Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification

Anisha Gunjal **Greg Durrett**
The University of Texas at Austin
anishagunjal@utexas.edu

Abstract

Automatic factuality verification of large language model (LLM) generations is becoming more and more widely used to combat hallucinations. A major point of tension in the literature is the granularity of this fact-checking: larger chunks of text are hard to fact-check, but more atomic facts like propositions may lack context to interpret correctly. In this work, we assess the role of context in these atomic facts. We argue that fully atomic facts are not the right representation, and define two criteria for *molecular facts*: decontextuality, or how well they can stand alone, and minimality, or how little extra information is added to achieve decontexuality. We quantify the impact of decontextualization on minimality, then present a baseline methodology [1] for generating molecular facts automatically, aiming to add the right amount of information. We compare against various methods of decontextualization and find that molecular facts balance minimality with fact verification accuracy in ambiguous settings.

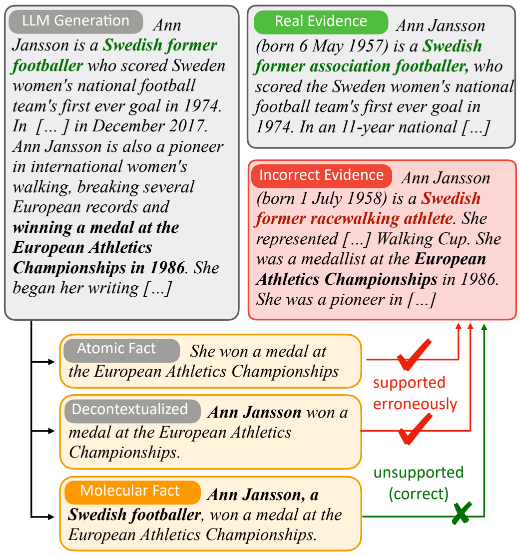

Figure 1: Breaking a paragraph into atomic facts can cause errors in attribution: facts out of context appear to be true when they are not. The right granularity of decontextualization, "molecular facts," balances contextual grounding with atomicity.

1 Introduction

Large language models (LLMs) have emerged as powerful tools for delivering knowledge to users, either via closed-book generation or retrieval-augmented systems. However, these systems may not always produce correct facts (Liu et al., 2023a), an instance of the "hallucination" problem (Zhang et al., 2024; Ji et al., 2022; Zhang et al., 2023). Recent research has shown the potential of LLMs to identify unfaithful content and enable automatic fact-checking and attribution against sources (Falke et al., 2019; Goyal and Durrett, 2021; Min et al., 2023; Wang et al., 2024; Chern et al., 2023; Wei et al., 2024; Chen et al., 2023a; Malaviya et al., 2024; Gao et al., 2023b; Tang et al., 2024).

A key step in this process is to break down generated content into individual atomic claims (Fabbri et al., 2022; Chen et al., 2023b; Kamoi et al., 2023b; Min et al., 2023). This decomposition allows for retrieval of evidence focused on a particular part of the generated content (Gao et al., 2023a; Wang et al., 2024; Chen et al., 2024) and also error localization by determining which parts of the content are supported or not. However, this step is not straightforward. Wanner et al. (2024) highlights that the effectiveness of automatic factuality verification is heavily dependent on the strategies employed for decomposing content into claims. In particular, LLMs have a propensity to incorrectly merge information about similarly named entities Lee et al. (2024) and current evaluation methods struggle to handle these ambiguities in atomic claims (Chiang and yi Lee, 2024). Figure 1 shows a possible issue: a fact that is "too atomic" can be validated against evidence that doesn't actually support it.

[1] https://github.com/anisha2102/molecular_facts

Findings of the Association for Computational Linguistics: EACL 2024, pages 3751–3768
November 12-16, 2024 ©2024 Association for Computational Linguistics

In this work, we address the problem of how to find minimal yet still unambiguous facts for LLM fact verification. We frame this problem as one of *decontextualization*, adding context to a sentence to make it stand alone while retaining its original meaning (Choi et al., 2021). This process draws on the idea of *specificity* from discourse (Louis and Nenkova, 2012), specifically whether sentences can express key information about the participants without ambiguity (Li et al., 2016). However, making a claim unambigous is not enough: when escalating from simple pronoun replacement in atomic facts to elaborations like *a Swedish footballer* in Figure 1, we must balance the specificity of the fact with how easy it will be to verify. It is not trivial to select the "right" information to elaborate on a claim without compromising the ease of verification.

We define two criteria needed in this fact-checking setting: *decontextuality*, where the claim should uniquely specifying entities, events, and context, and *minimality*, maintained by avoiding excessive additional information that could complicate verification. We propose a notion of *molecular facts*, which balances these two criteria: molecular facts should be fully specific while compatible with the maximum number of possible evidence documents. We explore these criteria and our molecular facts in two settings. First, we address the question of how much *non-minimality* could be a problem for error localization with standard decontextualization techniques. We devise a synthetic fact-checking experiment where particular nuances of an output generation are unsupported and show that an average of 6% of claims may pose problems for error localization. In a setting with LLM responses of 5 sentences with 3 claims each, this would lead to localization errors in a large fraction of responses. We then evaluate the opposite problem, whether decontextualization is *too minimal*. We study a dataset of fact-checking with ambiguous entity names presented in Chiang and yi Lee (2024). We show that our method of molecular fact generation balances accuracy under ambiguous entity references with minimality of claims.

Our main contributions are: (1) We re-examine the decontextualization process for fact-checking and define *molecular claims* following the desiderata of decontextuality and minimality. (2) We investigate the loss of minimality due to claim decontextualization and its impacts on error localization. (3) We find that molecular claims are more performant and minimal for long-form generations than existing decontextualization methods.

2 Desiderata for Decontextualization

We propose desiderata to determine the optimal level of decontextualization required for atomic facts. An *atomic fact* is defined as a discrete unit of information, derived from a broader claim, and variously described in the literature as propositions, subclaims, summary content units, or atomic content units (Nenkova and Passonneau, 2004; Liu et al., 2023b; Zhang and Bansal, 2021; Chen et al., 2023b; Min et al., 2023; Kamoi et al., 2023b).

Although an atomic fact theoretically represents a singular conceptual unit, recent NLP work using this does not typically give this a rigorous definition from the standpoint of semantics. Wanner et al. (2024) demonstrate a high variation in the number of subclaims generated by different decomposition methods, with the macro-average of subclaims per biography ranging from 20.2 using the method by Kamoi et al. (2023b) to 32.9 with the approach by Chen et al. (2023b). Note that in Figure 1, *She was a medallist at the European Athletics Championships in 1986* could be kept as one unit or broken into three facts evaluating her status as a medallist, the venue, and the date.

2.1 Desiderata

Preliminaries We define $\mathbf{r}$ as a response from a language model to an input prompt $\mathbf{x}$, consisting of a series of claims $(\mathbf{c}_1, \ldots, \mathbf{c}_n)$ to be verified. Claims are extracted through an upstream process of decomposition and potentially filtering for "check-worthiness" (i.e., does the claim present factual content or does it present an opinion?). We describe the prompting in Appendix A.

We assume that in the context of $\mathbf{r}$ and $\mathbf{x}$, a claim $\mathbf{c}_i$ can be fully interpreted with a truth-conditional meaning $I(\mathbf{c}_i \mid \mathbf{x}, \mathbf{r})$. In the terminology of Rashkin et al. (2021) and Choi et al. (2021), $I(\mathbf{c}_i \mid \mathbf{x}, \mathbf{r})$ represents $\mathbf{c}_i$ interpreted in the *linguistic context* of $\mathbf{x}$ and $\mathbf{r}$.

We can construct a *standalone proposition* with truth conditional meaning equivalent to I by being sufficiently specific. For example, the statement in Figure 1 could be completely specified as *Ann Jansson, the Swedish footballer born on 6 May 1957 who played for Hammarby IF, won a medal at the European Athletics Championship, the biennial event organized by the European Athletics Association, in 1986.*

Decontextualization Our goal in this work is to produce rewritten *molecular claims*. Denote by $\mathbf{m}_i$ the rewritten form of $\mathbf{c}_i$, which should have semantics I when interpreted as a standalone proposition. As in Figure 1, this requires adding disambiguating information that could provide information needed to identify an entity (specifying that Jansson is a *Swedish footballer*), identify an event (specifying that the event happened in 1986), specify a qualification (*in the field of biochemistry, ...*), or more.

Criterion 1 (Decontextuality) *When interpreted as a standalone statement,* $\mathbf{m}_i$ *must have the truth conditional meaning* $I(\mathbf{c}_i, \mathbf{x}, \mathbf{r})$. *That is, it should uniquely specify entities, events, and other context such that the claim* $\mathbf{c}_i$ *is now interpretable.*

This criterion is equivalent to Definition 1 from Choi et al. (2021). For the settings we consider, the level of added information needed to specify the meaning of a statement like that in Figure 1 may be higher than in past applications like Choi et al. (2021). It is not sufficient to replace the pronoun *she* with *Ann Jansson*; we need to specify *Ann Jansson, the Swedish footballer*. Similarly, the city *George Town* could refer to a city in the Cayman Islands or Malaysia, therefore it must be decontextualized appropriately with a descriptor like *George Town, a city in Cayman Islands*.

Other work such as question answering frameworks based on clarifying questions can target this information (Newman et al., 2023), but may fail to integrate the minimal new information needed, which we describe next.

Minimality Adding too much information to a claim makes it less minimal. For instance, replacing "*Ann Jansson*" with "*Ann Jansson, a Swedish footballer*" requires verifying that a context referring to Ann Jansson is indeed talking about the Swedish footballer. Taken further, the reference "*Ann Jansson, the Swedish footballer born on 6 May 1957 who played for Hammarby IF*" is clearly suboptimal. It requires verifying Jansson's birthdate as an additional detail, and crucially, this detail won't be frequently reported in documents about Ann Jansson.

Define $\mathcal{E}^*(I(\mathbf{c}, \mathbf{x}, \mathbf{r}))$ as the set of set of evidence documents that support the statement I with an *oracle* understanding of the entities involved. For instance, this would contain a document describing the correct Ann Jansson, even if it did not confirm all the details about her life. Define $\mathcal{E}(\mathbf{m}_i) \subset \mathcal{E}^*$ to be the set of evidence documents that fully support a statement $\mathbf{m}_i$. For instance, in the case of Ann Jansson above, the document would need to specify Jansson's birthdate if this is contained in $\mathbf{m}$.

Criterion 2 (Minimality) *Given a set of statements* $\mathcal{M}$ *that all decontextualize a claim* $\mathbf{c}_i$*, we should select* $\operatorname{argmax}_{\mathbf{m} \in \mathcal{M}} |\mathcal{E}(\mathbf{m})|$ *to maximize the size of the set of supporting evidence documents.*

This criterion means that, when selecting distinguishing details for an entity, we should choose those that can typically be inferred from evidence. For instance, "*Jason Martin*" may be characterized either as a "*rugby player*" or specifically as a "*former player for North Queensland Cowboys*." Since "*rugby player*" is a more enduring and widely recognized description, yet still specific enough to indicate Jason Martin, it is more likely to be supported by a larger number of documents.

Past work like Choi et al. (2021) instructs annotators to make minimal edits to statements. However, they do not provide guidance on what criteria should be used to choose from among multiple candidate edits.

Molecular facts These two criteria suggest two things. First, atomic facts can be "too atomic:" they may need to be decontextualized. However, it is still valuable to have a reasonably minimal fact so it can be supported by many possible evidence documents.

Molecular Fact *A molecular fact is a statement* $\mathbf{m}_i$ *corresponding to claim* $\mathbf{c}_i$ *that obeys criteria 1 and 2: it should uniquely specify the interpretation of* $\mathbf{c}_i$ *even when considered on its own, while adding as little information as possible to do so.*

2.2 Task Definition: Fact-checking LLMs

Recall our setting where an LLM has generated a response $\mathbf{r}$ to input prompt $\mathbf{x}$, and $\mathbf{r}$ has associated claims $(\mathbf{c}_1, \ldots, \mathbf{c}_n)$. For each $\mathbf{c}_i$, we have a corresponding set of k evidence documents, $D_i = (D_{i,1}, \ldots, D_{i,k})$, that are referenced to assess the accuracy of $\mathbf{c}_i$. Furthermore, we have access to a gold standard of human-annotated labels for each atomic fact, represented as $L = (l_1, \ldots, l_n)$, where each l_i can be either `SUPPORTED` or `NOT_SUPPORTED`. **Our goal is to make judgments about the supportedness of the** $\mathbf{c}_i$, which requires appropriately decontextualizing each fact.

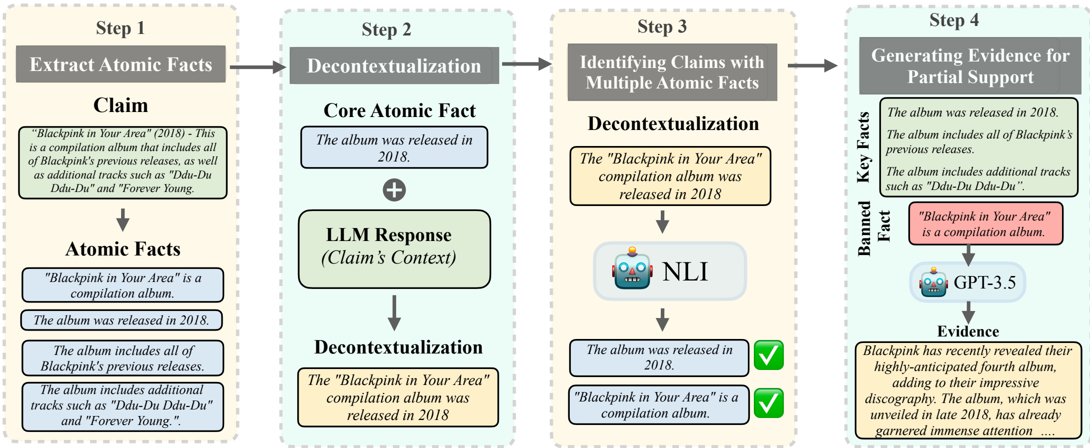

Figure 2: Controlled evidence generation framework for illustrating error localization introduced by decontextualization for atomic fact verification.

We augment each atomic claim $\mathbf{c}_i$ to a corresponding molecular claim $\mathbf{m}_i$ as described in Section 3, resulting in a set of facts $(\mathbf{m}_1, \ldots \mathbf{m}_n)$. We represent the model's factuality judgment prediction as a set of supported documents $p_i = \text{Check}(D_{i,j}, \mathbf{m}_i)$, for all $j \in \{1, 2, \ldots, n\}$ in D_i. In other words, the prediction of Check() is accurate when it supports the molecular claim with the same evidence docs as humans.

3 Method: Producing Molecular Facts

We use a two-step process to refine an atomic fact into a molecular fact using `gpt-4-turbo-2024-04-09` (Achiam et al., 2023). Our methodology makes the assumption that the ambiguity is typically restricted to a single entity in the claim. This is the case for the datasets we study in this work, described in Section 4.5.

Stage 1: Identifying Ambiguity We identify the primary subject of the claim and to assess potential ambiguities based on its parametric knowledge: does the model know of multiple entities with this name? This step identifies the main subject $\mathbf{s}_i$ of the claim $\mathbf{c}_i$ and provides a disambiguation criteria $\mathbf{b}_i$ for the subject $\mathbf{s}_i$. The disambiguation criteria $\mathbf{b}_i$ can be 'None' when there is no ambiguity, or a type of criteria such as profession, birthyear, or location when disambiguation is required.

For example, if the claim is about 'Charles Osgood', with multiple possible referents, $\mathbf{s}_i$ is *'Charles Osgood'*, while $\mathbf{b}_i$ could be *'profession'* or *'birthyear'* to clarify which Charles Osgood is being referred to. Conversely, if the claim concerns the unambiguous '*Julius Robert Oppenheimer*', $\mathbf{s}_i$ is *'Julius Robert Oppenheimer'*, and $\mathbf{b}_i$ is *'None'*.

Stage 2: Molecular Facts Generation We then prompt the LLM to disambiguate the subject $\mathbf{s}_i$ within the claim $\mathbf{c}_i$, harnessing both the identified disambiguation criteria $\mathbf{b}_i$ and the claim's context $\mathbf{r}$.The output of this stage is a molecular fact $\mathbf{m}_i$ for the atomic claim $\mathbf{c}_i$.

The specifics regarding the prompts used are elaborated upon in Appendix 6 and 7.

3.1 Baselines

We analyze the robustness of fact verification across various systems on the defined criteria of *minimality* and *decontextuality*. Outputs for baselines are generated with `gpt-4-turbo-2024-04-09`.

`ATOMIC`: Atomic claims are generated from the LLM's response using Min et al. (2023).

`SIMPLE-DECONTEXT`: Atomic claims are decontextualized with a prompt described in 8 using the LLM's generated response as context for the atomic claim.

`SAFE-DECONTEXT`: Decontextualization of atomic claims is performed using the revision prompt described in Wei et al. (2024).

`MOLECULAR-DECONTEXT`: This approach follows a two-stage process described in section 3 to identify disambiguation criteria and subsequently decontextualize the atomic claim.

Examples of outputs from each method can be found in Figure 3. With this task definition and baseline methodologies, we structure our experiments to analyze the two criteria presented in Section 2.1 in the following sections.

4 Experiment: Minimality & Localization

We begin our analysis of decontextualization with a controlled experiment to illustrate problems with error localization due to loss of minimality discussed in Criterion 2 in Section 2.1. Minimality is more difficult to evaluate than decontextuality. Less minimal facts impact error localization and can potentially lead to errors where an ancillary part of the claim leads to the whole claim being judged as wrong (Kamoi et al., 2023a). However, precisely measuring the harms of this is not easy without taking into account the downstream uses of error localization systems such as answer refinement (Xu et al., 2023) or fine-tuning (Wu et al., 2024; Roit et al., 2023).

To measure the effects in a controlled way, we design a method for synthetic evidence generation as summarized in Figure 2. **Our goal is to illustrate when decontextualized atomic facts actually contain *multiple* facts in a way that could impact error localization.** We then study how many of these cases truly show this problem. To study the impact of information addition, we consider two baselines SIMPLE-DECONTEXT and SAFE-DECONTEXT which respectively have less and more restrictive prompts for including new information from the context to revise an atomic claim.

4.1 Controlled Dataset Construction

We now detail the dataset construction process as illustrated in Figure 2. We take a dataset D of 812 claims from the Factcheck-Bench dataset (Wang et al., 2024) which consists of long form ChatGPT responses with human-annotated factuality labels.

Step 1: Extract Atomic Facts For each response $\mathbf{r} \in D$, we extract atomic facts $(\mathbf{c}_1, \dots, \mathbf{c}_n)$ using the method of Min et al. (2023).

Step 2: Decontextualization: We perform decontextualization of the extracted atomic facts using SIMPLE-DECONTEXT and SAFE-DECONTEXT. Let the decontextualization for claim $\mathbf{c}_i$ be denoted as $\mathbf{d}_i$. We refer to the $\mathbf{c}_i$ that $\mathbf{d}_i$ was created from as its *core atomic fact*; however, note that $\mathbf{d}_i$ might support other facts as well.

Step 3: Identifying Claims with Multiple Atomic Facts: We identify decontextualized claims that entail information of more than one atomic fact. We use the entailment model from Liu et al. (2022) to determine $e(\mathbf{d}_i, \mathbf{c}_j) \in \{\text{supported}, \text{unsupported}\}$; is each $\mathbf{c}_j$ supported by $\mathbf{d}_i$? We retain cases where $e(\mathbf{d}_i, \mathbf{c}_i) = \text{supported}$ and where $|\{j : e(\mathbf{d}_i, \mathbf{c}_j) = \text{supported}\}| \geq 2$; that is, at least two atomic facts are supported by $\mathbf{d}_i$. For example, in Figure 2, the claim $(\mathbf{d}_i)$, *'The "Blackpink in Your Area" compilation album was released in 2018'*, is a decontextualized claim derived from the core atomic claim $(\mathbf{c}_i)$, *'The album was released in 2018.'*. The decontextualized claim $(\mathbf{d}_i)$ entails the core atomic fact $(\mathbf{c}_i)$ and an additional atomic fact $(\mathbf{c}_j)$ *' "Blackpink in Your Area" is a compilation album'*. Let D' denote this filtered set.

Step 4: Generating Evidence for Partial Support: Whenever multiple atomic facts are merged, we could *theoretically* see a loss in localization capability from a model: if one fact is not supported, the entire claim will be determined to be not supported. To demonstrate this possibility, we now **generate** evidence that partially supports our multi-fact claims. As an example, in Figure 2, our goal in step 4 is to generate a paragraph that *should not* include details about "Blackpink in Your Area" being a compilation album. Then, if the statement *'The album was released in 2018'* is decontextualized to include information about it being a compilation album, this paragraph will enable us to identify this: the evidence will no longer support the decontextualized fact, reflecting a failure of error localization.

By construction of D', $\mathbf{d}_i$ is supported by at least two facts, its core atomic fact and auxiliary atomic fact(s). From this set of auxiliary atomic fact(s), we sample a *banned fact* $\mathbf{c}_b$. For each $\mathbf{d}_i$, we sample a set of *key facts* $C_i = \{\mathbf{c}_{i,1}, \dots, \mathbf{c}_{i,m}\}$ such that C_i contains the all atomic facts of the response r except $\mathbf{c}_b$. We then prompt the LLM to generate an evidence article supporting the facts C_i and not supporting the fact c_j. Each of these evidence articles ideally should support *all* the key facts and not support the *banned fact*.

The prompt for this step is detailed in Figure 10 and other filtering criteria are described in Appendix F. Denote this set where evidence generation is feasible as E'.

4.2 Evaluation Criteria

We evaluate the impacts of loss of minimality on the recall of fact-checking. We measure the percentage of cases that change their label from SUPPORTED to NOT_SUPPORTED after decontextualization on the set E'. We employ the `roberta-large` from AlignScore (Zha et al., 2023) as our $\text{Check}()$ function.[2]

[2]We conducted preliminary analysis with GPT-4 as well, and found it gave very similar results.

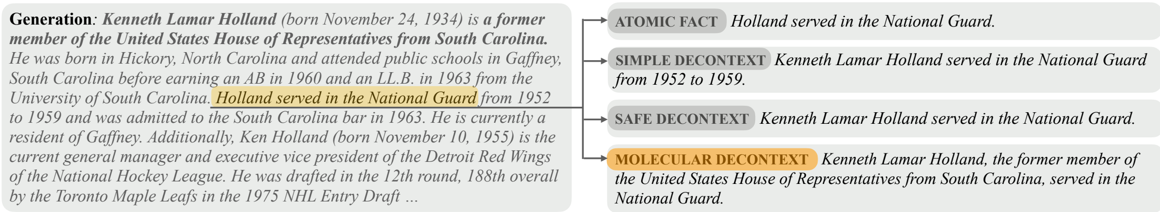

Figure 3: Example claims (right) generated by SIMPLE-DECONTEXT, SAFE-DECONTEXT, MOLECULAR-DECONTEXT for the atomic claim derived from the highlighted sentence in the LLM generation (left).

Baseline	Potential Non-minimal	Auto Non-minimal
SAFE-DECONTEXT	8.49%	3.94%
SIMPLE-DECONTEXT	23.39%	13.42%

Table 1: Percentage of overall dataset impacted by minimality loss due to decontextualization leading to prediction changes from SUPPORTED to NOT_SUPPORTED.

Category	Minimal	Non-minimal
SAFE-DECONTEXT	56.2%	43.8%
SIMPLE-DECONTEXT	27.5%	72.5%

Table 2: Human annotation for categorizing the Auto Non-minimal subset into minimal vs. non-minimal.

Using $\mathrm{Check}(\mathrm{D_i}, \mathrm{c_i})$, we identify cases where the *core key fact* is SUPPORTED by the generated evidence while the *decontextualization* and *banned fact* are NOT_SUPPORTED. We call this set *auto non-minimal*.

4.3 Results

Table 1 shows the fraction of claims which are included in the set E', which yields 8.49% for SAFE-DECONTEXT and 23.39% for SIMPLE-DECONTEXT. We refer to these claims as *potential non-minimal claims*: they have passed the checks in our pipeline and contain multiple atomic facts. Next we apply the $\mathrm{Check}()$ function to identify auto non-minimal claims, and find that they occur at a rate of 3.94% to 13.42% (Table 1).

4.4 Human Evaluation

Susceptibility to Error Localization We perform human evaluation on the *auto non-minimal* claims in Table 1. First, we categorize these into human judgments of whether a claim in this subset is minimal or not in Table 2. We categorize a decontextualization as minimal based on the criteria outlined in 2.1. This annotation is performed by the authors of the paper. We find that for SAFE-DECONTEXT, 43.8% of these cases are truly non-minimal in our judgment which represent 1.7% of the dataset D. For the SIMPLE-DECONTEXT baseline, we find that a staggering 72.5% of the auto non-minimal subset represents truly non-minimal claims. This represents 9.6% of the dataset D. We note that the remaining fraction of decontextualization cases not identified by the auto methods are those which entail more than one atomic fact but it is a necessary addition to make the atomic claim standalone.

Decontextualization and Loss of Minimality We highlight that addition of information to a claim does not always make it less entailed to the evidence. In fact, in many cases information addition makes the sentence more specific. This is evident from Table 2 which shows that automatically flagged cases for non-minimality have a large percentage of minimal claims after human evaluation. For instance, "*All taxes must be paid by April 15*" $\rightarrow$ "*In the US, all taxes must be paid by April 15*" is a necessary addition for claim specificity.

4.5 Conclusion: Problem of Non-minimality

We find through our controlled experiment and human evaluation that decontextualization can lead to non-minimal cases for between 1.7% to 9.6% of decontextualizations. These cases could cause error localization issues due to too much information added to the claims. In absolute terms, this is a low fraction for the baseline SAFE-DECONTEXT. However, we note that a biography from FActScore (Min et al., 2023) contains dozens of atomic facts, meaning that in a single response from an LLM, there can easily be a handful of facts posing localization problems. Given the increasing adoption of the decomposition and decontextualization pipeline for automatic fact verification systems, we argue that multiple localization errors per response is cause to re-examine that pipeline. Next, we analyze tradeoffs between minimality and decontextuality for fact checking of ambiguous biographies.

Subset	ACCURACY OVERALL	ACCURACY SUPPORTED	ACCURACY NOT_SUPPORTED	MODIFICATION RATE	AVG LENGTH (# of words)
ATOMIC	68.7%	77.5%	22.4%	-	7.61±3.03
SIMPLE-DECONTEXT	76.2%	84.3%	33.6%	99.5%	15.55±5.65
SAFE-DECONTEXT	73.4%	81.3%	31.9%	72.6%	9.86±4.38
MOLECULAR-DECONTEXT	74.7%	81.5%	38.8%	96.8%	14.96±5.6

Table 3: Accuracy measured by $\mathrm{Check}(D_i, m)$, assessing the effectiveness of claim revisions by each baseline against the ambiguous document set associated with claim's main entity.

Human Label→	SUPPORTED			NOT_SUPPORTED	
Baseline Pred→	SUPPORTED	SUPPORTED	NOT_SUPPORTED	SUPPORTED	
Matching Type→ **Baseline** ↓	Multi-Evidence matched	Single-Evidence Wrong Entity	No Evidence matched	Single/Multiple Evidence matched	**Overall** ↓
ATOMIC	16.2%	0.8%	1.8%	12.4%	31.1%
SIMPLE-DECONTEXT	7.9%	1.5%	3.9%	10.6%	23.8%
SAFE-DECONTEXT	12.0%	1.0%	2.8%	10.9%	26.6%
MOLECULAR-DECONTEXT	9.2%	1.5%	4.8%	9.8%	25.3%

Table 4: Fine-grained error analysis categorizing baseline mistakes based on human label of SUPPORTED/NOT_SUPPORTED along with categorization of <Single/Multi/No>-Evidence based on the number of ambiguous evidence docs that support the claim.

5 Experiment: Ambiguous Biographies

We now analyze to what extent our molecular facts add the correct information to decontextualize on an existing dataset with ambiguous entity references.

Dataset We use the ambiguous biographies dataset introduced in Chiang and yi Lee (2024) which comprises biographies generated by LLMs for multiple entities that share similar names, such as *Dick Hanley (swimmer)* and *Dick Hanley (footballer)*. In this dataset we represent the biographies generated by the LLMs as $\mathbf{r}$ and $\mathbf{c}_i$ correspond to atomic claims generated using the methodology outlined in (Min et al., 2023). For this setting, we define each claim to have a subject $\mathbf{s_i}$, which is ambiguous due to the nature of the dataset. The dataset provides a set of evidence documents sourced from Wikipedia page of the subject disambiguation, $D_i = \{D_{i,2}, D_{i,2}, ...\}$ for subjects sharing similar names as $\mathbf{s}_i$. This dataset is suitable for evaluating *decontextuality* as it consists of two properties: (i) atomic claims that require decontextualization (such as entity specification, noun completion), (ii) multiple entities with the same name that require additional disambiguation such as specifying location, occupation, or time-period.

Our goal is to verify the claims with the set of documents using $\mathrm{Check}()$. We randomly sample 726 claims from the human-annotated set for this study which belong to either SUPPORTED or NOT_SUPPORTED categories. For each claim we construct a revision using the methods and baselines described in section 3 and compare the prediction with human labels.

Baseline	Minimal ↑	Non-Minimal↓	Ambig.↓
SIMPLE	16.0%	56.0%	28.0%
SAFE	24.0%	0.0%	76.0%
MOLECULAR	52.0%	24.0%	24.0%

Table 5: Human analysis of decontextualized claims for all baselines on the axis of minimality and ambiguity.

Evaluation Criteria We evaluate our judgment of a claim on two axes: (1) whether it aligns with the human annotation of SUPPORTED or NOT_SUPPORTED, and (2) whether it is supported by the correct evidence. For each evidence associated with the claim, we compute $p_{i,k} = \mathrm{Check}(D_{i,k}, c_i)$ where c_i is the claim processed by the particular baseline and k represents the kth ambiguous subject related document for the claim. We consider the judgment $p_{i,k}$ to be correct only if the prediction of the claim matches the human label *and* the prediction is supported by the correct entity's evidence document.

6 Results: Ambiguous Biographies

Table 3 presents the results of this experiment. All methods of decontextualization baselines yield higher accuracy rates compared to atomic claims, across all subsets. We see that Molecular and Simple decontextualization methods have a higher proclivity to modify the atomic claims than the SAFE decontextualization baseline. Consequently, the average sentence lengths of the former methods is

Baseline Pair	Overlap
ATOM & SIMPLE-DECONTEXT	7%
ATOM & SAFE-DECONTEXT	44%
ATOM & MOLECULAR-DECONTEXT	15%
SIMPLE-DECONTEXT & SAFE-DECONTEXT	27%
SIMPLE-DECONTEXT & MOLECULAR-DECONTEXT	36%
MOLECULAR-DECONTEXT & SAFE-DECONTEXT	32%

Table 6: Information overlap between baselines as measured by bi-directional entailment.

also larger than the SAFE baseline. Higher degrees of modification generally lead to higher accuracy. All three methods are on a Pareto frontier of length versus accuracy.

However, accuracy using the $\mathrm{Check}()$ function does not incorporate minimality. We investigate the minimality of the baselines by performing a human evaluation of randomly sampled 25 claims in Table 5. We see that the baseline SIMPLE-DECONTEXT has a large fraction of non-minimal and ambiguous claims as compared to MOLECULAR-DECONTEXT. Analysis in Section 4.4 shows that SAFE-DECONTEXT is more minimal than SIMPLE-DECONTEXT; however, it struggles with ambiguity.

Overall, we observe that molecular claims strike a balance by maintaining minimality with ambiguity removal and improving accuracy. They are significantly more minimal than SIMPLE-DECONTEXT and more performant in ambiguous generations than SAFE-DECONTEXT.

Error breakdown To analyze the nature of errors encountered, we detail a case-wise error distribution in Table 4. Specifically, we study the behavior of various baselines to mispredict the label as SUPPORTED or NOT_SUPPORTED in comparison to human annotation. Note that due to the ambiguous nature of this dataset, claims may be erroneously validated by several distracting pieces of evidence. Therefore, we further partition the error analysis table to reflect the model's prediction on (i) `Single/Multi/No Evidence`: whether a claim is supported by single, multiple, or no pieces of evidence, and (ii) `(Correct/Wrong Entity)`: whether the set of supporting evidence contains the accurate evidence with which the claim ought to be aligned. Overall, all decontextualization methods show a lower error rate than atomic claims.

Information Overlap We perform an information overlap analysis shown in Table 6 using the model from Liu et al. (2022) to check bidirectional entailment of the fraction of cases where the information is equivalent between two baselines (Gunjal and Durrett, 2023). We find in a large fraction of cases each baseline adds different information to modify the atomic claim. SAFE-DECONTEXT has least amount of modification albeit suffers with ambiguity and SIMPLE-DECONTEXT has most amount of modification at the cost of minimality loss.

7 Related Work

Recent research in factuality verification of LLM generations advocates decomposing LLM generations into atomic facts or subclaims and verifying each against retrieved evidence (Min et al., 2023; Kamoi et al., 2023b; Fabbri et al., 2022). End-to-end pipelines for factuality verification have been proposed, involving steps such as claim extraction, revision, determining checkworthiness, evidence retrieval, and verification (Wang et al., 2024; Chern et al., 2023; Wei et al., 2024; Chen et al., 2024). These papers often evaluate on recently-released datasets of errors in generations Liu et al. (2023a); Malaviya et al. (2024); Chen et al. (2023a). Our work comments on the decontextualization step frequently used in these pipelines.

Our work fits into a broader ecosystem of techniques in this area. Gao et al. (2023b) enable LLMs to generate text with citations. For faithful LLM generations, Gao et al. (2023a) use evidence retrieval for revision, and He et al. (2022) utilize chain-of-thought coupled with retrieval for faithful explanations. Fine-tuned systems, such as that by Zha et al. (2023), predict alignment scores for verification, while Tang et al. (2024) propose LLM-AggreFact for sentence-level factuality labels. Wanner et al. (2024) find that evaluation metrics for fact verification are sensitive to the claim decomposition method used.

Prior work on decontextualization has investigated basic notions like anaphora resolution (Choi et al., 2021), question answering frameworks (Newman et al., 2023), and extract-then-decontextualize methods for summarization (Potluri et al., 2023). In fact verification, atomic claims are made standalone before evidence retrieval via decontextualization (Wang et al., 2024) or claim revision (Wei et al., 2024). Decontextualization is also used to resolve ambiguity Zhang and Choi (2021); Lee et al. (2024); our work shares this focus.

8 Conclusion

We introduce molecular facts and the desiderata of decontextualization in LLM fact verification. We define the criteria of decontextuality and minimality in this context. Through a controlled experiment, we show that localization errors due to loss of minimality by decontextualization is sensitive to the method used. We propose a method of "molecular facts" and find that they improve fact verification precision for claims from generation about ambiguous entities. We show that molecular facts strike a balance between maintaining minimality and accuracy of fact-verification.

Limitations

Scope We illustrate the phenomenon of ambiguity in atomic claims; however, our main evaluation of molecular facts is in the domain of English-language biographies. This is due to the availability of the dataset, Wikipedia evidence, and the prevalence of biography benchmarks in recent work. Conceptually, the ambiguity in the subject or predicate of the claim can be extended to other realistic datasets, but we leave that exploration to future work. Relatedly, we focus on entity ambiguity for illustration of our method. There may be other types of ambiguities that molecular fact generation can address in other contexts and other datasets.

Furthermore, we focus our experiments on high-performing LLMs in this work. The extension of decontextualization and molecular fact generation to smaller, open-source models and the improvement in this regime is a good subject for further study.

Finally, we believe our approach should be evaluated fully end-to-end in an LLM pipeline that generates responses and then verifies their factuality. However, despite substantial research in these directions, we are not aware of an off-the-shelf experimental pipeline that is usable for this setting.

Decomposition Quality We do not consider the errors introduced due to poor decomposition of atomic facts in this work. It is possible that some of these errors are resolved due to decontextualization or disambiguation implicitly, but we do not make any specific claims about this.

Coverage of Domains and Languages The datasets utilized for ambiguous biographies are limited to English-language claims focused on English-centric concepts within Wikipedia. Similarly, the synthetic data generation experiment for minimality analysis is confined to English language outputs and relies on GPT-4's parametric knowledge, which may limit the breadth of topics and domains covered.

Acknowledgments

We thank Jessy Li for comments and feedback on the initial draft of this work. This work was supported by NSF CAREER Award IIS-2145280 and the NSF AI Institute for Foundations of Machine Learning (IFML). This material is also based on research that is in part supported by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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 North American Chapter of the Association for Computational Linguistics*.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023a. FELM: Benchmarking Factuality Evaluation of Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528*.

Cheng-Han Chiang and Hung yi Lee. 2024. Merging Facts, Crafting Fallacies: Evaluating the Contradictory Nature of Aggregated Factual Claims in Long-Form Generations. *arXiv 2402.05629*.

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.

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

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

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 *Empirical Methods in Natural Language Processing (EMNLP)*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of NAACL*.

Anisha Gunjal and Greg Durrett. 2023. Drafting Event Schemas using Language Models. *arXiv preprint arXiv:2305.14847*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Ryo Kamoi, Tanya Goyal, and Greg Durrett. 2023a. Shortcomings of question answering based factuality frameworks for error localization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 132–146.

Ryo Kamoi, Tanya Goyal, Juan Rodriguez, and Greg Durrett. 2023b. 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.

Yoonsang Lee, Xi Ye, and Eunsol Choi. 2024. Ambigdocs: Reasoning across documents on different entities under the same name. *arXiv 2404.12447*.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3921–3927.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. 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. 2023b. 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.

Annie Louis and Ani Nenkova. 2012. A corpus of general and specific sentences from news. In *LREC*, volume 1818, page 10. Citeseer.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the North American Chapter of the 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.

Ani Nenkova and Rebecca J 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.

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.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal nlp. *arXiv preprint arXiv:2306.05240*.

Abhilash Potluri, Fangyuan Xu, and Eunsol Choi. 2023. Concise answers to complex questions: Summarization of long-form answers. *arXiv preprint arXiv:2305.19271*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and D. Reitter. 2021. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49:777–840.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Frank Schilder. 1998. An underspecified segmented discourse representation theory (usdrt). In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1188–1192.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. *arXiv preprint arXiv:2404.10774*.

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. *arXiv preprint arXiv:2403.11903*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2023. Pinpoint, not criticize: Refining large language models via fine-grained actionable feedback. *arXiv preprint arXiv:2311.09336*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Michael J.Q. Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How Language Model Hallucinations Can Snowball. In *Forty-first International Conference on Machine Learning*.

Shiyue Zhang and Mohit Bansal. 2021. Finding a Balanced Degree of Automation for Summary Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Prompts

We give details on all the prompts used throughout this work.

Decontextuality Experiment Prompts The stepwise molecular facts generation prompts for `MOLECULAR_DECONTEXT` are in Figure 6, 7. For the simple decontextualization baseline `SIMPLE_DECONTEXT`, the prompts are provided in 8.

Minimality Experiment Prompts The prompt for generating controlled evidence for the minimality experiment is given in Figure 10.

B Additional Related Work

Decomposition in Text Summarization Decomposition of responses is also prevelant in the text summarization literature. Nenkova and Passonneau (2004) introduced the Pyramid protocol for summarization evaluation which extracts weighted

Summarization Content Units (SCUs) which represent the importance of various facts present in multiple human-generated summaries of a text. Zhang and Bansal (2021) propose using Semantic Triplet Units (STUs), which are summary content units generated automatically using SRL parsers, to evaluate generated summaries with textual entailment models. Similarly, Liu et al. (2023b) propose Atomic Content Units (ACUs) as a new summarization salience protocol that allows for higher inter-annotator agreement. Chen et al. (2023b) propose using entailment judgments on a set of sentence propositions within a document.

Decontextualization and Specificity Decontextualization is a process of making sentences stand-alone by resolving missing context while preserving its meaning (Choi et al., 2021). A related phenomenon is the notion of *specificity*. Louis and Nenkova (2012) presented the first corpus of sentences distinguished on the criteria of being *general* or *specific*. Their idea of classification was based on examples and intuition by defining *general* sentences to be broad statements about a topic that would need additional evidence or examples for a reader to understand, whereas, *specific* sentences can stand by themselves. Li et al. (2016) make this definition more specific by grounding specificity for a sentence to three requirements: (i) it is easy to understand the meaning and identify of the intended references without ambiguity; (ii) the truth of the statement can be assessed based on the sentence itself and general shared knowledge; and (iii) the sentence fully expresses key information about the participants and causes of an event. Another related notion is underspecification in discourse, which is an intentional feature to maintain communication efficiency (Schilder, 1998). This has been annotated by Li et al. (2016) and highlighted in a multimodal setting by Pezzelle (2023).

C Human Annotation Criteria for Categorizing the Non-minimal Subset

We describe the criteria for annotating the auto non-minimal subset into minimal vs. non-minimal as shown in Table 2. For each instance, we compare the original claim, the decontextualization, and the banned fact. We label cases as *minimal* when either of the following applies: (1) the banned fact is closely related the atomic fact and it is a necessary addition to the atomic claim to make it standalone. In other words, the banned fact is a necessary addition to the atomic claim to add context and/or resolve ambiguity. For example, "*The album is their first full-length studio album.*" is decontextualized to "*The album released in 2020 is Blackpink's first full-length studio album.*" and the banned fact is "*The album was released in 2020.*". The information in the banned fact is necessary addition to disambiguate "*the album*" in this case. (2) The banned fact entailed by the decontextualization, but it is due to an entailment error. For example, the decontextualization *"Mey Eden, one of the largest bottled water companies in Israel, offers flavored water products."* is erroneously entailed by the banned fact *"Mey Eden offers still water products."*.

D Human Analysis Criteria for Categorizing Minimality and Ambiguity

We describe the criteria for the human analysis for on the decontextualization of each baseline on the axis of minimality and ambiguity shown in Table 5. We categorize a claim decontextualization as *non-minimal* when it contains additional information that goes beyond making the sentence stand-alone and can potentially cause loss of error-localization. We categorize a claim decontextualization as *ambiguous* when it lacks clarifications for entities that could refer to different ambiguous subjects or add enough context to disambiguate the main entity. If both of the above conditions are not violated, we categorize the decontextualization as *minimal*.

E Models, Datasets and Computation Cost

The `gpt-4-turbo-2024-04-09` model was employed for running baselines and generating outputs, while the `gpt-3.5-turbo` model was used for evaluation through FActScore (Achiam et al., 2023). For generation experiments, we set the temperature to 0.75. The total cost for generating decontextualizations and evaluating the ambiguous biography experiment was approximately $120.

In the minimality experiment, `gpt-3.5-turbo` was used to extract atomic facts, and `gpt-4-turbo-2024-04-09` was used for decontextualization and generation tasks. This resulted in a total cost of around $100. We use a NVIDIA A40 GPU for evaluation using AlignScore (Zha et al., 2023) and entailment computation using WANLI (Liu et al., 2022),

We use ChatGPT for improving writing format-

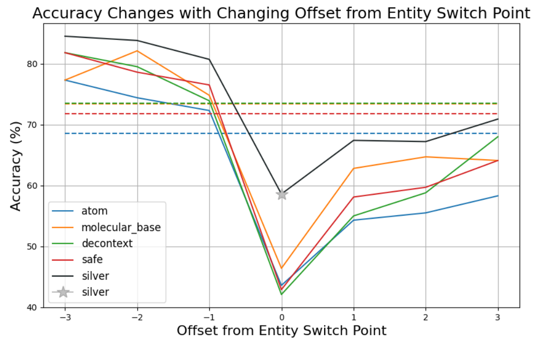

Figure 4: Variation in accuracy for different fact-checking methods as the offset from the entity switch point changes. Each line represents a method, with the solid lines indicating the method's accuracy at different offsets, and the dashed lines representing the overall accuracy of the method. The **silver** star represents the performance of human-in-the-loop molecular claim generation.

Figure 5: Changing preferences of selection of disambiguating fact by molecular decontextualization for long-form generation with hallucinations.

ting and generating boilerplate code for figure generation in this paper.

We use the open-source dataset published by Wang et al. (2024) under the Apache 2.0 license. We also use the open-source code-base of FactScore (Min et al., 2023) for evaluations which is published under MIT license and AlignScore (Zha et al., 2023) published under MIT License.

F Controlled Experiment on Minimality Generation Details

Filtering Criteria applied in Step 3 Before filtering claims which are supported by more than two atomic facts, we do not consider cases where one atomic fact is a substring of another one.

Filtering Criteria applied in Step 4 We detail the filtering criteria applied in evidence generation for partial support detailed in 4.1. After we sample a set of *key facts* $C_i = \{\mathbf{c}_{i,1}, \ldots, \mathbf{c}_{i,m}\}$ such that C_i contains the all atomic facts of the response r except $\mathbf{c}_b$, we also apply a filtering criteria to remove cases where the *banned fact* and **any** of the *key facts* is similar; i.e., for $\mathbf{c}_{i,k}$ in C_i, we filter cases where $e(\mathbf{c}_{i,k}, \mathbf{c}_b) = \text{supported}$. At the end of step 4 after we prompt the LLM to generate an evidence article, we also account for generation errors and remove the cases where banned fact is supported by the generated evidence.

G Remaining Challenges

To shed light on the remaining challenges, we focus on one of the most challenging scenarios for decontextualization. In the ambiguous biography dataset from Chiang and yi Lee (2024), we often observe what we call an *entity switch point*: a claim $\mathbf{c}_i$ that draws on information about entity B, when sentences $\mathbf{c}_{<i}$ all refer to entity A. This is where decontextualization is crucial to recognize that $\mathbf{c}_i$ in context does not refer to the correct entity.

Molecular claims recover fastest at the entity-switching point We investigate the performance of baselines under the lens of ambiguity resolution. Note that these results are reported on baselines tested with `gpt3.5-turbo`. We find that the dataset of ambiguous biographies becomes the most confusing at the entity switch point. Figure 4 shows a significant performance drop at the switch across all methods. Basic decontextualization methods (DECONTEXT, SAFE-DECONTEXT) perform the worst, underperforming the ATOMIC baseline at the switch, but molecular claims, which incorporate richer disambiguation information, show relative robustness, improving by 3.5% over the most effective decontextualization approach (SAFE-DECONTEXT).

Gap from human performance To estimate the upper bound of ideal performance at the entity switch point in Figure 4, we generate molecular claims at the entity-switch point with weak supervision human-in-the-loop supervision. We use the prompt shown Figure 9 in which has access to gold disambiguations from Wikipedia about the entities in the passage. This method's performance even with weak human supervision is significantly better than automated decontextualization methods, bringing attention to this limitation of current fact-checking pipelines.

```
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:
- Identify the main SUBJECT within the claim.
- Determine if the SUBJECT is ambiguous according to the provided AMBIGUITY CRITERIA.
- Utilize your world knowledge to enumerate potential DISAMBIGUATIONS for the identified SUBJECT.
- Specify the TYPE of information employed for disambiguation based on the list of DISAMBIGUATIONS.
- If the SUBJECT does not have ambiguous interpretations, return None
- Provide an explanation of the method used to arrive at the final response.

Format your response as a combination of explanation and a dictionary with the following structure:
##EXPLANATION##:
<step-by-step-explanations>
##RESPONSE##:
{"subject": <subject>, "disambiguations":[ <instance-1>,  <instance-2>..], "disambiguation_type": <type>}

Example 1:
##CLAIM##: David Heyman, born in 1961 in England, is the founder of Heyday Films.
##EXPLANATION##:
The SUBJECT of the claim is "David Heyman". Based on my world knowledge, there are multiple individuals who share similar names,\
 such as "David Heyman - the British film producer" and "David Heyman - the Chairman of the Board of UK HPA." \
 To differentiate between them, it is crucial to consider their respective occupations. This criterion offers a \
 clearer disambiguation compared to nationality, as both individuals are British and thus nationality alone does not \
 provide sufficient distinguishing information.
##RESPONSE##:
{"subject": "David Heyman", "disambiguations": ["David Heyman - British film producer, founder of Heyday Films",\
 "David L. Heyman - Chairman of the Board of UK HPA"], "disambiguation_type": "Occupation"}

Example 2:
##CLAIM##: Ruth Bader Ginsburg served as a Supreme Court justice.
##EXPLANATION##:
The SUBJECT is "Ruth Bader Ginsburg". According to my world knowledge, this is a unique individual and I am not aware \
of any other individuals/entities with a similar name. Hence, there are no ambiguous interpretations of this SUBJECT and the claim requires no further disambiguation.
##RESPONSE##:
{"subject": "Ruth Bader Ginsburg", "disambiguations": "None"}

Example 3:
##CLAIM##: Charles Osgood, the american television commentator, is best known for hosting CBS News Sunday Morning.
##EXPLANATION##:
The SUBJECT in focus is "Charles Osgood". Based on my world knowledge, there are two notable individuals with similar names: "Charles Osgood - American radio
and television
commentator" and "Charles E. Osgood - American psychologist." Given the ambiguity surrounding the name, \
specifying the individual's profession serves as an apt disambiguation method.
##RESPONSE##:
{"subject": "Charles Osgood", "disambiguations": ["Charles Osgood - American radio and television commentator", \
"Charles E. Osgood - American psychologist"], "disambiguation_type": "Profession"}

Similarly, disambiguate the following claim by detecting the main SUBJECT and disambiguation information for the SUBJECT using your world knowledge. \
Generate an EXPLANATION followed by dictionary-formatted RESPONSE.
##CLAIM##: [claim]
##EXPLANATION##:
```

Figure 6: Ambiguity detection prompt for detection of ambiguous entities and generating disambiguation guideline for generation of molecular claims for the baselines `MOLECULAR` and `MOLECULAR-GPT4`.

```
DECONTEXTUALIZATION CRITERIA: Decontextualization adds the right type of information to a CLAIM to make it standalone \
and contain relevant disambiguating information. This process can modify the original CLAIM in the following manners:
- Substituting pronouns or incomplete names with the specific subject being referred to.
- Incorporating the most important distinguishing details such as location/profession/time-period to distinguish the subject from others who might share similar names.
- Should not omit information from the original CLAIM.

Instructions:\
- Use the "subject" and the CONTEXT to substitute any incomplete names or pronouns in the CLAIM.
- Use the "disambiguation_type" and the CONTEXT to resolve ambiguity by adding clarification phrases about the SUBJECT to the claim.
- If information from disambiguation_type is already present in the CLAIM, no decontextualization necessary, return the original claim as is.

Example 1:
##CLAIM:##: He is best known for hosting CBS News Sunday Morning.
##DISAMBIGUATION GUIDELINE##:  {"subject": "Charles Osgood", "disambiguation_type": "Occupation"}
##CONTEXT##: Charles Osgood, a renowned American radio and television commentator and writer, was born on January 8, 1933, in the Bronx, New York City. He is best known for hosting "CBS News Sunday Morning" for over 22 years and "The Osgood File" radio commentaries for over 40 years. Osgood also authored several books, including "The Osgood Files", "See You on the Radio", and "Defending Baltimore Against Enemy Attack". He was born to Charles Osgood Wood, III and his wife Jean Crafton, and grew up with five siblings. Osgood graduated from Fordham University in 1954 with a bachelor of science degree in economics.
##EXPLANATION##: The SUBJECT "He" pertains to "Charles Osgood". The DISAMBIGUATION GUIDELINE indicates that there are \
multiple individuals named "Charles Osgood", distinguishable by their occupations. The context clarifies that the \
referenced subject in this claim is Charles Osgood, who is "a renowned American radio and television commentator and \
writer". Opting for minimal disambiguating information, "a commentator" aligns well with the claim concerning hosting \
a news show.
##DECONTEXTUALIZED CLAIM##: Charles Osgood, the commentator, is best known for hosting CBS News Sunday Morning.

Example 2:
##CLAIM:##: Heyman is the founder of Heyday Films.
##DISAMBIGUATION GUIDELINE##: {"subject": "David Heyman", "disambiguation_type": "Occupation"}
##CONTEXT##: David Heyman is a renowned film producer and founder of Heyday Films, known for producing the entire \
"Harry Potter" film series and collaborating with director Alfonso CuarÃ³n on "Harry Potter and the Prisoner of Azkaban" \
and "Gravity". He was born on July 26, 1961, in London. His family has a background in the film industry, with his parents\
 being a producer and actress. Heyman studied Art History at Harvard University and began his career in the film industry\
  as a production assistant. Throughout his career, he has received numerous awards and nominations, including an Academy \
  Award nomination for Best Picture and a BAFTA Award for Best British Film.
##EXPLANATION##: The SUBJECT "Heyman" refers to "David Heyman". The DISAMBIGUATION GUIDELINE indicates that there are \
multiple individuals named "David Heyman", distinguishable by their occupations. The CONTEXT clarifies that the \
referenced SUBJECT in this claim is Heyman which refers to David Heyman and the subject's occupation is film producer. \
We opt for minimal disambiguating information by adding "a film producer" as a disambiguation for the SUBJECT.
##DECONTEXTUALIZED CLAIM##: David Heyman, the film producer, is the founder of Heyday Films.

Now generate an EXPLANATION and DECONTEXTUALIZED CLAIM for the following. Ensure that \
only minimal information is added to eliminate ambiguity, such as adjusting pronouns or including \
clarifying details. When faced with multiple options for disambiguation under "disambiguation_type," \
prioritize information consistent with the CONTEXT. Avoid repeating information if the claim already \
includes information suggested by the "disambiguation_type."

##CLAIM:##: [claim]
##DISAMBIGUATION GUIDELINE##:[disambiguation]
##CONTEXT##: [context]
##EXPLANATION##:
```

Figure 7: Molecular decontextualization prompt for the baselines MOLECULAR and MOLECULAR-GPT4.

DECONTEXTUALIZATION CRITERIA: Decontextualization adds the right type of information to a CLAIM to make it standalone. This process can modify the original CLAIM in the following manners:
- Substituting pronouns or incomplete names with the specific subject being referred to.
- Including contextual information to provide more context about the subject.

Instructions:- Identify the "subject" of the claim and locate the claim within the context.
- Use the CONTEXT to substitute any incomplete names or pronouns in the CLAIM.
- If there is no decontextualization necessary, return the original claim as is.
- The decontextualization should minimally modify the claim by only adding necessary contextual information.
- Refer to the following examples to understand the task and output formats.

Example 1:
CONTEXT: Almondbury Community School bullying incident: The clip shows the victim, with his arm in a cast, being dragged to the floor by his neck as his attacker says "I'll drown you" on a school playing field, while forcing water from a bottle into the victim's mouth, simulating waterboarding. The video was filmed in a lunch break. The clip shows the victim walking away, without reacting, as the attacker and others can be heard continuing to verbally abuse him. The victim, a Syrian refugee, had previously suffered a broken wrist; this had also been investigated by the police, who had interviewed three youths but took no further action.
CLAIM: The victim had previously suffered a broken wrist.
DECONTEXTUALIZED CLAIM: The Syrian refugee victim in the Almondbury Community School bullying incident had previously suffered a broken wrist.

Example 2:
CONTEXT: Isaiah Stewart: Stewart was born in Rochester, New York. He grew up playing soccer and boxing.
CLAIM: He grew up playing boxing.
DECONTEXTUALIZED CLAIM: Isaiah Stewart grew up playing boxing.

Example 3:
CONTEXT: Arab Serai: According to S.A.A. Naqvi, Mughal emperor Humayun's widow Haji Begum built this "serai" in c. 1560/61 to shelter three hundred Arab mullahs whom she was taking with her during her "hajj" to Mecca; however, Y.D. Sharma opines that the word Arab in the title is a misnomer as this building was built for the Persian craftsmen and workers who built the Humayun's Tomb. In January 2017, the Aga Khan Trust for Culture started a project to conserve the "serai". The restoration was completed in November 2018. In March 2019, the trust announced a planned project to conserve the "baoli" (stepwell) of the serai with the help of funds from the embassy of Germany.
CLAIM: The planned project is to conserve the "baoli" (stepwell) of the serai.
DECONTEXTUALIZED CLAIM: The Aga Khan Trust for Culture's planned project in March 2019 is to conserve the "baoli" (stepwell) of the Arab Serai.

Example 4:
CONTEXT: Mason Warren: Warren was born in Doncaster, South Yorkshire and started his career with Rotherham United, where he progressed from the youth team to sign a professional contract in May 2015. He was taken with the first team on the pre-season tour of Scotland and became a regular with the development squad before he was sent to NPL Division One South side Sheffield on a two-month youth loan deal. He was a prominent figure in the side making six appearances during his loan spell before he was recalled in early January 2016. In February 2016, he was loaned out again joining National League North side Harrogate Town on a one-month loan deal. After picking up the Player of the Month award for Harrogate during February, his loan was extended until April. He went on to make a total of eleven appearances for Town. Upon his return to Rotherham in April, he signed a new two-year contract extension until 2018.
CLAIM: He signed a new two-year contract extension until 2018.
DECONTEXTUALIZED CLAIM: Mason Warren Warren signed a new two-year contract extension until 2018 with Rotherham United.

Example 5:
CONTEXT: Lost Girls (band): Lost Girls is a band that primarily consists of Patrick Fitzgerald and Heidi Berry. They formed in 1998 after Fitzgerald left Kitchens of Distinction and Berry left 4AD, which had released three of her albums after her appearance on This Mortal Coil's 1991 album "Blood".
CLAIM: 4AD had released three of her albums.
DECONTEXTUALIZED CLAIM: 4AD had released three of Heidi Berry's albums before she left to form Lost Girls.

Example 6:
CONTEXT: Bernard Joseph (politician): He was a member of the Congress of the People before he joined the Economic Freedom Fighters. Joseph said that he left the party because he felt that the party lacked leadership and movement. He joined the Economic Freedom Fighters to implement the party's policies.
CLAIM: He joined the party to implement the party's policies.
DECONTEXTUALIZED CLAIM: Bernard Joseph joined the Economic Freedom Fighters to implement the party's policies.

Example 7:
CONTEXT: Ham Sandwich (song): On February 20, 2019, the song was self-released as a digital download on international digital stores, as well as being released through various music streaming services. The song was released partially as a response to fans who were displeased with Getter's album "Visceral", released in late 2018. It was also released shortly before the launch of his "Visceral Tour", based off of his album of the same name.
CLAIM: The album and tour are both named "Visceral".
DECONTEXTUALIZED CLAIM: Getter's album and tour are both named "Visceral".

Similarly, generate a decontextualized claim for the following pair of CLAIM and CONTEXT making minimal alterations to the original structure of the CLAIM while ensuring clarity and coherence.

CONTEXT: <context>
CLAIM: <claim>
DECONTEXTUALIZED CLAIM:

Figure 8: Decontextualization prompt for the baseline `SIMPLE-DECONTEXT`.

Instructions:

You are provided with information about different individuals who share similar names under "WIKI DISAMBIGUATIONS." Following this, a PASSAGE about one or more of these subjects is presented. CLAIMS extracted from this passage are then listed.

Task:

1. Identify the number of main entities introduced in the passage provided without referring to the WIKI DISAMBIGUATIONS. Determine if the passage describes a single individual or multiple individuals based on the information given.

2. For a passage that initially seems to describe one person:
 - Identify a core fact from the initial description of the entity in the passage that disambiguates this individual from all the entities in the WIKI DISAMBIGUATIONS.
 - Use this core fact (such as date of birth, specific educational background, or other unique identifiers) consistently to revise all claims related to that individual until the passage explicitly introduces a new person with the same name.

3. If the passage explicitly introduces a new person (e.g., stating "There is another person called XYZ"), identify a new core fact about this individual directly from the passage. Use this core fact to revise all subsequent claims related to this new individual.

4. Consistently apply the selected type of disambiguation (such as profession, birth date, or specific identifier) for each claim about an entity, ensuring all claims about the same entity are revised uniformly.

5. Make a revision for each claim with consistent disambiguation information added.

Output Format:

Think step by step, and finally provide all claim revisions in a structured list format, where each revised claim is clearly matched with its original claim. Each revised claim should be nested within <revision> tags, ensuring that the chosen method of disambiguation is applied uniformly to all claims about the same entity and that the revision remains succinct. E.g.
- Claim: <revision> revised-claim </revision>

WIKI DISAMBIGUATIONS: <disambigs>

PASSAGE: <passage>

CLAIMS: <claims>

Figure 9: Silver labels ambiguity detection prompt for detection of ambiguous entities and generating disambiguation guideline for generation of molecular claims for the baselines `MOLECULAR` and `MOLECULAR-GPT4`.

Generate a news article that contains the following **key facts**:
1. Key Fact 1
2. Key Fact 2
3. […]

The document should NOT have the following **banned facts**:
1. Banned Fact

Remember that your document must avoid mentioning any of the banned facts, ensuring that they are not hinted at or implied throughout the content. You can use general knowledge and common sense knowledge to create realistic articles that contain the the key facts.

Figure 10: Prompt for controlled evidence generation to generate articles that incorporate key facts and avoid banned facts.