasets

In this section, we show all the results reported in §5 plus two additional sets of results. Table 10, 11, and 12 show the full results.

Few-shot results with random examples and examples from Wanner et al. (2024). In §5, we showed few-shot prompting results with dynamically selected examples. In this section, we also add few-shot prompting results with few-shot examples randomly selected per test example. We

Metric	Property	Example	Expected Score	Calculated Score	Explanation
Reference-less precision	Supported	Passage = "The price of the books are all less than ten dollars, and they download before you can get up for a cup of coffee." – Predicted Propositions = ["The books download before you can get up for a cup of coffee."]	1	0.9999	The predicted proposition is entailed by the passage.
Reference-less recall	Comprehensive	Passage = "The price of the books are all less than ten dollars, and they download before you can get up for a cup of coffee." – Predicted Propositions = ["The price of the books are all less than ten dollars."]	0	0	The predicted propositions do not cover all the information in the passage.
Reference-based precision	Self-contained	Predicted Propositions = ["The price of the books are all less than ten dollars.", "They download before you can get up for a cup of coffee."] – Gold Propositions = ["The price of the books are all less than ten dollars.", "The books download before you can get up for a cup of coffee."]	0.5	0.5	The second predicted proposition is not as self-contained as the second gold proposition ("They" vs "The books"). Therefore, the second predicted proposition should get a score of 0 when calculating reference-based precision.
Reference-based recall	Self-contained	Predicted Propositions = ["The price of the books are all less than ten dollars.", "They download before you can get up for a cup of coffee."] – Gold Propositions = ["The price of the books are all less than ten dollars.", "The books download before you can get up for a cup of coffee."]	0.5	0.5	The second gold proposition is more self-contained than the second predicted proposition ("The books" vs "They"). Therefore, the second gold proposition should get a score of 0 when calculating reference-based recall.
Reference-based precision	Atomic	Predicted Propositions = ["The price of the books are all less than ten dollars. They download before you can get up for a cup of coffee."] – Gold Propositions = ["The price of the books are all less than ten dollars.", "The books download before you can get up for a cup of coffee."]	0	0	The predicted proposition is not as atomic as any of the gold proposition. Therefore, it gets a score of 0 when calculating reference-based precision.
Reference-based recall	Atomic	Predicted Propositions = ["The price of the books are all less than ten dollars. They download before you can get up for a cup of coffee."] – Gold Propositions = ["The price of the books are all less than ten dollars.", "The books download before you can get up for a cup of coffee."]	0	0	The gold propositions are more atomic than the predicted proposition. Therefore, they both get a score of 0 when calculating reference-based recall.

Table 6: Examples with expected and calculated metrics. For each example, we provide the propositions property that is mainly measured by the metric. In addition, we explain how the property affects the score.

performed the experiment 5 times. In most cases, the dynmaic approach outperforms the random approach. This is expected since the LLM can learn more from more similar few-shot examples than

Domain	Length	Text
reddit post	paragraph	Hey there everyone! I often see people asking where to start when getting into prog metal, so I thought instead of answering every one of them individually I'd make a list. I'm not going into too much depth because otherwise this will become endless, but I'll try to give a brief explanation of all styles I'm going over. So let's get started!

Table 7: A few-shot example used to generate synthetic multi-domain text. The example has a domain, a length, and a text.

INPUT TEXT
Packs of wild boar are hunting newborn lambs in Britain, experts claim. **Boar at the Forest of Dean usually feed only on plants and dead animals.** But in recent weeks, groups of boar have reportedly killed four lambs. Serious implications for animal health and spread of disease, vet says.
PROPOSITIONS
• newborn lambs are hunted. • Packs of wild boar are hunting in Britain. • Packs of wild boar are hunting, experts claim. • **Boar usually feed only on plants.** • **Boar usually feed only on dead animals.** • The boar is from the Forest of Dean. • groups of boar have reportedly killed lambs. • In recent weeks, four lambs are killed. • Serious implications for animal health. • Serious implications for spread of disease. • They are serious implications, vet says.

Table 8: Example from the ROSE dataset where propositions are not supported by the input text, but we filter the example out. The relevant sentence and unsupported propositions is boldfaced.

INPUT TEXT
Wembley was almost full for England's 4-0 win over Lithunia. **Raheem Sterling linked well with Wayne Rooney and Danny Welbeck. Roy Hodgson must prepare his side for the stiffer tests at Euro 2016. Italy are a different proposition to the side that beat England last summer.**
PROPOSITIONS
• Wembley was almost full. • England won. • The score was 4-0. • England played Lithuania.

Table 9: Example from the ROSE dataset where propositions are not comprehensive, but we filter the example out. The sentences that are not covered by propositions are boldfaced.

random examples.

We also experimented with 21 few-shot prompts from Wanner et al. (2024). These examples are annotated based on Bertrand Russell's theory of logical atomism (Russell, 2014) and neo Davidsonian analysis (Davidson, 1967; Parsons, 1990). This few-shot prompting approach led to generally worse results on all datasets and all metrics, including RL_r (with only one exception).

Gemma 2B results. In §5, we trained Gemma 7B on ROSE and also trained it as a student on distillation data. In this section, we additionally report the results with Gemma 2B. Gemma 2B generally performs slightly worse than Gemma 7B, but we obtain the same trends for Gemma 2B as Gemma 7B. For example, Gemma 2B student models obtain similar results to teacher models and generally obtain better results than Gemma 2B trained on ROSE.

F Few-shot Models Recall Issues

Table 13 shows examples where a few-shot model (Gemini Pro with dynamically selected examples) does not cover some of the facts from the input text, but our student model (Gemma 7B distilled from Gemini Pro data) successfully covers those facts.

	Reference-Less Metrics			Reference-Based Metrics			
	Precision	Recall	F1	Precision	Recall	F1	# Props
Gold	99.71	96.54	97.53	100.00	100.00	100.00	5.84
Sentence	100.00	100.00	100.00	24.71	18.24	20.42	2.52
Few Shot							
Gemini Pro Ran	94.58	93.11	91.05	45.96	48.25	45.49	6.01
Gemini Pro R-ND	94.63	90.43	89.48	40.08	50.82	42.58	8.00
Gemini Pro Dyn	99.21	93.26	94.31	47.10	41.41	43.20	4.22
Gemini Ultra Ran	89.10	91.01	85.89	44.86	50.71	45.58	8.39
Gemini Ultra R-ND	97.70	89.55	90.92	33.97	51.24	39.72	8.93
Gemini Ultra Dyn	99.37	89.75	91.88	49.49	47.49	47.74	5.29
Trained on ROSE							
Gemma 2B UG	96.49	92.64	92.67	51.75	49.75	50.20	5.56
Gemma 2B G	97.46	94.49	94.39	53.29	51.59	51.89	5.62
Gemma 7B UG	98.09	96.57	96.54	52.16	50.93	51.02	5.57
Gemma 7B G	98.57	97.48	97.48	53.70	51.43	51.93	5.61
Gemini Pro UG	99.51	97.84	98.20	54.76	52.48	53.02	5.54
Gemini Pro G	99.31	96.66	97.23	55.96	54.87	54.83	5.66
Gemini Ultra UG	99.46	98.05	98.33	**57.69**	56.32	56.45	5.72
Gemini Ultra G	**99.53**	**98.16**	**98.50**	57.62	**56.50**	**56.49**	**5.77**
Gemma 2B Distilled Models							
Gemini Pro Data	98.20	96.40	96.61	54.13	52.29	52.61	5.46
Gemini Ultra Data	97.55	97.31	96.92	54.73	53.04	53.30	5.64
Gemma 7B Distilled Models							
Gemini Pro Data	98.98	97.91	98.02	55.14	53.02	53.50	5.53
Gemini Ultra Data	98.93	98.08	98.23	56.82	55.18	55.41	5.65

Table 10: Full results on the ROSE dataset. The methods are split into 5 blocks. The first block has gold and sentence baselines. The second one has few-shot baselines with randomly (Ran) selected examples, examples from Wanner et al. (2024) based on based on Russellian and neo-Davidsonian theories (R-ND), and dynamically (Dyn) selected examples. The third block has baselines directly trained on ROSE with ungrouped (UG) and grouped (G) propositions. The fourth and fifth blocks contain Gemma 7B and Gemma 2B distilled models results, respectively. The best result for each metric (excluding gold and sentence baselines) are boldfaced.

	Reference-Less Metrics			Reference-Based Metrics			
	Precision	Recall	F1	Precision	Recall	F1	# Props
Gold	98.72	99.22	98.86	100.00	100.00	100.00	10.70
Sentence	100.00	100.00	100.00	32.99	17.80	22.43	4.40
Few Shot							
Gemini Pro Ran	97.06	80.48	84.36	54.10	47.24	49.79	8.92
Gemini Pro R-ND	94.27	75.33	81.57	54.83	**50.98**	**51.98**	11.60
Gemini Pro Dyn	**100.00**	83.40	89.31	**56.98**	42.03	47.74	7.30
Gemini Ultra Ran	97.87	71.44	78.37	48.69	43.35	45.04	9.04
Gemini Ultra R-ND	96.48	64.60	73.07	43.51	41.70	42.00	11.40
Gemini Ultra Dyn	98.99	72.61	80.44	54.59	44.73	48.53	8.15
Trained on ROSE							
Gemma 2B G	93.95	97.49	95.30	33.78	32.22	32.49	10.40
Gemma 7B G	98.25	98.73	98.35	35.49	38.30	36.57	10.25
Gemini Pro G	98.80	94.41	96.08	41.80	43.44	42.06	**10.65**
Gemini Ultra G	99.68	96.66	97.83	40.82	44.69	42.39	11.20
Gemma 2B Distilled Models							
Gemini Pro Data	98.85	97.08	97.87	46.78	46.07	45.57	11.00
Gemini Ultra Data	99.54	99.71	**99.61**	40.00	42.36	40.84	10.80
Gemma 7B Distilled Models							
Gemini Pro Data	99.47	97.08	98.00	45.22	48.08	46.20	10.90
Gemini Ultra Data	98.88	**99.96**	99.40	40.43	43.21	41.46	11.00

Table 11: Full results on the REDDIT dataset. See Table 10's caption for details.

	REFERENCE-LESS METRICS			REFERENCE-BASED METRICS			
	Precision	Recall	F1	Precision	Recall	F1	# Props
Gold	100.00	99.98	99.99	100.00	100.00	100.00	6.55
Sentence	100.00	100.00	100.00	37.94	24.50	29.21	3.55
FEW SHOT							
Gemini Pro Ran	99.14	66.64	74.33	46.51	43.43	44.50	6.09
Gemini Pro R-ND	97.48	55.97	65.97	38.90	43.89	40.80	7.75
Gemini Pro Dyn	99.54	62.97	71.82	50.02	46.74	47.89	6.10
Gemini Ultra Ran	95.80	49.00	57.19	42.62	39.81	40.22	6.82
Gemini Ultra R-ND	98.75	44.38	56.02	29.99	36.42	32.47	8.45
Gemini Ultra Dyn	85.09	53.10	57.17	44.41	44.38	42.06	11.60
TRAINED ON ROSE							
Gemma 2B G	97.60	98.98	98.01	50.89	49.53	49.99	6.35
Gemma 7B G	**99.98**	**99.99**	**99.99**	51.90	48.88	49.83	6.25
Gemini Pro G	99.53	97.98	98.50	55.62	54.09	54.52	6.60
Gemini Ultra G	99.83	96.98	98.09	**56.75**	57.08	**56.65**	6.80
GEMMA 2B DISTILLED MODELS							
Gemini Pro Data	98.99	94.72	96.55	56.29	56.87	56.35	6.70
Gemini Ultra Data	99.77	96.30	97.77	54.19	53.97	53.89	6.60
GEMMA 7B DISTILLED MODELS							
Gemini Pro Data	98.30	96.72	97.30	56.00	**57.09**	56.25	7.05
Gemini Ultra Data	99.27	99.16	99.21	53.43	53.03	53.09	**6.55**

Table 12: Full results on the AMAZON REVIEW dataset. See Table 10's caption for details.

Input text	**Few-shot**	**Gemma 7B distilled from Gemini Pro**
In the 1940's, the NAACP charged the film with giving the impression of "an idyllic master-slave relationship."	- The NAACP charged the film with giving the impression of "an idyllic master-slave relationship."	- In the 1940's, the NAACP charged the film. - the NAACP charged the film with giving the impression of "an idyllic master-slave relationship."
The scope of such restrictions and requirements varies widely **from state to state and business to business**, but in almost all cases there are a ton of boxes which need to be checked before anyone can move forward.	- The scope of restrictions and requirements varies widely. - There are a ton of boxes which need to be checked before anyone can move forward.	- The scope of governmental restrictions varies widely from state to state. - The scope of governmental requirements varies widely from state to state. - The scope of governmental restrictions varies widely from business to business. - The scope of governmental requirements varies widely from business to business. - In almost all cases there are a ton of boxes which need to be checked. - There are a ton of boxes which need to be checked before anyone can move forward.

Table 13: Examples where a few-shot model (Gemini Pro with dynamically selected examples) does not cover some the facts (boldfaced) from a sentence. However, our student model (Gemma 7B distilled from Gemini Pro data) covers those facts.