DnDScore: Decontextualization and Decomposition for Factuality Verification in Long-Form Text Generation

Miriam Wanner, Benjamin Van Durme, Mark Dredze

Johns Hopkins University

{mwanner5,vandurme,mdredze}@jhu.edu

Abstract

The decompose-then-verify strategy for verification of Large Language Model (LLM) generations decomposes claims that are then independently verified. Decontextualization augments text (claims) to ensure it can be verified outside of the original context, enabling reliable verification. While decomposition and decontextualization have been explored independently, their interactions in a complete system have not been investigated. Their conflicting purposes can create tensions: decomposition isolates atomic facts while decontextualization inserts relevant information. Furthermore, a decontextualized subclaim presents a challenge to the verification step: what part of the augmented text should be verified as it now contains multiple atomic facts? We conduct an evaluation of different decomposition, decontextualization, and verification strategies and find that the choice of strategy matters in the resulting factuality scores. Additionally, we introduce DnDScore, a decontextualization aware verification method that validates subclaims in the context of contextual information.

1 Introduction

Factuality evaluations measure the correctness of language model generations. Recent measures of factual precision utilize a decompose-then-verify framework, where text is first decomposed into atomic subclaims and then validated against a trusted source document (Min et al., 2023; Jiang et al., 2024). However, decomposition may remove information necessary to understand the claim. For example, in Figure 1, the decomposition of the sentence "He was one of the most influential directors in 1930s cinema." would include the subclaim "He was a director.", which lacks entity context (who does "He" refers to?). These ambiguities prevent successful claim verification, and could lead to false positives.

Decontextualization is the process of augmenting subclaims with necessary context to an ensure an accurate claim verification independent of the rest of the generation. Decontextualization can include pronoun replacement, name completion, or addition of information from the original text (Choi et al., 2021). Decontextualizing decomposed atomic subclaims provides necessary information, but introduces several other problems. By introducing new information to the subclaim, the claim becomes less atomic, making it unclear which part of the new claim requires verification. For example, in Figure 1, decontextualizing "He was a director." results in "Quentin Tarantino was an influential 1930s director". As a result, the verification task changes from validating whether Tarantino was a directer, to validating whether he was a *1930s* director. In addition, the potential incorrectness of information added at the decontextualization step can mask the correctness of the original subclaim. Finally, by decontextualizing a set of decomposed claims, we are at risk of building a set of redundant claims that closely resemble the original claim (as shown in top right set of claims in Figure 1). Decontextualization alone still may not solve the original problem, where the subclaim does not stand alone from the original context and therefore cannot be verified.

Issues with factuality evaluation persist even if we flip the verification pipeline from decompose-then-decontextualize to decontextualize-then-decompose, where the original claim is first decontextualized, then decomposed, and finally verified. This approach may result in verifying repeated context across claims, inflating the resulting factuality score. For example in the bottom right of Figure 1, the subclaim "Quentin Tarantino is an American filmmaker." is generated for both Sentence 3 and Sentence 4. Additionally, decomposing the decontextualized claim results in a loss of context for every subclaim added in

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23621–23638
November 4-9, 2025

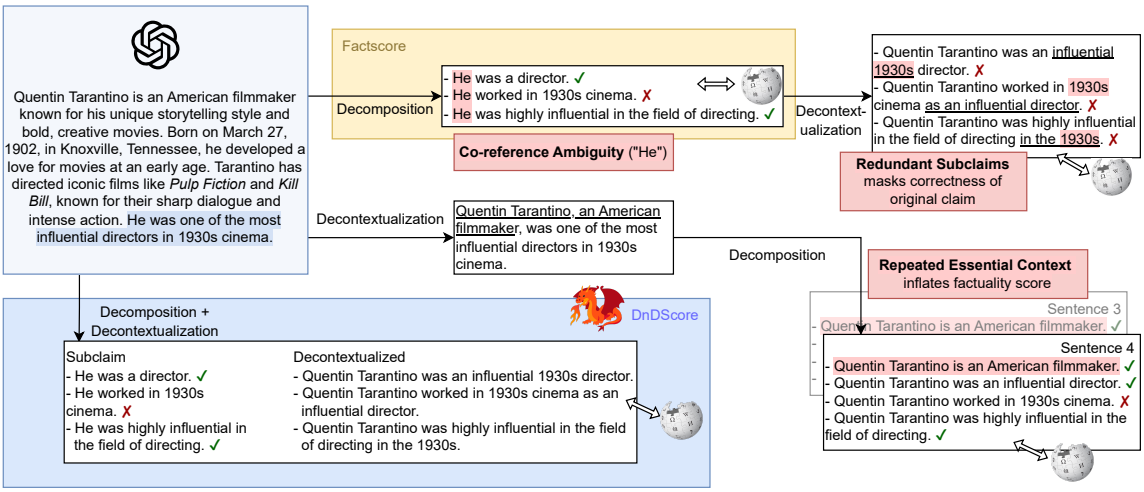

Figure 1: Current claim verification methods evaluate subclaims out of context, but adding this context into the claim verification pipeline is not trivial. We introduce DNDSCORE, a method that evaluates subclaims given the decontextualized claim or another context to help verify.

the decontextualization step. If there was another person named Quentin Tarantino, we would need additional information to disambiguate which Quentin Tarantino was being discussed. Current research has validated the decomposition (Wanner et al., 2024) and decontextualization (Gunjal and Durrett, 2024) processes independently, but have not considered these interactions.

We evaluate three methods for performing both decomposition and decontextualization in the same system. First, we consider decomposition then decontextualization, which enables us to first isolate the atomic subclaims and then add the relevant context. We next consider decontextualization then decomposition, where we ensure claims have necessary context before decomposing them into subclaims. Finally, we propose a new joint method, DnD, which jointly performs decomposition and decontextualization.

Since the resulting decontextualized claim is no longer atomic, it presents an inherent ambiguity to the verified: which portion of the claim is being validated? We propose a new verification method DNDSCORE (Decomposition and Decontextualization Score) that considers both an atomic subclaim and its decontextualized form for verification, indicating the specific claim to verify and its relevant context. We find that using our new contextualized fact verification method changes if claims are verified, highlighting problems with current decompose-then-verify scores. We also show examples and common settings where DNDSCORE and decompose-then-verify methods differ, and where claims are easier or more difficult to verify.

2 Related Works

2.1 Decomposition

There are several systems that implement the decompose-then-verify framework, varying in decomposition and verification strategy. Min et al. (2023) introduce FACTSCORE, which utilizes a LLM-prompted decomposition step for claim verification. Jing et al. (2023) use a similar pipeline, but introduce FAITHSCORE, a variation for evaluating the output of vision-language models. Song et al. (2024) introduce a new flavor of FACTSCORE, called VERISCORE, which extracts and verifies only claims which are verifiable. They include a contextualized method with a sliding window for verifiable claim extraction, but do not use context in the verification step.

Several studies have tried alternate strategies to factuality scoring. Kamoi et al. (2023) introduce a new textual entailment dataset, WICE, and use an automatic LLM Claim-Split approach, separating claims into sub-sentence units used for entailment classification. Using both the dataset and decomposition strategy, they tackle verification and retrieval. Chen et al. (2024) generate yes/no questions aligning to specific aspects of a claim for claim verification. A similar system is built in Chen et al. (2022), also incorporating implied sub-questions. Chen et al. (2023b) introduce a sub-sentence encoder, a contextual embedding model that creates

distinct embeddings for atomic facts within a sentence, and show its use in supporting fact retrieval and text attribution. Li et al. (2024) introduce SELF-CHECKER, a factuality verification method which extracts claims, generates queries from those claims, which are then used for retrieval and ultimately verification of claims. Tang et al. (2024) propose an efficient method for fact-checkin LLMs called MiniCheck. They construct GPT-4 synthetic training data to train a sentence-level fact-checker.

We follow the decompose-then-verify framework, where the decomposition step can have a significant effect on downstream factuality scores. We build on previous work that explores different decomposition methods. We use the decomposition technique of Wanner et al. (2024), who propose a method grounded in Russellian and Neo-Davidsonian theory, and show its benefits over other methods. CORE is a method for subclaim selection that filters based on uniqueness and informativeness (Jiang et al., 2024). Hu et al. (2024) examine the effect of including decomposition in fact verification pipelines, and find there exists a trade-off between accuracy and noise when using decomposition.

2.2 Decontextualization

Decontextualization is the technique of rewriting a sentence to stand alone, interpretable outside of the context of the passage in which it appears. Previous work (Choi et al., 2021) laid out a framework for decontextualization by editing a sentence in four different ways: (1) name completion, pronoun/NP swap, (2) discourse marker removal, (3) bridging global scope, and (4) addition. The last two edits involve adding text, including prepositional phrases or background information.

Gunjal and Durrett (2024) examine the balance between decontextualization, making the sentence stand alone, and minimality, how little information is added. They examine this in the setting of biographies of ambiguous entities, where one name could denote multiple different entities, specifically for fact verification. Wang et al. (2024) use decontextualization as a part of an end-to-end factuality annotation solution for fact checking LLM responses. Wei et al. (2024) develop a method called Search-Augmented Factuality Evaluator (SAFE), for evaluating the factuality of long-form LLM responses. Lee et al. (2024) build a dataset of ambiguous entities (AmbigDocs), relevent to our case even though they do not include decontextualization.

Beyond the application of fact verification, Newman et al. (2023) use decontextualization to rewrite snippets from scientific documents to stand alone. They use three steps: (1) question generation, (2) question answering, (3) rewriting. Potluri et al. (2023) propose an extract-and-decontextualize approach for generating summaries of long-form answers to complex questions. Kane and Schubert (2023) experiment with zero- and few-shot decontextualization using LLMs.

3 Methods: Claim Decomposition and Decontextualization

We now describe methods for decontextualization and decomposition that we adopt in our experiments. Sections 3.1 and 3.2 consider extenstively validated methods from prior work for decontextualization and decomposition, respectively, and Section 3.3 proposes a prompt-based method for joint decontextualization and decomposition (DnD).

3.1 Decontextualization: Molecular Facts

Gunjal and Durrett (2024) evaluate and compare existing decontextualization methods (Wei et al., 2024) on minimality and decontextuality. They call their method "Molecular Facts", which they recommend above other methods of decontexualization due to its balance in minimality. This method uses a two-step prompt for decontextualizing a sentence: first using an LLM to identify ambiguities in the sentence, extracting the ambiguities in an outputted dictionary, and then prompting an LLM with these ambiguities to decontextualize the text. Because this decontextualization method was developed for human biographies, we adopt the Molecular Facts method for our experiments. We use GPT-4o mini for both disambiguation and decontextualization.

3.2 Decomposition: $\mathcal{D}_{\text{R-ND}}$

Wanner et al. (2024) investigated various decomposition methods and their faithfulness, coherence, and atomicity. Here, faithfulness denotes if the subclaim is consistent with the original claim. Coherence ensures the subclaim is entailed by the original claim. Atomicity refers to how *much* a claim is broken down into components. They introduced a method grounded in Russellian and Neo-Davidsonian theory, $\mathcal{D}_{\text{R-ND}}$, and find this method to have the highest atomicity in comparison with other methods, while still remaining faithful and coherent to the original claim. Their prompt-based method

Claim: He first gained recognition in the mid-1990s for his starring role in the film "Schindler's List," directed by Steven Spielberg.

Decomposition	**Decomposition → Decontextualization**
He gained recognition in the mid-1990s.	Liam Neeson, the actor from Northern Ireland, gained recognition in the mid-1990s.
He gained recognition for his starring role.	Liam Neeson, the actor from Northern Ireland, gained recognition for his starring role in the film "Schindler's List."
His starring role was in the film Schindler's List.	Liam Neeson's starring role was in the film Schindler's List.
Schindler's List is a film.	"Schindler's List," directed by Steven Spielberg, is a film.
Steven Spielberg directed Schindler's List.	Steven Spielberg directed Schindler's List.
He gained recognition for his role in Schindler's List in the mid-1990s.	Liam Neeson gained recognition for his role in Schindler's List in the mid-1990s.

Decontextualized Claim	**Decomposition of Decontextualized Claim (Decontextualization → Decomposition)**
Liam Neeson first gained recognition in the mid-1990s for his starring role in the film "Schindler's List," directed by Steven Spielberg.	- Liam Neeson gained recognition in the mid-1990s. - Liam Neeson's recognition was for his role in Schindler's List. - Liam Neeson had a starring role in Schindler's List. - Schindler's List is a film. - Schindler's List was directed by Steven Spielberg. - Liam Neeson gained recognition for his role in Schindler's List in the mid-1990s. - Liam Neeson's recognition came after Schindler's List was released.

Decomposition and Decontextualization Jointly: DnD

Subclaims	*Decontextualized*
He gained recognition.	Liam Neeson gained recognition.
He gained recognition in the mid-1990s.	Liam Neeson gained recognition in the mid-1990s.
His recognition was for a starring role.	Liam Neeson's recognition was for a starring role.
His starring role was in the film 'Schindler's List.'	Liam Neeson's starring role was in the film 'Schindler's List.'
'Schindler's List' is a film.	'Schindler's List' is a film directed by Steven Spielberg.
'Schindler's List' was directed by Steven Spielberg.	'Schindler's List' was directed by Steven Spielberg.
The film 'Schindler's List' contributed to his recognition.	The film 'Schindler's List' contributed to Liam Neeson's recognition.
The time period of the mid-1990s refers to the years around 1995.	The time period of the mid-1990s refers to the years around 1995.

Table 1: Examples of each method of decomposition and decontextualization. Decomposition and DnD subclaim sets are similar, and the decontextualized sets are similar. Decomposition is done using the $\mathcal{D}_{\text{R-ND}}$ method described in section 3.2. We use the Molecular Facts method described in section 3.1 for decontextualization. We use DnD from section 3.3 for joint decomposition and decontextualization.

uses high quality in-context decomposition examples. We use $\mathcal{D}_{\text{R-ND}}$ for all decompositions in our experiments. We use `gpt-3.5-turbo-instruct` for decomposition, as used in Wanner et al. (2024).

3.3 Joint Decontextualization and Decomposition: DnD

Running decomposition and decontextualization in sequence can create problems. These two steps fundamentally interact, so considering each step independently poses limitations on the decisions made within the step. Additionally, running these in sequence doubles the number of LLM calls. Therefore, we develop a new method called DnD (Decontextualization and Decomposition) to obtain two sets of claims with just one LLM call. The first contains the decomposed subclaims of a sentence and the second set of claims contains the decontextualized form of each of the decomposed subclaims. Each claim in the subclaim set has a corresponding decontextualized claim in the decontextualized set. The prompt we developed for this method appears in Appendix A.3. We use GPT-4o mini for DnD.

3.4 Decomposition Evaluation

Previous work has proposed several different ways to measure the results of a decomposition method (Wanner et al., 2024). We use DECOMPSCORE, introduced by Wanner et al. (2024), which measures the average number of supported (highly correlated with NLI entailment) subclaims per passage produced. This metric indicates which method generates the most subclaims that cohere with the sentence being decomposed.

4 Methods: Claim Verification

We consider two decompose-then-verify methods for claim verification. First, we adopt the widely used FACTSCORE from previous work. We then propose a new method that utilizes decontextualization to better isolate the specific claim to verify.

4.1 FACTSCORE

Decompose-Then-Verify metrics have become increasingly popular, including the introduction of FACTSCORE (Min et al., 2023). FACTSCORE is an LLM-based fact-checking pipeline that verifies claims decomposed from a passage against a trusted reference source (e.g., Wikipedia), and the percentage of decomposed claims supported is the FACTSCORE. We use FACTSCORE as a baseline, however, it is insufficient for evaluating subclaims that require more context. We use Inst-LLAMA from FACTSCORE, which is a LLAMA 7B trained on the Super Natural Instructions dataset (Wang et al., 2022; Touvron et al., 2023), and run FACTSCORE with the Inst-LLAMA + retrieval + NPM setting.

4.2 DNDSCORE

When a claim is augmented with context, it contains multiple atomic facts. When passed to a verifier, such as FACTSCORE, which fact is being verified? How do we ensure that the same context is not re-verified across multiple claims?

We propose DNDSCORE, which modifies the prompt to include the source document, the subclaim, and the augmented, decontextualized claim. The method is asked to verify the specific subclaim using the relevant context against the source document. The prompt appears in Appendix A.5. We use the same Inst-LLAMA described above.

5 Results

5.1 Data

We use LLM generated biographies (FACTSCORE) and QA responses (LFQA) in our experiments.

FACTSCORE Data The released data from Min et al. (2023) includes generated biographies from 12 language models of varying sizes. Entities for these biographies range from very rare to very frequent and span different nationalities. We do not generate new or modify existing biographies. The entity Wikipedia page is the reference source.

LFQA Data The released data from Chen et al. (2023a) contains generated responses to ELI5 dataset (Fan et al., 2019), a dataset of dataset of threads from the Reddit forum "Explain Like I'm Five". Reference sources are retrieved by humans, models, or randomly selected. We use the test split of the data as collected by Tang et al. (2024).

5.2 Methods of Decomposition and Decontextualization

We use the decomposition ($\mathcal{D}_{\text{R-ND}}$) and decontextualization (Molecular Facts) methods (Section 3) in sequence to evaluate the interactions of these two methods. We then evaluate our method DnD, using one LLM call for joint decomposition and decontextualization. We evaluate these approaches using DECOMPSCORE and a qualitative analysis.

The DECOMPSCORE results are shown in Table 2, with full results in Table 10 in the Appendix. The DECOMPSCORE remains high for each of the methods, indicating a high number of facts entailed by the original sentence. Despite more information being added to decontextualized claims, we see only a small increase DECOMPSCORE. This suggests the added information for decontextualized claims is supported by the original claim, or a minimal enough addition that it does not change the entailment judgment.

Method	Avg DECOMP-SCORE (%)	Avg # Subclaims
Decomp	95.69	43.48
Decomp → Decontext	95.87	43.48
Decontext → Decomp	96.80	45.48
DnD Subclaim	96.14	36.03
DnD Decontext	96.58	36.03

Table 2: DECOMPSCORE results for each decomposition and decontextualization method on FACTSCORE data. We report the two sets of claims returned by our joint approach (DnD) alongside both sequential approaches. Scores remain high even with the use of decontextualization. Full results are in table 10.

Table 1 includes an example sentence and the result of each decomposition and decontextualization method. The decomposition and subclaims sets from DnD contain almost identical claims, indicating the new DnD is aligned with previously validated forms of decomposition. Similarly, the Decomposition → Decontextualization set, Decontextualization → Decomposition set, and the DnD decontextualized set all contain similar claims, although the latter two contain more facts. Because

the claim is decontextualized first in Decontextualization → Decomposition, the decontextualized subclaim decomposes into more subclaims than Decomposition → Decontextualization. The decontextualized claim only replaces the pronoun and does not add any extra external information, which is expected in a simple sentence like this one.

5.3 Fact Verification

We evaluate FACTSCORE using (1) decomposition only (as in the original paper (Min et al., 2023)), (2/3) using the decomposition and decontextualization methods in sequence (both decomposition → decontextualization and decontextualization → decomposition), and then the (4) subclaim set and (5) decontextualized set obtained from our DnD method. We then use our proposed fact verification method DNDSCORE, which requires context for each verified subclaim. When applying DNDSCORE to the decomposition-only method, we use the original sentence as context and the decomposition for the claim. For evaluating Decomposition → Decontextualization, we can use the pairs of decomposed subclaims and their corresponding decontextualized claim as context. We use the output of Decontextualization → Decomposition as the verified subclaim, and the decontextualized sentence as context. For DnD, we can just use the subclaim and corresponding decontextualized subclaim as context for verification.

FACTSCORE and DNDSCORE results are shown in Table 3, with full FACTSCORE results in Table 8 and full DNDSCORE results in Table 9 in the Appendix. The original FACTSCORE method yields an average score of 33.00% and 43.48 average subclaims per paragraph. By design, the Decomposition → Decontextualization method has the same number of subclaims, but almost 13% higher FACTSCORE. The Decontextualization → Decomposition generates the highest number of subclaims, with more information added into the original sentence in the decontextualization step, but still achieves around the same FACTSCORE as Decomposition → Decontextualization. The DnD subclaim set receives a similar FACTSCORE as decomposition only, and the DnD decontextualized set has a similar FACTSCORE as Decomposition → Decontextualization, but with fewer subclaims.

The DNDSCORE evaluation is on average higher than the FACTSCORE counterpart, while the number of subclaims is the same, by design. The DNDSCORE is highest for the Decontextualization → Decomposition as verified subclaim, and decontextualized claim as context at 61.51%. Verifying the decomposed subclaims with both the original sentence as context and the Decomposition → Decontextualization as context achieves similar scores, within five percent. Evaluating the DnD method subclaim and decontextualization sets with DNDSCORE results in a score between the others (51.60%.)

We show that factuality scores change when using FACTSCORE evaluated decontextualized subclaims instead of decomposed subclaims, however using only these decontextualized claims is insufficient. DNDSCORE addresses the ambiguity problems of using decomposed subclaims, and the redundancy and loss of atomicity problems of using decontextualized subclaims. The FACTSCORE and DNDSCORE evaluations in table 3 demonstrate that these factuality scores are changing. In the following section, we aim to understand *what* causes these changes.

6 Analysis

We provide quantitative analyses in section 6.1 in order to understand how, how much, and why judgments change. Qualitative analyses in section 6.2 look at specific examples and their corresponding judgments. We analyze redundancy in section 6.3, and there are further analyses on tradeoffs and failure cases in appendix section C.

6.1 Verification Changes: Quantitative Evaluation

Based on the change in factuality scores, we aim to understand: what makes a claim verifiable and what causes the changes of these factuality scores for the same passage? We perform analyses on FACTSCORE generated biography data. We examine the DNDSCORE evaluation, and DnD FACTSCORE evaluation of the subclaim set and decontextualization set. We specifically study these results, because we can align them at a subclaim level and examine the different subclaims and corresponding decontextualized claims evaluated by FACTSCORE, and the pair evaluated by DNDSCORE. Additionally, the decomposition only and the DnD subclaim set achieve a similar factuality score, and the Decomposition → Decontextualization achieves a similar factuality score as DnD decontextualized set, and therefore we expect these sets of subclaims to be similar.

Fact-uality Score	Decompose Method		FACTSCORE Data				LFQA Data			
			Avg Score (%)	Avg # Sub-claims	Avg Score + CORE (%)	Avg # Sub-claims + CORE	Avg Score (%)	Avg # Sub-claims	Avg Score + CORE (%)	Avg # Sub-claims + CORE
FS	DP Only		33.00	43.48	32.27	23.97	68.9	28.2	67.6	13.8
	DP → DT		45.97	43.48	43.51	20.93	74.9	28.2	72.2	12.4
	DT → DP		44.60	45.48	40.89	24.08	71.5	29.7	69.0	13.6
	DnD DP		35.92	36.03	35.43	22.39	70.4	26.9	68.9	12.3
	DnD DT		47.70	36.03	45.70	19.76	77.9	26.9	76.6	12.9
DS	Context	Verified Subclaim								
	Original Sent.	DP	41.53	43.48	41.38	23.97	79.2	28.2	81.8	13.8
	DP → DT	DP	46.44	43.48	46.80	23.97	87.7	28.2	89.2	13.8
	DT Sent.	DT → DP	61.51	45.48	59.18	24.08	80.2	29.7	82.9	13.6
	DnD DT	DnD DP	51.60	36.03	51.56	22.39	90.8	26.9	90.6	12.3

Table 3: Fact verification results on FACTSCORE and LFQA data using different combinations of decomposition and decontextualization, and with two fact verification methods: FACTSCORE (FS) and DNDSCORE (DS), a contextualized version of the former. We report the two sets of claims returned by our joint approach (DnD) alongside both sequential approaches (DP denotes decomposition, DT denotes decontextualization). These scores are averaged across different language model splits. The full data can be found in Tables 8 and 9 in the appendix. We additionally report the deduplicated FACTSCORE, DNDSCORE, and average number of subclaims for each method, filtered with CORE (Jiang et al., 2024).

We first examine the percentage of claims whose support judgment changed. We find that between DnD subclaim and DnD decontextualization FACTSCORE evaluated sets, there is a 19.11% change in judgment, and 16.25% change from false to true when decontextualizing. 48.52% of these contain a pronoun replacement[1], suggesting that judgements become verifiable due to entity disambiguation. DnD subclaim and DnD decontextualization FACTSCORE evaluation changes from true to false only 3.26% of the time, 11.82% of these containing pronoun replacement. Incorrectly verified claims were less of an issue.

Between the DnD subclaim FACTSCORE evaluations and DNDSCORE evaluations, 16.97% of subclaim judgments change, with 16.50% changing from false to true, and 0.48% changing from true to false. The addition of context rarely changes judgments to false, and instead only adds additional background or pronoun replacement that makes the claim easier to verify as correct. Human evaluation (section C.3) on a subset of the data support these findings. **DNDSCORE evaluation helps verify ambiguous claims, most often with pronoun replacement.**

[1]Subclaims which contain one or more common pronouns ("she", "her", "hers", "herself", "he", "him", "his", "himself", "they", "them", "theirs", and "themself"), which was not found in the decontextualized claim are indicated as having pronoun replacement.

6.2 Verification Changes: Qualitative Evaluation

Context is important for fact verification, shown in Table 4. Here we look at DnD subclaim and decontextualized subclaim pairs, each evaluated separately with FACTSCORE, and corresponding DNDSCORE judgments. We present examples where factuality judgments disagree, demonstrate where the addition of context helps disambiguate the claim, and show how DNDSCORE can handle these cases.

Example 1 in Table 4 shows an example where judgment changes from false to true, without pronoun replacement. "Prince Daniel" is a name which could refer to the current member of the Swedish royal family (born 1973), the prince of Galicia (1201–1264), the Russian prince from 1261–1303, or Prince Daniel of Saxony (born 1975). The added information to the verified subclaim disambiguates which "Prince Daniel" is the subject of the paragraph. Example 2 shows an example where FACTSCORE validation changes from false to true *with pronoun replacement*. The pronoun "She" is replaced with "Susan Sarandon" disambiguating the subject of the sentence. The DNDSCORE for the first two examples is also correctly judged as true with context.

Examples 3 and 4 in Table 4 had a FACTSCORE judgment changed from true to false. The third example, with no pronoun replacement, disambiguates the referenced sitcom, changing if the

Ex. #	Subclaims	FS	DS
1	**Subclaim**: Prince Daniel is a member of the Swedish royal family. **Decontextualized**: Prince Daniel, who is a sibling of Prince Carl Philip, is a member of the Swedish royal family.	False True	True
2	**Subclaim**: She has appeared in several television series. **Decontextualized**: Susan Sarandon has appeared in several television series.	False True	True
3	**Subclaim**: The sitcom featured Matthew Perry as a lead actor. **Decontextualized**: 'The Matt Payne Show' featured Matthew Perry as a lead actor.	True False	False
4	**Subclaim**: She wrote for the school's newspaper. **Decontextualized**: Nikole Hannah-Jones wrote for the newspaper of Wesleyan University.	True False	False
5	**Subclaim**: He began his wrestling career. **Decontextualized**: Fuerza Guerrera, also known as Juan Conrado Aguilar Jáuregui, began his wrestling career.	True False	True
6	**Subclaim**: A bit can be abbreviated. **Decontextualized**: A bit, which is the smallest unit of information in computing and digital communications, can be abbreviated.	False True	True
7	**Subclaim**: Some of these factors include infections. **Decontextualized**: One of the factors that may contribute to less saliva production and/or thickened saliva is infections.	False True	True

Table 4: Examples from FACTSCORE data (examples 1-5) and LFQA data (examples 6-7) of subclaims and decontextualized claims as generated by DnD, and FACTSCORE (FS) evaluations, and the DNDSCORE (DS) evalutation for the pair of claims. Context helps evaluate the factuality of claims, and DNDSCORE can handle cases where incorrect context is added to the text. Examples 1 and 3 show decontextualization with added information, and examples 2, 4, and 5 demonstrate pronoun replacement.

subclaim is verified[2]. The fourth example replaces "She" with "Nikole Hannah-Jones". Both examples are correctly judged as false using DNDSCORE.

Examples 6 and 7 in Table 4 both come from the LFQA dataset. In both examples, the added information helps verify the subclaim. The first example disambiguates what "a bit" refers to: the unit of information in computing as opposed to a small amount. The second example enumerates "the factors" which are unknown from just the subclaim, but interpretable with the decontextualized claim. In both these examples DNDSCORE correctly marks the claim as true, showing the effectiveness of DNDSCORE outside the human biographies domain.

There is a lot of movement between the FACTSCORE judgments of subclaims and the FACTSCORE judgments of decontextualized claims. The additional context can help, as in the examples shown, however, there exist examples where decontextualization is not sufficient. In some cases, added contextual information is wrong, and masks the correctness of the subclaim being verified. Example 5 in Table 4 concerns the Mexican wrestler Fuerza Guerrera where the atomic fact is true, but the information added to the decontextualized claim is false.[3] The FACTSCORE evaluation of the subclaim is true, but false for the decontextualized claim, because "Conrado Aguilar Jáuregui" refers to a different Mexican wrestler. Despite this incorrect addition to the decontextualized subclaim, DNDSCORE evaluates the subclaim as true. DNDSCORE verifies the subclaim, while still including the context. This ensures that the context helps to verify claims, but does not overshadow the subclaim being verified. By handling these nuances, DNDSCORE proves to be a more robust method for fact verification. **DNDSCORE helps both confirm true subclaims and refute false subclaims originally judged incorrectly by FACTSCORE.**

6.3 CORE for Subclaim Deduplication

CORE (Jiang et al., 2024) filters sets of decomposed subclaims and formats subclaim subselection as a constrained optimization problem to eliminate duplicated facts, which we apply at a generation level. Deduplication results are in Table 3. We use this filtering to analyze the different methods of decontextualization, decomposition, and verification, however, it can (and should) be used to filter sets

[2]Matthew Perry did not appear in 'The Matt Payne Show.'

[3]Juan Conrado Aguilar Jáuregui is the name of a different wrestler, and is not another name for Fuerza Guerrera.

of subclaims for verification to ensure scores are not inflated with many duplicated facts.

The average set of filtered subclaims for all decomposition and decontextualization methods is between 19 and 25. Filtered subclaim size has less variance because CORE filters to a *core* set of subclaims. The DnD decontextualized set had more subclaims removed than the DnD subclaim set, indicating a more redundant decontextualized set. Figure 1 shows that the decontextualized form of subclaims is less atomic and includes information from other subclaims. The redundant subclaims are removed in the deduplication process.

Filtering the subclaim DnD set results in the least number of facts, and decontextualize-then-decompose have the most remaining subclaims, due to added context having slightly different wording, and not fully being filtered out. For example, in the set of decontextualized-then-decomposed subclaims from Vicuna-7B on Dr. Dre, multiple claims stating something equivalent to "Dr. Dre was a rapper" were kept in the filter process, despite being semantically equivalent. This is an example of the repeated essential context added to each decontextualized sentence in the paragraph before decomposing as shown in the bottom right of Figure 1. The DnD sets contained fewer subclaims, and thus had fewer subclaims removed during filtering. FACTSCORE changes are within two percent for all deduplicated subclaim sets, except for decontextualize-then-decompose, which drops almost four percent. As shown in Figure 1, this method is at risk of containing duplicated essential context across sentences.

7 Conclusion

In this work, we consider the interactions of decomposition and decontextualization in a fact verification system. We introduce DnD, a prompt-based method for extracting subclaims and corresponding decontextualized forms. Using pairs of subclaims and decontextualized claims, we propose a new decompose-then-verify method, DNDSCORE, which validates claims with a given context. We demonstrate cases where verification judgment differs, and show DNDSCORE is able to handle context better than previous measures of factuality.

Limitations

Although we demonstrate the robustness of DNDSCORE in handling ambiguities, which is especially important in verification of generations with many entities, we only consider its use in fact verification of generated biographies and QA settings. Our work does not show the use of DNDSCORE in any other domains.

DNDSCORE is not intended to handle debatable claims, such as opinions. Previous work has examined factuality scores of only verifiable claims (Song et al., 2024), which could be integrated into DNDSCORE, such that DNDSCORE only evaluates these verifiable claims.

The reference documents are static in this study, but using decontextualized subclaims for reference document retrieval is a possible line of future work, that would allow this method to work on arbitrary generations, instead of generations on a predetermined topic.

A important limitation of our evaluation is the underlying dataset, a set of biographies for which the correct source (Wikipedia) documents are available for verification. In a production system, we may find source document selection errors due to the information retrieval process. We hypothesize that these will increase verification confusions, and the differences between our approaches and previous work will grow. Additionally, while biographies center on a single individual, other generation types, e.g. news updates, may include multiple entities and increase verification confusion.

This work is only done in the language of English, although we expect these results would hold in other languages.

Ethics Statement

LLMs are prone to hallucination, which can result in the spread of misinformation. Mitigating these hallucinations is important and still actively being researched. The decomposition and decontextualization methods in this work are at risk of injecting hallucinated information not in the original claim. Evaluation of these untrue generations are necessary to ensure the reliability of them, and care should be taken when trusting these generations.

References

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2023a. Understanding retrieval augmentation for long-form question answering.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification

with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

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, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2023b. Sub-sentence encoder: Contrastive learning of propositional semantic representations. *arXiv preprint arXiv:2311.04335*.

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.

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.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in llm fact verification.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2024. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance?

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2024. Core: Robust factual precision with informative subclaim identification.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models.

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.

Benjamin Kane and Lenhart Schubert. 2023. Get the gist? using large language models for few-shot decontextualization.

Yoonsang Lee, Xi Ye, and Eunsol Choi. 2024. Ambigdocs: Reasoning across documents on different entities under the same name.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

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.

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*, pages 3194–3212, Singapore. Association for Computational Linguistics.

Abhilash Potluri, Fangyuan Xu, and Eunsol Choi. 2023. Concise answers to complex questions: Summarization of long-form answers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9709–9728, Toronto, Canada. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan

Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models.

A Prompts

A.1 Decomposition Prompt

We use the $\mathcal{D}_{\text{R-ND}}$ prompt from Wanner et al. (2024). Their prompt uses dynamically retrieved in-context decompositions.

A.2 Decontextualization Prompt

We use the Molecular Facts prompt from Gunjal and Durrett (2024). They use a two step prompting methods, with the first prompt identifying ambiguities in a claim, and the second prompt using the identified ambiguities to decontextualize the claim.

A.3 DnD Prompt

The DnD prompt for extracting pairs of subclaims and their decontextualized form can be found in Tables 5-7. We use the ambiguity criteria outlined in the Molecular Facts prompt and decomposed in-context examples from $\mathcal{D}_{\text{R-ND}}$.

A.4 FACTSCORE Prompt

We use the FACTSCORE prompt from Min et al. (2023) shown below. This prompt provides a reference document and asks if an atomic fact is true or false.

```
Answer the question about [TOPIC] based on the given
context.

Title: [REFERENCE DOC SECTION TITLE]
Text: [REFERENCE DOC CONTENT]

Input: [ATOM] True or False?
Output:
```

A.5 DNDSCORE Prompt

The following is the DNDSCORE prompt used for factuality verification of claims given the decontextualized form of the claim. This prompt is adapted from FACTSCORE, but includes the decontextualized claim as context for the atomic claim.

```
Answer the question about [TOPIC] based on the given
reference document and context.

Reference Document:
[REFERENCE DOC]

Given the following context: "[DECONTEXT CLAIM]"
Input: Is "[ATOM]" True or False?
Output:
```

A.6 Compute

Decomposition and decontextualization experiments were run on a GPU cluster with Quadro RTX 6000. We estimate experiments took around 400 GPU-hours.

B Full Results

B.1 FACTSCORE Results

The full FACTSCORE results for each language model split from the FACTSCORE data are in Table 8.

B.2 DNDSCORE Results

The full DNDSCORE results for each language model split from the FACTSCORE data are in Table 9.

B.3 DECOMPSCORE Results

The full DECOMPSCORE results for each language model split from the FACTSCORE data are in Table 10.

C Further Analysis

C.1 Tradeoffs

Due to the longer prompt lengths in DNDSCORE, the cost of using this method is slightly higher. Based on the data of this paper, running FACTSCORE on one passage averages a total of 78,057 input tokens, with an average total of 89,335 input tokens for DNDSCORE. FACTSCORE uses about 6,212 input tokens on atomic fact generation and 71,845 on evaluation. DNDSCORE uses about 15,878 input tokens on DnD and 73,457 on evaluation. This difference is minimal, and less than a cent difference per passage. The tokens from the knowledge sources in the evaluation step overshadow the difference in token length for atomic fact generation in FACTSCORE, and the respective DnD in DNDSCORE. The prompt and response for DnD is longer than the FACTSCORE decomposition prompt, but there are also fewer subclaims extracted, and thus fewer calls to the verifier model. Although DNDSCORE is more expensive it crucially disambiguates cases with many entities, which is necessary for verification. The computational overhead of DNDSCORE is marginally more than FACTSCORE, but results in disambiguated passage verification.

C.2 Failure Cases

The only case where DNDSCORE may introduce error is when the FACTSCORE evaluation on the subclaim is true, FACTSCORE evaluation on the decontextualized subclaim is false, and DNDSCORE evaluation is true. In this case, the decontextualized information is incorrect, but the DNDSCORE

```
Ambiguity Criteria: Ambiguity manifests in diverse forms, including:
- Similar names denoting distinct entities.
- Varied interpretations stemming from insufficient information.
- Multiple understandings arising from vague or unclear information.

Instructions:
- You are given a paragraph, and one sentence from the paragraph to decompose and decontextualize.
- First decompose the sentence into subclaims. Only use information from the sentence, and do not
add any external information.
- Then using those subclaims, write a decontextualized version of each subclaim.
- In the decontextualized version, include all necessary information to disambiguate any entities
or events in the subclaim using the ambiguity criteria above.
- In the decontextualized version, only use information from the paragraph. Do not add any external
information.
- Provide an explanation of what ambiguities need to be resolved

Format your response as a combination of decomposition and a dictionary with pairs of context and
subclaims:
##PARAGRAPH##: <paragraph>
##SENTENCE##: <sentence>
##SUBCLAIMS##:
<list-of-subclaims>
##EXPLANATION##:
<explanations>
##CONTEXT-SUBCLAIM PAIRS##:
[
    {"subclaim": <subclaim1>, "decontextualized": <context1>},
    {"subclaim": <subclaim2>, "decontextualized": <context2>},
    ...
]

Example 1:
##PARAGRAPH##: Michael Collins (born October 31, 1930) is a retired American astronaut and test
pilot who was the Command Module Pilot for the Apollo 11 mission in 1969. He orbited the Moon
in the command module Columbia while Neil Armstrong and Buzz Aldrin made their historic landing.
Born in Rome, Italy, Collins graduated from the U.S. Military Academy in 1952, joining a family
tradition of military service, and went on to become a test pilot in the U.S. Air Force. Selected
as an astronaut in 1963, he flew two space missions, Gemini 10 in 1966 and Apollo 11 in 1969,
making him one of only 24 people to travel to the Moon. Collins was an accomplished astronaut,
becoming the fourth person to conduct a spacewalk and the first to perform multiple spacewalks.
After leaving NASA in 1970, he served as Assistant Secretary of State for Public Affairs, later
directing the National Air and Space Museum. He also held senior roles at the Smithsonian and in
private aerospace, eventually founding his own consulting firm. Collins and his Apollo 11 crewmates
received the Presidential Medal of Freedom in 1969 and the Congressional Gold Medal in 2011.
##SENTENCE##: Michael Collins (born October 31, 1930) is a retired
American astronaut and test pilot who was the Command Module Pilot
for the Apollo 11 mission in 1969.
##SUBCLAIMS##:
- Michael Collins was born in October.
- Michael Collins was born on the 31st day of a month.
- Michael Collins was born in 1930.
- Michael Collins is retired.
- Michael Collins is American.
- Michael Collins was an astronaut.
- Michael Collins was a test pilot.
- Michael Collins participated in the Apollo 11 mission.
- Michael Collins's participation in the Apollo 11 mission occurred in 1969.
- The Apollo 11 mission was active in 1969.
- The day of Michael Collins's birth occurred before his year of participation
in the Apollo 11 mission.
- The Apollo 11 mission had a Command Module Pilot.
- Michael Collins's role in the Apollo 11 mission was as the Command Module Pilot.
##EXPLANATION##:
"Michael Collins" needs to be disambiguated as the astronaut associated with the
Apollo 11 mission to distinguish him from other potential individuals with
similar names.
```

Table 5: 1/3 of the DnD method for extracting subclaims and corresponding decontextualized subclaims. More details can be found in appendix section A.3.

method is still able to verify the subclaim as correct. We find that in our experiments this is a rare occurrence, only happening 3% of the time. We expect this proportion may change with different data.

C.3 Human Evaluation

We perform human annotation on a small 100 instance subset of the LFQA data, specifically where FACTSCORE judgments flip (i.e. subclaim is true and decontextualized subclaim is false, or vice versa). We assess the groundedness with respect to the grounding document, which we assume is fac-

```
##CONTEXT-SUBCLAIM PAIRS##:
[
    {"subclaim": "Michael Collins was born in October.",
     "decontextualized": "Michael Collins, the retired American astronaut and test pilot, was born
     in October."},
    {"subclaim": "Michael Collins was born on the 31st day of a month.",
     "decontextualized": "Michael Collins, the retired American astronaut and test pilot, was born
     on the 31st day of a month."},
    {"subclaim": "Michael Collins was born in 1930.",
     "decontextualized": "Michael Collins, the retired American astronaut and test pilot, was born
     in 1930."},
    {"subclaim": "Michael Collins is retired.",
     "decontextualized": "Michael Collins, the retired American astronaut and test pilot, is
     retired."},
    {"subclaim": "Michael Collins is American.",
     "decontextualized": "Michael Collins, the American astronaut, is American."},
    {"subclaim": "Michael Collins was an astronaut.",
     "decontextualized": "Michael Collins, the retired American astronaut and Command Module Pilot
     for the Apollo 11 mission, was an astronaut."},
    {"subclaim": "Michael Collins was a test pilot.",
     "decontextualized": "Michael Collins, the retired American astronaut and test pilot, was the
     Command Module Pilot for the Apollo 11 mission in 1969."},
    {"subclaim": "Michael Collins participated in the Apollo 11 mission.",
     "decontextualized": "Michael Collins, the retired American astronaut and test pilot,
     participated in the Apollo 11 mission."},
    {"subclaim": "Michael Collins's participation in the Apollo 11 mission occurred in 1969.",
     "decontextualized": "Michael Collins's participation in the Apollo 11 mission as the Command
     Module Pilot occurred in 1969."},
    {"subclaim": "The Apollo 11 mission was active in 1969.",
     "decontextualized": "The Apollo 11 mission, which involved human spaceflight to the Moon, was
     active in 1969."},
    {"subclaim": "The day of Michael Collins's birth occurred before his year of participation in
    the Apollo 11 mission.",
     "decontextualized": "The day of Michael Collins's birth on October 31, 1930, occurred before
     his year of participation in the Apollo 11 mission."},
    {"subclaim": "The Apollo 11 mission had a Command Module Pilot.",
     "decontextualized": "The Apollo 11 mission had Michael Collins as its Command Module Pilot."},
    {"subclaim": "Michael Collins's role in the Apollo 11 mission was as the Command Module
    Pilot.",
     "decontextualized": "Michael Collins's role in the Apollo 11 mission was as the Command Module
     Pilot."}
]

Example 2:
##PARAGRAPH##: Stephen Miller (born August 23, 1985) is an American political advisor who served
as a senior advisor for policy and director of speechwriting to President Donald Trump. Miller has
been described as the architect of Trump's controversial immigration policies, and has previously
worked for Alabama Senator Jeff Sessions on immigration issues. Miller was instrumental in shaping
several of Trump's key policies, including the travel ban, a reduction in refugee admissions, and
family separations at the border. He began his career in communications roles for conservative
legislators, including Senators Jeff Sessions, Michele Bachmann, and John Shadegg. As Trump's
speechwriter, Miller helped draft the inaugural address and served as a trusted advisor from the
early days of the administration. He also played a significant role in the resignation of Secretary
of Homeland Security Kirstjen Nielsen, whom he deemed insufficiently strict on immigration. As a
White House spokesperson, Miller made several unsubstantiated claims about election fraud and
promoted content from white nationalist sources, leading to his inclusion on the Southern Poverty
Law Center's list of extremists.
```

Table 6: 2/3 of the DnD method for extracting subclaims and corresponding decontextualized subclaims. More details can be found in appendix section A.3.

tual. We provide the annotator with (1) the query, (2) the passage, (3) the subclaims, (4) the decontextualized subclaims, and (5) the grounding document. So as not to bias the annotator, we do not provide the FACTSCORE or DNDSCORE judgments. For each subclaim and decontextualized subclaim we ask the annotator:

1. Given the grounding document, is the subclaim from the passage True or False? (options: True/False/Not enough information)

2. Given the grounding document, is the decontextualized subclaim from the passage True or False? (options: True/False/Not enough information)

3. Does the subclaim contain enough information to correctly judge it as true/false? (options: Yes/No)

4. Does the decontextualized subclaim contain enough information to correctly judge it as true/false? (options: Yes/No)

We find that the subclaim FACTSCORE and annotator agree 16% of the time, and the decontextu-

```
##SENTENCE##: Miller has been described as the architect of Trump's controversial immigration
policies, and has previously worked for Alabama Senator Jeff Sessions on immigration issues.
##SUBCLAIMS##:
- Miller has been described.
- Miller has been described as an architect.
- Miller has been described as an architect of Trump's controversial
immigration policies.
- Trump has immigration policies.
- Trump's immigration policies are controversial.
- Miller worked for Jeff Sessions.
- Jeff Sessions is a Senator.
- Jeff Sessions represents Alabama.
- Miller worked on immigration issues.
- Miller's work for Jeff Sessions involved immigration issues.
##EXPLANATION##:
"Miller" needs to be disabiguated as Stephen Miller, a political advisor for Donald Trump, to avoid
confusion with other individuals with the same name. Clarify that "Trump's immigration policies"
refers specifically to policies developed during Donald Trump's presidency, as "Trump" alone may be
ambiguous in a different context.
##CONTEXT-SUBCLAIM PAIRS##:
[
    {"subclaim": "Miller has been described.",
     "decontextualized": "Miller, the architect of Trump's controversial immigration policies, has
     been described."},
    {"subclaim": "Miller has been described as an architect.",
     "decontextualized": "Miller, who has been described as the architect of Trump's controversial
     immigration policies, has been described as an architect."},
    {"subclaim": "Miller has been described as an architect of Trump's controversial immigration
    policies.",
     "decontextualized": "Stephen Miller has been described as an architect of Trump's
     controversial immigration policies."},
    {"subclaim": "Trump has immigration policies.",
     "decontextualized": "Donald Trump has immigration policies."},
    {"subclaim": "Trump's immigration policies are controversial.",
     "decontextualized": "Donald Trump's immigration policies are controversial."},
    {"subclaim": "Miller worked for Jeff Sessions.",
     "decontextualized": "Miller, the architect of Trump's controversial immigration policies,
     worked for Jeff Sessions."},
    {"subclaim": "Jeff Sessions is a Senator.",
     "decontextualized": "Jeff Sessions is a Senator from Alabama."},
    {"subclaim": "Jeff Sessions represents Alabama.",
     "decontextualized": "Jeff Sessions represents the state of Alabama."},
    {"subclaim": "Miller worked on immigration issues.",
     "decontextualized": "Miller, the architect of Trump's controversial immigration policies,
     worked on immigration issues."},
    {"subclaim": "Miller's work for Jeff Sessions involved immigration issues.",
     "decontextualized": "Stephen Miller's work for Jeff Sessions involved immigration issues."},
]

Your task:
##PARAGRAPH##: [paragraph]
##SENTENCE##: [sentence]
##SUBCLAIMS##:
```

Table 7: 3/3 of the DnD method for extracting subclaims and corresponding decontextualized subclaims. More details can be found in appendix section A.3.

alized subclaim FACTSCORE and annotator agree 77% of the time (from question 1/2). The DND-SCORE and annotator agree 76% of the time (from question 1). From question two, the percentage of subclaims with enough information to judge are 54% and decontextualized subclaims with enough information to judge are 88%, indicating that the decontextualization step is important. These results indicate that when FACTSCORE judgments flip, the original subclaim tends to be ambiguous, and the decontextualization step crucially disambiguates the subclaim.

Decompose Method	LM Split	FACTSCORE (%)	# Subclaims
Decomp Only	Alpaca 7B	35.0	22.2
	Alpaca 13B	38.9	22.0
	Alpaca 65B	44.0	22.2
	ChatGPT	48.2	44.2
	Dolly 12B	16.5	33.0
	GPT-4	51.1	77.7
	InstructGPT	40.1	36.3
	MPT-Chat 7B	24.8	49.0
	Oasst-pythia 12B	20.1	57.7
	StableLM 7B	13.8	40.4
	Vicuna 7B	32.4	59.8
	Vicuna 13B	31.1	57.3
Decomp → Decontext	Alpaca 7B	49.8	22.2
	Alpaca 13B	52.8	22.0
	Alpaca 65B	60.6	22.2
	ChatGPT	66.6	44.2
	Dolly 12B	22.3	33.0
	GPT-4	71.8	77.7
	InstructGPT	57.0	36.3
	MPT-Chat 7B	33.7	49.0
	Oasst-pythia 12B	29.4	57.7
	StableLM 7B	17.8	40.4
	Vicuna 7B	46.5	59.8
	Vicuna 13B	43.4	57.3
Decontext → Decomp	Alpaca 7B	48.3	24.2
	Alpaca 13B	51.2	23.7
	Alpaca 65B	57.2	24.1
	ChatGPT	61.6	45.5
	Dolly 12B	24.1	34.7
	GPT-4	64.6	79.9
	InstructGPT	53.5	38.4
	MPT-Chat 7B	33.9	50.8
	Oasst-pythia 12B	31.0	59.1
	StableLM 7B	21.4	42.5
	Vicuna 7B	45.1	61.2
	Vicuna 13B	43.3	61.7
DnD Subclaim	Alpaca 7B	36.8	20.7
	Alpaca 13B	41.3	20.0
	Alpaca 65B	46.5	20.2
	ChatGPT	53.6	37.9
	Dolly 12B	18.1	30.1
	GPT-4	56.3	61.8
	InstructGPT	43.4	31.7
	MPT-Chat 7B	26.5	40.6
	Oasst-pythia 12B	21.5	43.1
	StableLM 7B	15.1	33.5
	Vicuna 7B	35.9	47.5
	Vicuna 13B	36.0	45.2
DnD Decontextualized	Alpaca 7B	52.0	20.7
	Alpaca 13B	54.4	20.0
	Alpaca 65B	62.3	20.2
	ChatGPT	69.2	37.9
	Dolly 12B	23.8	30.1
	GPT-4	73.7	61.8
	InstructGPT	59.8	31.7
	MPT-Chat 7B	34.9	40.6
	Oasst-pythia 12B	29.7	43.1
	StableLM 7B	18.0	33.5
	Vicuna 7B	48.3	47.5
	Vicuna 13B	46.3	45.2

Table 8: The FACTSCORE for different language model splits, the results of which are aggregated in Table 3. The factuality ranking of these language models stays consistent, however the scores differ.

Decompose Method	LM Split	DNDSCORE (%)	# Subclaims
Context: Original Sentence **Verified Subclaim:** Decomp	Alpaca 7B	43.6	22.2
	Alpaca 13B	46.8	22.0
	Alpaca 65B	51.0	22.2
	ChatGPT	53.8	44.2
	Dolly 12B	25.6	33.0
	GPT-4	57.8	77.7
	InstructGPT	47.8	36.3
	MPT-Chat 7B	36.8	49.0
	Oasst-pythia 12B	31.5	57.7
	StableLM 7B	22.0	40.4
	Vicuna 7B	41.7	59.8
	Vicuna 13B	40.0	57.3
Context: Decomp → Decontext **Verified Subclaim:** Decomp	Alpaca 7B	49.9	22.2
	Alpaca 13B	52.7	22.0
	Alpaca 65B	57.9	22.2
	ChatGPT	58.9	44.2
	Dolly 12B	30.6	33.0
	GPT-4	62.6	77.7
	InstructGPT	53.9	36.3
	MPT-Chat 7B	40.7	49.0
	Oasst-pythia 12B	33.8	57.7
	StableLM 7B	25.3	40.4
	Vicuna 7B	46.0	59.8
	Vicuna 13B	45.0	57.3
Context: Decontext Sentence **Verified Subclaim:** Decontext → Decomp	Alpaca 7B	68.2	24.2
	Alpaca 13B	68.8	23.7
	Alpaca 65B	73.1	24.1
	ChatGPT	69.6	45.5
	Dolly 12B	46.5	34.7
	GPT-4	73.5	79.9
	InstructGPT	70.2	38.4
	MPT-Chat 7B	55.8	50.8
	Oasst-pythia 12B	54.8	59.1
	StableLM 7B	38.9	42.5
	Vicuna 7B	62.4	61.2
	Vicuna 13B	56.3	61.7
Context: DnD Decontextualized **Verified Subclaim:** DnD Subclaim	Alpaca 7B	53.9	20.7
	Alpaca 13B	57.0	20.0
	Alpaca 65B	61.5	20.2
	ChatGPT	65.4	37.9
	Dolly 12B	37.2	30.1
	GPT-4	69.4	61.8
	InstructGPT	58.6	31.7
	MPT-Chat 7B	44.9	40.6
	Oasst-pythia 12B	38.3	43.1
	StableLM 7B	29.4	33.5
	Vicuna 7B	51.7	47.5
	Vicuna 13B	51.9	45.2

Table 9: The DNDSCORE for different language model splits, the results of which are aggregated in Table 3. The factuality ranking of these language models stays consistent, however the scores differ.

Decompose Method	LM Split	DECOMPSCORE (%)	# Subclaims
Decomp Only	Alpaca 7B	98.7	22.2
	Alpaca 13B	98.6	22.0
	Alpaca 65B	98.6	22.2
	ChatGPT	93.0	44.2
	Dolly 12B	97.4	33.0
	GPT-4	96.2	77.7
	InstructGPT	98.1	36.3
	MPT-Chat 7B	96.5	49.0
	Oasst-pythia 12B	98.3	57.7
	StableLM 7B	89.2	40.4
	Vicuna 7B	94.8	59.8
	Vicuna 13B	88.9	57.3
Decomp → Decontext	Alpaca 7B	99.1	22.2
	Alpaca 13B	99.0	22.0
	Alpaca 65B	98.9	22.2
	ChatGPT	92.9	44.2
	Dolly 12B	97.8	33.0
	GPT-4	96.3	77.7
	InstructGPT	98.6	36.3
	MPT-Chat 7B	96.3	49.0
	Oasst-pythia 12B	98.5	57.7
	StableLM 7B	88.1	40.4
	Vicuna 7B	95.3	59.8
	Vicuna 13B	89.7	57.3
Decontext → Decomp	Alpaca 7B	99.0	24.2
	Alpaca 13B	99.1	23.7
	Alpaca 65B	99.1	24.1
	ChatGPT	95.1	45.5
	Dolly 12B	98.1	34.7
	GPT-4	96.8	79.9
	InstructGPT	98.5	38.4
	MPT-Chat 7B	97.0	50.8
	Oasst-pythia 12B	98.5	59.1
	StableLM 7B	92.4	42.5
	Vicuna 7B	96.4	61.2
	Vicuna 13B	91.6	61.7
DnD Subclaim	Alpaca 7B	98.9	20.7
	Alpaca 13B	99.0	20.0
	Alpaca 65B	99.1	20.2
	ChatGPT	92.7	37.9
	Dolly 12B	98.1	30.1
	GPT-4	96.0	61.8
	InstructGPT	98.9	31.7
	MPT-Chat 7B	97.1	40.6
	Oasst-pythia 12B	99.1	43.1
	StableLM 7B	89.8	33.5
	Vicuna 7B	95.6	47.5
	Vicuna 13B	89.4	45.2
DnD Decontextualized	Alpaca 7B	99.4	20.7
	Alpaca 13B	99.3	20.0
	Alpaca 65B	99.4	20.2
	ChatGPT	94.0	37.9
	Dolly 12B	98.6	30.1
	GPT-4	96.4	61.8
	InstructGPT	99.4	31.7
	MPT-Chat 7B	96.9	40.6
	Oasst-pythia 12B	99.1	43.1
	StableLM 7B	88.4	33.5
	Vicuna 7B	96.7	47.5
	Vicuna 13B	91.3	45.2

Table 10: The DECOMPSCORE for different language model splits, the results of which are aggregated in Table 2. The DECOMPSCORE remains high despite additional new information for decontextualized subclaims.