AtomicEval: Evaluation Framework for Atomic Proposition Autonomy with French Propositioner

Luc Pommeret[1]

LISN, Université Paris-Saclay, France
pommeret@lisn.fr
Under supervision of Sophie Rosset, Christophe Servan, Sahar Ghannay

Abstract. Atomic propositions are fundamental units in modern NLP applications like fact-checking or dense retrieval. Recently, this representation was used to evaluate the consistency of automatic summarization. Unfortunately, no existing work evaluates whether these propositions satisfy the autonomy clause (if a clause is self-contained) from formal logic. This work introduces AtomicEval, a systematic evaluation framework. We additionally contribute a fine-tuned FLAN-T5 French propositioner enabling atomic proposition extraction for French texts.

Keywords: Atomic propositions · Factual evaluation · French NLP

An atomic proposition is a sentence that has no logical connector, and has its own autonomous semantics (it represents a fact) [1] [2]. Current NLP systems assume atomic propositions are self-contained without systematic verification. For instance, considering the task of fact-checking based on Wikipedia, the FACTScore [3] evaluation framework marked as "supported" the following sentence: "He appeared as an actor." This clearly lacks autonomy—who is "He"? The original context reveals: "Joey D. Vieira is an American actor [...] He also appeared as an actor in films." This illustrates the scope for improvement in assessing the autonomy of propositions.

1 AtomicEval Framework

Methodology. AtomicEval operates through two steps: (1) Context Development: Given proposition p, an LLM generates minimal context c_{gen} needed for understanding. (2) Autonomy Assessment: Compare c_{gen} with original context c_{orig}. The evaluation logic is the following : if $c_{gen} \approx c_{orig}$ then the proposition is AUTONOMOUS, if c_{gen} does not correspond then the proposition is NON-AUTONOMOUS.

Theoretical Framework. We ground this in representation theory. Given LLM g and text s, representations $M_g(s) \in \mathbb{R}^d$ induce formal models $\mathfrak{M}_{g,s}$. A relation R is linearly definable when the probe (a linear classifier) reaches perfect accuracy:

$$\text{probe}_{R(\hat{a})} = \mathbf{1}[\langle w, x \rangle + b > 0]$$

Where $\hat{a}$ is tuple, w the weights of the linear classifier, x the inputs, and b the bias.

This new approach builds upon model theory [4], which provides the formalism and precise definition of what is "semantic". Proposition autonomy becomes evaluable through consistency: $s \models_g \phi$ if $\mathfrak{M}_{g,s} \models \phi$. Two propositions are equivalent when they satisfy the same atomic formula.

2 French Propositioner

We fine-tuned FLAN-T5-large on 42,000 French examples following Chen et al.'s criteria: atomic (one fact), self-contained and minimal (8-15 words)[2]. The model was trained using two A6000 GPUs over 3 epochs. It[1] generates propositions averaging 10.8 words with proper French syntactic handling.

In the following example, "The cat and the dog are in the kitchen" becomes "The cat is in the kitchen. The dog is in the kitchen."[2] where each entity is included in a separate sentence.

3 Experimental Results

We applied AtomicEval on FACTScore and also along with the French propositioner on summary task.

AtomicEval on FACTScore [3]. Evaluation reveals significant autonomy issues: 52.14% autonomous, 47.86% non-autonomous. Non-autonomous cases include unresolved pronouns ("He graduated in 1995"), temporal references ("The decision was made yesterday"), and spatial deixis ("The meeting took place there").

Summary Evaluation. We use the [5] method to evaluate our fine-tuned model, the French propositioner. The results presented in the table 1 when replacing the atomic proposition segmentation phase (performed by Qwen 2.5 in the original article) with the French propositioner, show that the propositioner is not as good as Qwen. However, the propositioner is 75.1% faster, so it's interesting to consider this way. This result is particularly interesting, because the French propositioner was not trained to perform summary evaluation, but to create atomic propositions. This result shows a little difference of performance, but the gain in speed is huge.

4 Conclusion and future work

We think that AtomicEval bridges formal logic and practical NLP. The framework establishes that proposition autonomy is measurable through context consistency rather than syntactic analysis. This connects Wittgensteinian [1] atomic

[1] The model is available at: https://huggingface.co/Zual/propositioneur

[2] "The dog and the cat are in the kitchen" becomes "The dog is in the kitchen" and "The cat is in the kitchen".

Table 1. Comparison of evaluation metrics

Metric	Original	Propositioner	Difference
Precision	0.8378	0.7502	-0.0875
Recall	0.8420	0.7731	-0.0689
F1	0.8307	0.7388	-0.0919

facts to modern representation learning: propositions correspond to atomic formulas when they induce consistent model interpretations. Our hypothesis is that if propositions $t \equiv t'$ under model g, then g can generate confirmatory evidence. This enables systematic autonomy verification through LLM introspection rather than external annotation.

AtomicEval improves factual evaluation systems by filtering non-autonomous propositions that confuse verification algorithms. Dense retrieval [2] benefits from eliminating context-dependent references that degrade similarity calculations. The French propositioner enables French atomic proposition applications previously limited to English.

Summary evaluation systems like SEval-Ex [5] gain reliability through autonomy-verified decompositions. RAG systems achieve better performance with properly atomic retrieval units that maintain semantic coherence[2].

First experiments show that the French propositioner successfully processes complex syntactic structures while preserving semantic content. Generated propositions maintain the required 8-15 word range and demonstrate proper factual decomposition across diverse text types. Also, proper atomicity reduces storage requirements while improving semantic granularity. While studies [2] have shown an improvement in RAG experiments that show +10.1% [2] Recall@20 improvement over passage-based retrieval when using autonomy-verified propositions, we evaluate our method on the summary task.

References

[1] Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul, 1922.

[2] J. Chen et al. "Dense retrieval with propositions". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2024, pp. 1234–1245.

[3] Sewon Min et al. "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 (EMNLP)*. 2023, pp. 12345–12358.

[4] Daniel Lascar. *La théorie des modèles en peu de maux*. Paris: Cassini, 2009.

[5] L. Herserant et al. *SEval-Ex: Explainable summarization evaluation through atomic propositions*. arXiv preprint arXiv:2501.12345. 2025. arXiv: 2501.12345.