Dissecting Atomic Facts: Visual Analytics for Improving Fact Annotations in Language Model Evaluation

Manuel Schmidt *
University of Konstanz

Daniel A. Keim †
University of Konstanz

Frederik L. Dennig ‡
University of Konstanz

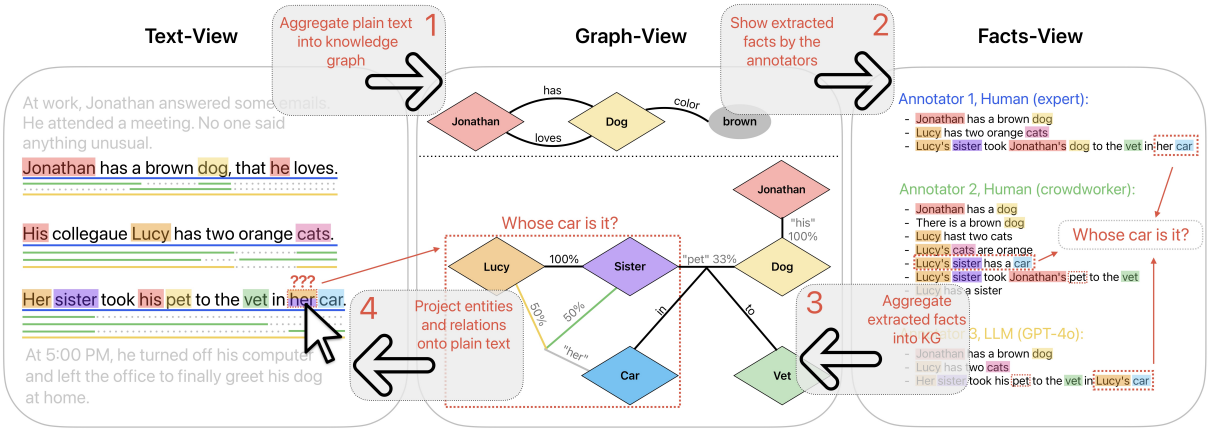


Figure 1: Whose car is it? Our conceptual design illustrates a visual interface revealing inconsistencies in factual annotations across human annotators and language models. Entity highlighting, atomic facts, and corresponding knowledge graphs are shown side by side. Ambiguous references, e.g., the unresolved ownership of "her car", are selected for inspection. Small-multiple graphs enable interactive exploration of the local context, providing insight into alternative interpretations and annotation uncertainty.

Abstract

Factuality evaluation of large language model (LLM) outputs requires decomposing text into discrete "atomic" facts. However, existing definitions of atomicity are underspecified, with empirical results showing high disagreement among annotators, both human and model-based, due to unresolved ambiguity in fact decomposition. We present a visual analytics concept to expose and analyze annotation inconsistencies in fact extraction. By visualizing semantic alignment, granularity and referential dependencies, our approach aims to enable systematic inspection of extracted facts and facilitate convergence through guided revision loops, establishing a more stable foundation for factuality evaluation benchmarks and improving LLM evaluation.

Index Terms: Visual analytics, atomic facts, text annotation.

1 Introduction

Recent advances in large language models (LLMs) led to increased reliance on evaluating their *factuality*, which describes the LLMs' ability to generate outputs that are accurate, verifiable, and coherent. Validating an LLMs factuality is crucial for critical use cases in which reliable outputs are mandatory, such as in public services, healthcare or law. However, evaluation pipelines such as *FActScore* [3], *TruthfulQA* [2], or *FEVER* [7] rely heavily on extracted atomic facts or claims: minimal, discrete units of meaning against which outputs can be evaluated. This raises a foundational issue: What constitutes an atomic fact?

*e-mail: manuel.schmidt@uni.kn
†e-mail: daniel.keim@uni.kn
‡e-mail: frederik.dennig@uni.kn

Despite being central to factuality evaluation, there is no agreed definition of an atomic fact. Atomic facts inherit some structure from *Summarization Content Units (SCUs)* of earlier summarization evaluation work (Pyramid Method) [4], yet remain underspecified and lack consistent formalization. This ambiguity has practical consequences. In our annotation experiments, we observe substantial disagreement between both human and LLM-based annotators, particularly in the amount and granularity of extracted facts. These inconsistencies suggest a deeper methodological gap: factual correctness cannot be evaluated if the annotation ground truth disagrees with itself. This lack of standardized definitions leads to substantial differences in fact decomposition methods, reducing the meaningfulness of LLM evaluation [8].

We argue that low inter-annotator agreement (IAA) is a signal of incomplete or ambiguous annotation guidelines. Definitions of atomicity trade off between *independence* (i.e., minimal dependency on other facts) and *simplicity* (i.e., short facts that do not carry long context). This creates structurally unstable representations: independent facts tend to become overly verbose, while simple facts often encode hidden dependencies. Philosophical approaches like Russel's *logical atomism* [6] or *neo-Davidsonian event semantics* offer theoretical insight, yet they remain infeasible in practice due to linguistic ambiguity and severe annotation overhead [1].

We propose to tackle this problem through visual analytics (VA): We present a conceptual framework for a visual fact annotation tool designed to expose disagreement (see Fig. 1), allowing for systematic annotation guideline refinement, until convergence. Once the guidelines reach a robust state in which both humans and LLMs reach a certain level of agreement, the tool can be used as a human-in-the-loop LLM-guided annotation tool, increasing both annotation accuracy and efficiency and ultimately improving LLM evaluation.

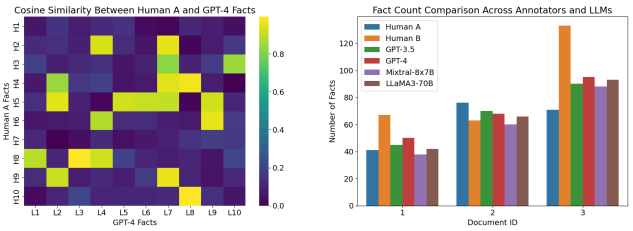


Figure 2: Initial analysis shows high disagreement between human annotators and LLMs both in semantics (left) and granularity (right).

2 Method

To study disagreements in human and LLM annotations, we manually annotated 13 German administrative documents from the platform *service-bw*. Each document was processed into a list of natural language atomic facts based on our annotation guidelines. Three of these documents were independently annotated by two human annotators to measure inter-annotator agreement (IAA). Our annotation guidelines contain instructions on how to deal with anaphora resolution, conditionals and conjunctions. The instruction fine-tuned LLMs (GPT-3.5, GPT-4, LlaMA3-8B/70B, Mistral-7B, Mixtral-8x7B/22B) were supplied with a plaintext version of our annotation guidelines and annotated 800 documents of *service-bw*.

To compare annotations, we developed an embedding-based fact matching pipeline. First, facts are embedded using SBERT [5]. Pairwise cosine similarities are computed between two fact lists. Then, the *Hungarian algorithm* is applied to find an optimal assignment of facts. We use a similarity threshold to discard low quality matches. The resulting IAA is the *Jaccard similarity* applied to the two binary matching vectors. The threshold was empirically optimized to minimize deviation from manual human-based matching. We also evaluated alternative embedding methods (TF-IDF, CLIP, FlagEmbedding), but SBERT consistently aligned closest with human annotations.

Despite multiple iterations of producing detailed annotation guidelines, IAA for all human-human, human-LLM and LLM-LLM pairs remained highly variable (see Fig. 2, left). Disagreement was especially pronounced in granularity (see Fig. 2, right), i.e., annotators differed in how finely they decomposed conjunctive and conditional structures, and referential dependency, i.e., whether contextual elements like conditions and entities should be replicated for completeness or omitted. Language models also diverged, often merging conceptually distinct facts, omitting conditions or overgeneralizing. These results demonstrate the fragility of atomic fact extraction when definitions are underconstrained, indicating the need for human intervention. To address these issues, we propose a VA system to support annotation convergence and guideline refinement. In the following, we describe the components of our approach.

Text-anchored fact highlighting: Highlights in text associated with extracted facts from multiple annotators, enabling direct comparison of extractions (see Fig. 1, *Text-View* and *Facts-View*).

Semantic similarity heatmap: Alignment matrix of annotator/model facts with color-coded similarities for comparative overview and detection of disagreements (see Fig. 2, left).

Fact count histogram: Comparison of fact granularity across multiple models and annotators (see Fig. 2, right).

Knowledge graphs: Comparison of facts and source text as small multiple entity-relation graphs to identify mismatches, uncertainties and omissions through visual aggregation (see Fig. 1, *Graph-View*).

Branching logic visualization: Visualization of parsed conditionals and conjunctions into tree-structured representations to explore semantic decompositions variants.

These views help analysts identify divergent guideline interpretations, missing or ambiguous dependencies and over- or under-specified facts. This workflow to identify disagreement is itself a four-stage loop (Fig. 1): (1) aggregate the plain text into a knowledge graph (KG) through entity and relation extraction to create an overview, (2) highlight these entities in the extracted facts, (3) aggregate extracted facts into small-multiple KGs and (4) project the semantics of the extracted facts back onto the original text to reveal gaps or overspecification. Ultimately, our approach aims to support a revision loop of nested disagreement exposure, guideline refinement, reannotation, and convergence measurement.

3 Outlook

We are developing a prototype of the proposed VA system, implementing our conceptual design. As the main goal is to facilitate annotation convergence, we plan to add features, such as semantic clustering of facts, and contradiction and redundancy detection. Once annotation convergence (high IAA) is achieved, the tools will support human-in-the-loop LLM-guided annotation for greater efficiency. LLM-based guidance will use either majority-vote based resolution of ambiguities or interactively highlight possible resolution pathways, to make the annotation process more robust and efficient. Finally, consistent annotations will serve as a foundation for more meaningful factuality evaluation by supplying facts to methods such as FActScore.

4 Conclusion

Even with detailed guidelines to extract atomic facts, our empirical results show substantial disagreement among human annotators, driven by ambiguity in decomposition, i.e. dependency resolution and granularity. We argue that such disagreement is not noise, but a signal of underspecified annotation practices. To this end, we propose a VA-based workflow that helps to surface inconsistencies and supports guideline revision and convergence. Our goal is enable stable and efficient factual annotation, which forms the necessary foundation for rigorous and interpretable evaluation of LLM outputs.

Acknowledgments

We acknowledge financial support by the Federal Ministry for Economic Affairs and Climate Action (BMWK, grant 03EI1048D).

References

[1] A. Gunjal and G. Durrett. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Find. Assoc. Comput. Linguist.: EMNLP 2024*, pp. 3751–3768. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.215 1

[2] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. 60th Annu. Meet. Assoc. Comput. Linguist.*, pp. 3214–3252, 2022. doi: 10.18653/V1/2022.ACL-LONG.229 1

[3] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. 2023 Conf. Empir. Methods Nat. Lang. Process.*, pp. 12076–12100, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.741 1

[4] A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Hum. Lang. Technol. Conf. North Am. Chapter Assoc. Comput. Linguist.*, pp. 145–152, 2004. 1

[5] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process.*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410 2

[6] B. Russell. *The Philosophy of Logical Atomism*. Routledge, 2010. 1

[7] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proc. 2018 Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, pp. 809–819, 2018. doi: 10.18653/V1/N18-1074 1

[8] M. Wanner, S. Ebner, Z. Jiang, M. Dredze, and B. V. Durme. A closer look at claim decomposition. In *Proc. 13th Jt. Conf. Lexical Comput. Semant.*, pp. 153–175, 2024. doi: 10.18653/V1/2024.STARSEM-1.13 1

Dissecting Atomic Facts

Visual Analytics for Improving Fact Annotations in Language Model Evaluation


Manuel Schmidt
University of Konstanz
manuel.schmidt@uni.kn


Daniel A. Keim
University of Konstanz
keim@uni.kn

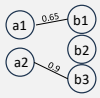

Frederik L. Dennig
University of Konstanz
frederik.dennig@uni.kn

Motivation

- LLMs are prone to making up false statements
  → Hallucinations
- Tools are needed to evaluate LLM outputs for factual correctness
- Meaningful LLM Evaluation paves the way for reliable AI in sensitive domains

LLM Evaluation using Atomic Facts

- Methods like FActScore [2] evaluate the factuality of LLMs
- Inspired by Summary Content Units from Summarization Evaluation [1]
- Utilize atomic facts
- How to match facts? Similarity based matching
- Precision $f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} \mathbf{1}[a \text{ is supported by } \mathcal{C}]$

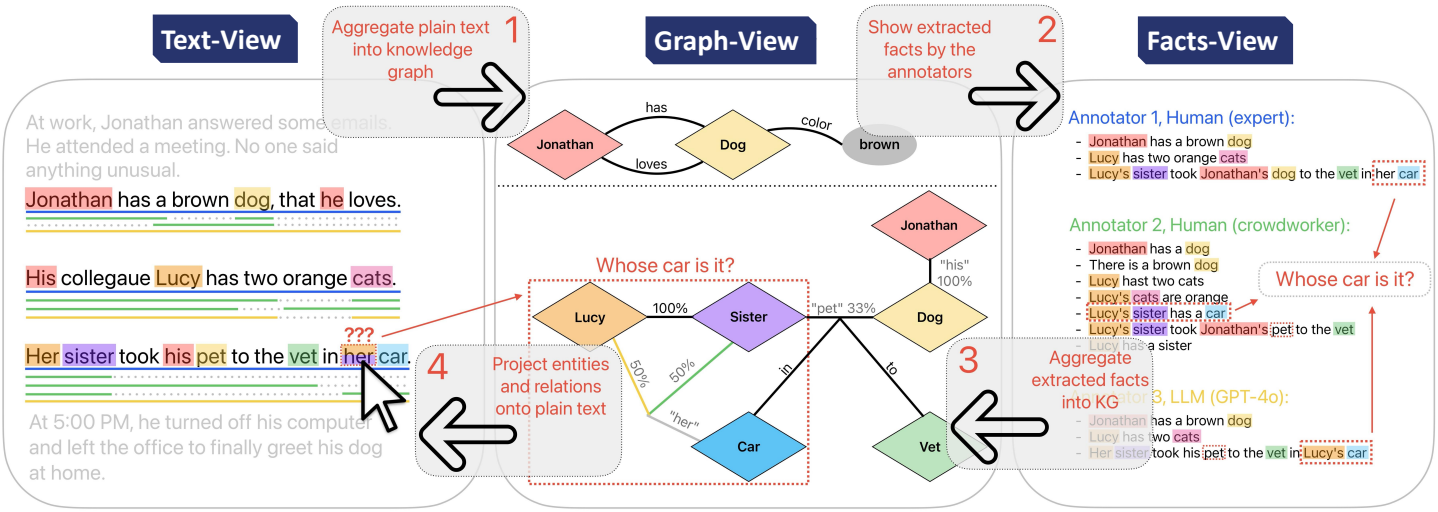


Atomic Facts

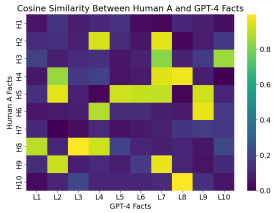
- Smallest indivisible factual statement
  → Minimal
- Self-contained statement expressing one relation or attribute
  → Independent
- What constitues an atomic fact?

Finding

- Difference in granularity of extracted atomic facts
  → Low inter-annotator agreement
- Annotators treat Dependency resolution, Conditionals and Conjunctions differently
- Ex.: The dog was brown **and** very tall – How many facts?

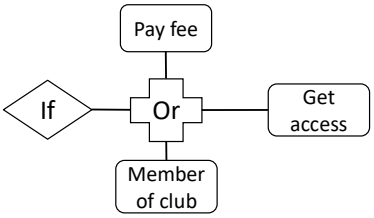


- Semantic highlighting of extracted facts
- Interactive selection highlights parts in the Graph-View



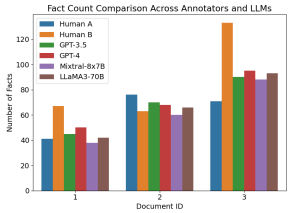
Similarity between Atomic Facts
→ Fact Matching

- Automatic knowledge graph (KG) generation
- Create KGs from original source text and extracted facts
- Comparison through small multiples



Visualize Conditionals with Tree-Viewer

- Shows facts extracted by human and LLM annotators
- Entity highlighting
- Ambiguity highlighting



Fact Count Histogram
→ Annotator Agreement

Outlook

- Visual Analytics Application under development
- Features such as semantic clustering help in finding ambiguities
- Contradiction and redundancy detection help to spot disagreements
- Once agreement converges: **Human-in-the-loop LLM-guided annotation**

Conclusion

- A Visual Analytics based workflow is necessary to address the problems
- Analytic features like **detection of annotation inconsistencies** are required
- Iteratively **revise annotation guidelines**
- Goal: **Converge inter-annotator-agreement**

[1] Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In Hum. Lang. Technol. Conf. North Am. Chapter Assoc. Comput. Linguist., pp. 145–152, 2004
[2] Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proc. 2023 Conf. Empir. Methods Nat. Lang. Process., pp. 12076–12100, 2023. doi: 10.18653/V1/2023.EMNLP-MAI

Universität Konstanz
Manuel Schmidt, Daniel A. Keim, Frederik L. Dennig
Data Analysis and Visualization Group, University of Konstanz, Germany, www.vis.uni-konstanz.de

Funded by: Federal Ministry for Economic Affairs and Climate Action (BMWK, grant 03EI1048D)

Data Analysis and Visualization

November 02-07, 2025, Vienna, Austria