

Verum Sententia: An Explainable Factual Verification Tool based on Natural Language Inference

Jairo Madrigal-Cutillas^{1,*†}, María Miró Maestre^{1†} and Elena Lloret^{1†}

¹Department of Software and Computing Systems, University of Alicante, Apdo. de correos, 99, Alicante, E-03008, Alicante, Spain.

Abstract

Factual verification has become increasingly important in document-grounded Natural Language Processing (NLP), since generated outputs may appear fluent while still introducing unsupported or incorrect information. To address this challenge we present Verum Sententia, an interactive tool for factual verification over documents. The system combines evidence retrieval, statement normalization, Natural Language Inference, and entity-aware validation to determine whether a candidate statement is supported by document content. Its architecture is motivated by the idea that semantic similarity alone is not sufficient for reliable verification, especially when factually incorrect passages remain topically related to the claim. Rather than treating factuality only as an offline evaluation task, the tool is designed as a user-facing application that makes the verification process inspectable through its interface. In the current evaluation setting, the underlying verification pipeline achieves around 80% accuracy, and human assessment indicates that only about 7% of the reviewed cases are misclassified. These results suggest that Verum Sententia is a practical tool for transparent and practical factual verification in real document-based NLP workflows.

Keywords

Factual Verification, Natural Language Generation, Evidence Retrieval, Natural Language Inference

1. Introduction

The versatility of Large Language Models (LLMs) has made them increasingly useful for a wide variety of tasks, within which generating information is each time receiving more attention in the field of Natural Language Generation (NLG) considering the challenges these models face when ensuring that the generated output remains faithful to the source material rather than merely plausible at the surface level.

This issue becomes especially important when the input consists of PDF documents, since these sources often contain dense descriptive content, named entities, dates, quantities, and layout or OCR artifacts such as broken ligatures, misinterpreted column flows, or hyphenation noise that can complicate downstream verification. In practice, a generated statement may be topically related to the source while still introducing unsupported details, omitting relevant constraints, or conflating facts from different passages [1].

For this reason, factual verification has become an important component of trustworthy document-based Natural Language Processing (NLP) systems [2]. A reliable verification workflow should not only determine whether a claim appears semantically compatible with a document, but also check whether the supporting evidence actually preserves the factual anchors required to justify that decision.

This paper presents Verum Sententia, a system demonstration focused on factual verification over PDF documents through an interactive and modular architecture. Rather than framing factuality only as an offline evaluation problem, the system is conceived as a user-facing tool that makes the verification process observable, inspectable, and reusable in practical document-grounded scenarios.

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

†These authors contributed equally.

✉ jmc39@gcloud.ua.es (J. Madrigal-Cutillas); maria.miro@ua.es (M. M. Maestre); elloret@dlsi.ua.es (E. Lloret)

🌐 url (M. M. Maestre); url (E. Lloret)

🆔 0009-0008-3026-4302 (J. Madrigal-Cutillas); 0000-0001-7996-4440 (M. M. Maestre); 0000-0002-2926-294X (E. Lloret)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Verum Sententia

Provide the processed document and the claim to verify. The demo will automatically select the most relevant sentences before evaluation.

Processed document

Paste the full processed document here...

The demo will automatically trim the most relevant fragment before evaluation.

Claim to verify

Paste the claim to verify here

It can be in Spanish, Valencian, or English.

EVALUATE

This view reproduces the evidence selection stage of the locations pipeline before verification.

Figure 1: Verum Sententia user interface showing the input orchestrator for document-grounded factual verification.

Verum Sententia is an interactive tool that analyzes PDF-based evidence through a pipeline combining evidence retrieval, Natural Language Inference (NLI), and entity-aware validation. The system ingests pre-extracted text from PDF sources to identify relevant evidence fragments. To ensure high-quality inference, the pipeline reformulates candidate statements when they lack necessary context, for instance, transforming a brief, fragmented response such as “*In 1928*” into a full declarative statement like “*Alguenya began its segregation from Pinoso in 1928*”, before applying the NLI-based verification stage and constraining the final decision through checks over named entities and numeric information.

This design allows users to inspect not only the final factuality label, but also the intermediate stages that lead to it, including the retrieved passage, the normalized statement, and the validation signals used to support the prediction.

The remainder of this paper is organized as follows. Section 2 summarises the salient works related to the task at hand. We present the architecture of Verum Sententia in Section 3. Section 4 shows the evaluation methodology conducted and, finally, Section 5 concludes the paper with final highlights and future work that could be pursued basing on our present study.

2. Related work

The literature has addressed this problem through several lines of work, including benchmark construction, hallucination detection, automatic factuality metrics, and evidence-based verification strategies [2, 3, 4, 5].

Among these approaches, NLI has become one of the most useful paradigms for factual verification, since it frames the problem as deciding whether a hypothesis is entailed, contradicted, or not fully supported by a premise [6]. This perspective is particularly suitable for document-grounded verification because it enables factuality decisions to be tied directly to retrieved evidence passages rather than to broad semantic similarity alone [5].

Related research has also shown that semantic inference often benefits from complementary validation signals. In particular, Named Entity Recognition (NER), keyword-based evidence selection, and semantic similarity cues have been used to improve the precision of factuality judgments and reduce errors caused by topically related but factually incorrect passages [7, 8]. In document collections containing locations, historical references, organizations, dates, or quantities, these additional constraints are

especially valuable because factual errors often arise from entity drift rather than from complete semantic mismatch.

Considering the focus on previous literature, comparatively less attention has been paid to interactive tools that expose the verification path to the user and make the internal stages of the process transparent in a demonstrable interface. Furthermore, most existing factuality frameworks and benchmarks are developed exclusively for English, leaving a significant gap for other languages. While some recent and pioneering efforts have begun to address this by proposing hallucination detection approaches specifically for Spanish, such as the work by [9], which combines linguistic analysis with factual verification against structured knowledge sources like Wikidata, there remains a lack of end-to-end, user-facing tools that operationalize these capabilities for document-grounded workflows. Verum Sententia is positioned within this line of work as a demonstrable tool centered on document evidence and interpretable decision-making. Its contribution is to show how a verification architecture based on evidence retrieval, NLI, and entity-aware validation can be effectively operationalized as an interactive system that bridges the gap between specialized detection models and practical, multilingual document analysis.

3. System architecture

The system is designed as a modular, document-grounded verification pipeline. Its primary objective is to contrast unstructured sources into a verifiable knowledge base against synthetic claims through a hybrid linguistic engine. The architecture is divided into a high-performance Python-based backend and a reactive web frontend.

3.1. Backend

The backend, implemented in Python, operates through a multi-stage process intended to maintain consistency even when dealing with regional languages.

1. **Translation and Pivot Logic.** To leverage NLI models with high performance on English benchmarks (such as MNLI corpus [10]) the backend incorporates a translation and alignment stage. By utilizing this English-centric pivot, the system gains a significant multilingual advantage, allowing it to process evidence and claims in various languages (such as Spanish or Valencian) without being restricted by the limited availability of high-quality NLI models in those specific languages. Furthermore, structurally weak inputs, such as short answer fragments or isolated dates, are reformulated into declarative statements using predefined syntactic templates. This ensures that the inference engine always receives a well-formed hypothesis, where the fragmented response is embedded into a complete sentence structure that preserves the original semantic intent before evaluation.
2. **Document-Scoped Evidence Retrieval.** When a statement is submitted, the system ranks the document sentences based on their relevance to the claim. This retrieval strategy combines:
 - **Semantic Similarity:** Using embeddings to find topically related concepts.
 - **Lexical and Numeric Overlap:** Checking for specific terms, toponyms (e.g., *Pinoso*, *Vinalopó*), and dates.

This dual-ranking approach ensures high recall even in documents with noisy formatting, selecting the most pertinent sentences to serve as the “Premise” for the inference stage.

3. **Verification Engine (NLI+NER).** The decision logic balances semantic meaning with factual precision through two layers:
 - **NLI Inference:** A RoBERTa-based model [11] estimates the logical relationship, providing a probability distribution between *Entailment* (Compatible), *Neutral* (No concluent), and *Contradiction*.

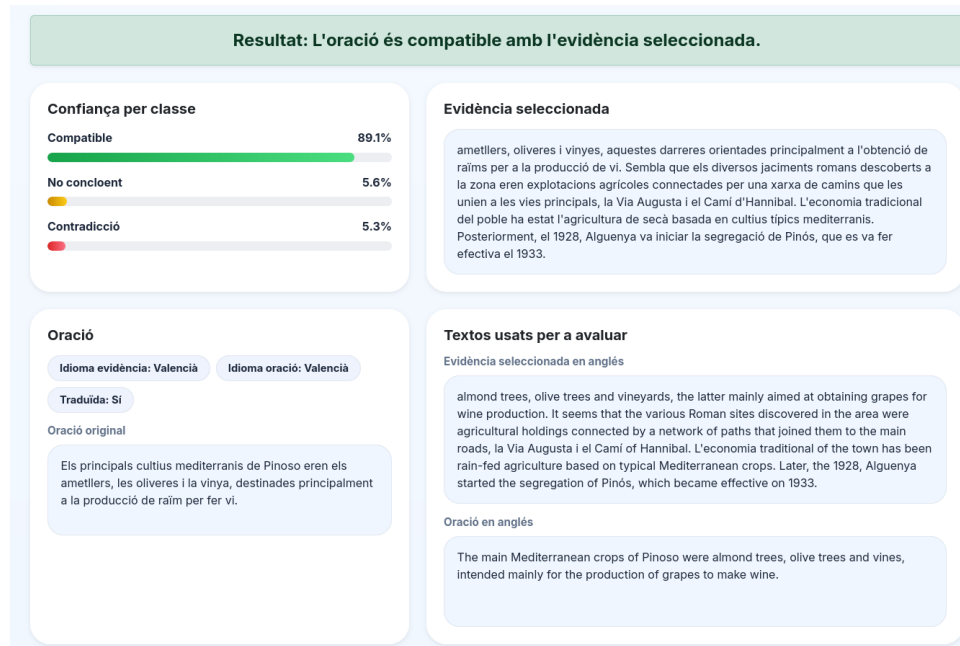


Figure 2: Verification of a factually correct claim using an example in Valencian.

- **Entity-Aware Layer:** This symbolic layer identifies named entities and numeric values using the spaCy NLP framework to extract locations, dates, and organizations. Specifically, it utilizes a pre-trained transformer-based pipeline. The system employs the `en_core_web_trf` model for English and `es_core_news_lg` for Spanish-based entity extraction.¹ If a significant entity in the claim (such as a specific date or product) lacks support in the retrieved premise—verified through a multi-stage process of exact canonical matching, numeric validation, and conservative fuzzy matching—the system adjusts the final confidence score or label to reflect this factual discrepancy.

3.2. Frontend

The frontend acts as a thin, reactive interaction layer designed to facilitate direct user engagement with the verification engine. Its primary role is to collect user input, manage the asynchronous communication with the API, and render the returned decisions along with their supporting signals. By delegating all heavy computational tasks—such as retrieval, translation, and neural inference—to the backend, the frontend remains a lightweight web application compatible with standard browser environments.

As illustrated in Figure 1, the interface is designed around a clean, dual-input layout that mirrors the logical structure of a verification task. The “Input Orchestrator” view provides two main text areas:

- **Processed Document:** A workspace where the user pastes the source material to be used as evidence. The interface explicitly informs the user that the system will automatically trim and select the most relevant fragments before evaluation.
- **Claim to Verify:** A dedicated field for the hypothesis or statement under analysis. To support regional and international use cases, the UI specifies that inputs are accepted in Spanish, Valencian, or English.

From a demonstration perspective, the frontend serves as an observability component rather than a simple output display. Once the “Evaluate” action is triggered, the interface transitions from a data-entry state to an analytical view. Instead of presenting an opaque “True/False” label, the system exposes the entire verification path. This allows users to inspect the specific evidence fragments extracted,

¹<https://spacy.io/models>

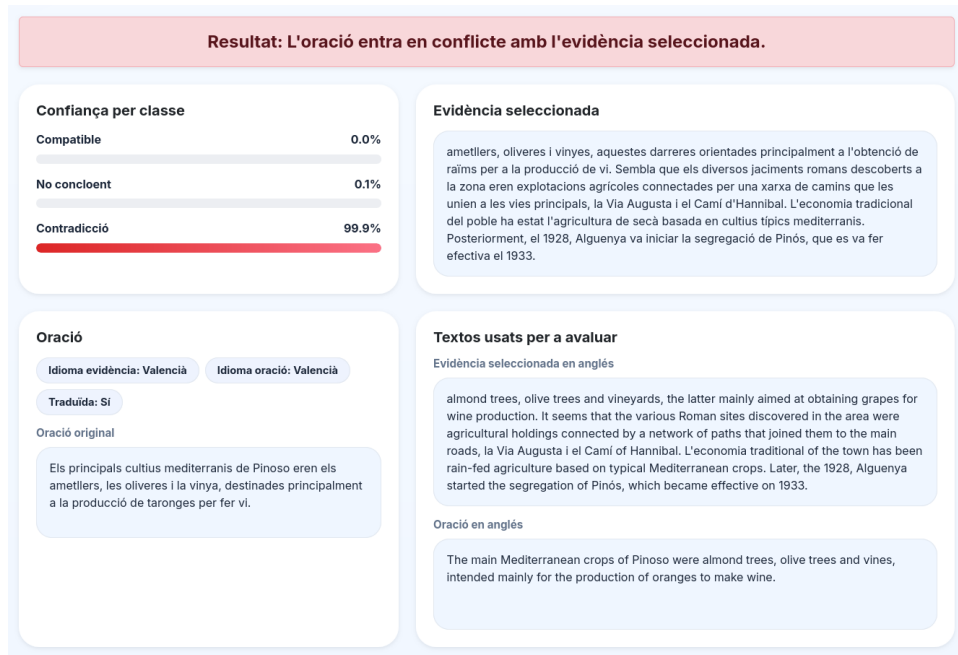


Figure 3: Detection of a factual contradiction using an example in Valencian. By substituting “grapes” with “oranges”.

the normalized versions of the text used for English-based inference, and the raw probability scores assigned to each NLI class (Entailment, Neutral, or Contradiction). This level of transparency is critical for document-grounded scenarios, as it enables the user to validate the tool’s reasoning and understand the internal linguistic stages that lead to the final factuality judgment.

3.3. Walkthrough: From Entailment to Contradiction

To demonstrate the practical utility and sensitivity of Verum Sententia, we present a comparative walkthrough of two verification scenarios based on a document describing the agricultural history of Pinoso (A town from Alicante’s province). This example illustrates how the system handles translation, evidence alignment, and the detection of factual “hallucinations” or errors.

3.3.1. Scenario A: Factual Alignment (Entailment)

As shown in Figure 2, the user provides a claim in Valencian regarding the traditional Mediterranean crops of Pinoso: *“Els principals cultius mediterranis de Pinoso eren els ametllers, les oliveres i la vinya, destinades principalment a la producció de raïm per fer vi.”* (The main Mediterranean crops of Pinoso were almond trees, olive trees and vines, intended mainly for the production of grapes to make wine).

- **Contextual Retrieval:** The system identifies the relevant narrative block in the source document which mentions “ametllers, oliveres i vinyes” and their historical role in the local economy (*almond trees, olive trees and vineyards*).
- **Cross-Lingual Mapping:** The “Textos usats per a avaluar” panel reveals the system’s internal pivot to English. The evidence is translated to describe “almond trees, olive trees and vineyards... aimed at obtaining grapes for wine production.”
- **Probabilistic Support:** The NLI engine detects a direct logical entailment between the source evidence and the claim. The interface reflects this with a 89.1% Compatibility score, confirming that the statement is factually grounded.

3.3.2. Scenario B: Sensitivity to Entity Drift (Contradiction)

The robustness of the system is best demonstrated by introducing a subtle factual error. In Figure 3, the claim is modified by replacing a single key entity: “**raim**” (grapes) is changed to “**taronges**” (oranges).

- **Semantic Overlap vs. Factual Conflict:** At a surface level, the sentence remains topically related to Pinoso and its agriculture. A simple similarity-based model might still grant this a high score. However, Verum Sententia’s NLI-based approach recognizes that the premise (vineyards for grapes) logically contradicts the hypothesis (vineyards for oranges).
- **Entity-Level Discrepancy:** The entity-aware layer flags the term “oranges” as unsupported by the retrieved evidence. This results in a dramatic shift in the confidence distribution: the Contradiction (*Contradicció*) probability surges to 99.9
- **Visual Feedback:** The interface transitions to a red alert state, explicitly stating that “The sentence conflicts with the selected evidence”. This immediate feedback loop allows researchers and users to identify exactly where a generated text has deviated from the source material.

By exposing these intermediate steps—the original text, the English pivot, and the per-class confidence—Verum Sententia moves beyond “black-box” validation, providing a transparent audit trail for the verification process.

4. System evaluation

The evaluation of Verum Sententia was designed to assess the practical behavior of the system in a document-grounded factual verification scenario. A sample dataset of 19,734 instruction-response pairs was used, and each instance was evaluated against its original source document. The source collection consisted of 184 tourism-related documents containing locality descriptions, historical references, and other factual information suitable for verification.

In the current evaluation setting, the system produced 41,565 sentence-level verification decisions and achieved around 80% accuracy. The prediction distribution was dominated by entailment (81.76%), followed by neutral (10.86%) and contradiction (7.38%), which indicates that the tool is able to identify both supported content and cases in which evidence is missing or inconsistent.

A complementary human assessment showed that only about 7% of the reviewed cases were misclassified, mainly in examples where the response was only partially supported by the original document. This metric was calculated by extracting a significant random sample from the total dataset to ensure statistical relevance. The evaluation shows that the tool provides sufficiently accurate predictions while also making the verification path inspectable through evidence retrieval, inference, and entity-aware validation.

From a demonstration perspective, these results support the usefulness of Verum Sententia as an interactive verification tool. This combination of prediction quality and transparency is especially relevant for document-based NLP workflows in which users need to inspect not only the final label, but also the evidence behind it.

5. Conclusions & Future Work

This paper presented Verum Sententia, an interactive tool for factual verification over pre-extracted text from documents that combines evidence retrieval, statement normalization, NLI, and entity-aware validation within a unified workflow. The system is designed to support document-grounded verification in an interpretable way, allowing users to inspect not only the final factuality label but also the intermediate stages involved in the decision process.

From a system-demonstration perspective, the main contribution of this work lies in showing how factual verification can be operationalized as an end-to-end application rather than only as an offline evaluation pipeline. By integrating semantic inference with additional checks over named entities and

numeric information, Verum Sententia aims to provide a more transparent and reliable verification process for PDF-based content analysis.

Future work will focus on extending the tool to additional document types, improving the robustness of document preprocessing and evidence selection, and refining the interface to better support user inspection of the verification path. Another relevant direction is the adaptation of the system to additional languages and domains, as well as the exploration of verification strategies that reduce dependence on translation-based processing.

Acknowledgments

This work has been partially funded by the Ministry of Education, Vocational Training and Sports through the collaboration grant (application no.: 25CO1/000382); by the Ministry of Science, Innovation and Universities, through the projects “SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP” (AIA2025-163322-C63), the project “Quantum mechanics for language understanding and generation” (PID2024-160791OB-I00) funded by MICIU/AEI/10.13039/501100011033 and by FEDER, and the project “Conscious Text Generation” (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF), under the slogan “A way of making Europe”, EU; by the ALIA Models Development Project within the framework of the National Language Technologies Plan -ENIA 2024 of the Ministry for Digital Transformation and Public Service and PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024; and by Fundación Ramón Areces through the project Evaluation Criteria for Quality Corpora in Artificial Intelligence (CRITERIA), developed within the framework of the 2nd National Call for Research Grants in the Humanities 2025, under the theme “Digital Humanities” (Reference: FRAHUMANIDADES25-01).

Declaration on Generative AI

During the preparation of this work, the author(s) used **Claude** in order to: **Grammar and spelling check**. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoenybi, B. Catanzaro, Factuality enhanced language models for open-ended text generation, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *Advances in Neural Information Processing Systems*, volume 35, Curran Associates, Inc., 2022, pp. 34586–34599. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/df438caa36714f69277daa92d608dd63-Paper-Conference.pdf.
- [2] P. West, C. Quirk, M. Galley, Y. Choi, Probing factually grounded content transfer with factual ablation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3732–3746. URL: <https://aclanthology.org/2022.findings-acl.294/>. doi:10.18653/v1/2022.findings-acl.294.
- [3] S. Chen, Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, J. He, FELM: Benchmarking factuality evaluation of large language models, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), *Advances in Neural Information Processing Systems*, volume 36, Curran Associates, Inc., 2023, pp. 44502–44523. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/8b8a7960d343e023a6a0afe37eee6022-Paper-Datasets_and_Benchmarks.pdf.
- [4] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, FActScore: Fine-grained atomic evaluation of factual precision in long form text generation, in: H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023,

pp. 12076–12100. URL: <https://aclanthology.org/2023.emnlp-main.741/>. doi:10.18653/v1/2023.emnlp-main.741.

- [5] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. N. Georgiev, R. J. Das, P. Nakov, Factuality of large language models: A survey, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 19519–19529. URL: <https://aclanthology.org/2024.emnlp-main.1088/>. doi:10.18653/v1/2024.emnlp-main.1088.
- [6] A. Shah, H. Shah, V. Bafna, C. Khandor, S. Nair, Validation and extraction of reliable information through automated scraping and natural language inference, *Engineering Applications of Artificial Intelligence* 147 (2025) 110284.
- [7] A. Martín, J. Huertas-Tato, Á. Huertas-García, G. Villar-Rodríguez, D. Camacho, Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference, *Knowledge-based systems* 251 (2022) 109265.
- [8] R. Panchendrarajan, A. Zubiaga, Entity-aware cross-lingual claim detection for automated fact-checking, in: V. Demberg, K. Inui, L. Marquez (Eds.), Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 17–33. URL: <https://aclanthology.org/2026.findings-eacl.2/>.
- [9] J. M.-C. y Elena Lloret, Aproximación basada en inferencia lingüística y hechos estructurados para la detección de la alucinación en textos, *Procesamiento del Lenguaje Natural* 76 (2026) 149–161. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6837>.
- [10] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, *CoRR abs/1704.05426* (2017). URL: <http://arxiv.org/abs/1704.05426>. arXiv:1704.05426.
- [11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).