Provenance Verification of Wikidata Statements at Scale

Prof. Elena Simperl
Department of Informatics
King's College London
elena.simperl@kcl.ac.uk

Dr. Odinaldo Rodrigues
Department of Informatics
King's College London
odinaldo.rodrigues@kcl.ac.uk

Dr. Albert Meroño-Peñuela
Department of Informatics
King's College London
albert.merono@kcl.ac.uk

Abstract

Wikidata is a very large Knowledge Graph containing over 1.65 billion statements, edited daily by over 24 thousand active editors. A significant portion of these statements need information that ensures their accuracy. However, external references are particularly challenging to verify. They need retrieval and parsing of the external document to select relevant passages, followed by the evaluation of the support stance of the passage for the statement, and some summarisation of the results in a form that is useful for Wikidata editors. Due to the number of statements in Wikidata and the fact that external references can change over time, manual verification is not scalable. We have conducted preliminary research, developed and deployed a MVP. In the process, we identified several important follow up questions both in terms of the technical architecture, and the user-centric design of the tool.

Introduction

Wikidata is one of the world's most important Knowledge Graphs (KGs). It is used by web search engines, virtual assistants such as Siri and Alexa, fact checkers, and in over 800 projects in the Wikimedia ecosystem. Ensuring KGs are trustworthy depends on well-documented and verifiable provenance of the information they encode [1]. This is not a trivial process. For example, we expect a statement of the form *⟨reelin, encoded by, RELN⟩* to be supported by a reference that indeed states that "the mammalian protein reelin is encoded by the *gene* RELN". This relies on assumptions about the nature of the relationship (predicate), the roles the objects play in it, and the ability to somehow verify the claim conveyed by this association.

Mechanisms that help to evaluate and ensure the quality of supporting information are thus crucial to the verifiability of KGs [1], [2], [3]. However, such processes are currently mostly performed manually [4] and do not scale with size. Yet, on vital KGs such as Wikidata and DBpedia manual verification is prohibitive due to their sheer size [2] – Wikidata has currently over 1.65 billion statements – and more support to assist with verification is needed.

In response to this need, we are building a user-centric AI assistant to verify and improve the references of Wikidata statements. We developed a MVP called ProVe (Provenance Verification), which allowed us to learn about technical and user challenges. ProVe leverages research findings designed and evaluated by the Wikidata community. It uses language models to assess if verbalisations of statements are supported by the text contained in their corresponding references – thus responding to important data assurance needs [5], [6], [7].

Despite encouraging user uptake, we identified the need to do fundamental research into 1) the design of more interactive user interfaces, able to improve on the explainability of the results of the AI model and learn from user feedback; and 2) the exploration of alternative NLP/LLM components to execute some tasks in the verification pipeline, where we identified performance could be improved. In so doing, we believe we are helping to fill a critical gap in the editing infrastructure of Wikidata. To encourage others to contribute to this line of research we will be releasing all software and data, in addition to continuing to provide access to updates of the existing prototype.

Date: From October 1, 2025 – September 30, 2026

Related work

FEVER (Fact Extraction and VERification) [8] is a large dataset containing over 185K claims, that can be used for fact verification against textual sources. Despite its general applicability, claim verification in KGs face additional challenges because statements need to be turned into sentences first and this process needs to consider important assumptions about the way information is represented. In the Reference Verification Process and Methodology section, we describe how ProVe employs FEVER in parts of some natural language tasks.

Several works have previously focussed on the verification of the quality of information in Knowledge Graphs [1], [2], [3], [4], [5], [6], [7].

However, to the best of our knowledge, ProVe is the only available tool integrated within Wikidata that automates the verification process and is based on published research. ProVe has been running since Summer 2024, even though we only started collecting detailed usage information since March 2025. Our statistics show over 10,000 requests since then (~400 requests/day) coming from 22 countries in all continents.

Reference Verification Process and Methodology

A Wikidata item can be seen as a subject "topic", consisting of one or more *statements* about it in the form <*subject, predicate, object*>. A statement is meant to convey information about the item. Although not all statements need to indicate where the information it conveys comes from, most are expected to supply a reference in support of the statement's claim.[1] Arguably, verifying the support of statements by *external* references is harder, because they are largely provided in natural language, and maintained independently from Wikidata. To verify the support of statements by references computationally we need a robust mechanism to translate the structured Wikidata statements into natural language sentences before verification, be able to retrieve the external text and evaluate the support for the statement's claim and then perform the verification every time the reference or document changes. For this to be useful to end users, we need to communicate effectively how the tool arrives at its results without hindering the users' natural workflow patterns.

ProVe interacts with users via a "gadget" – a user interface available within Wikidata during edition of items; and a Web API, which allows programmatic access to the tool's functionality (see Figure 1). It operates at the "item" level, analysing each statement about the item for which external references are provided. The main reference verification process is illustrated in Figure 2. Critical to this process are the models used in the natural language components (e.g., sentence selection and textual entailment recognition) and the communication of results to users. Improving these two aspects are the main objectives of this proposal.

The research we propose will be conducted around four key activities which will be described in more detail in the context of the reference verification process described below.

[1] https://www.wikidata.org/wiki/Help:Sources/Items_not_needing_sources

Key Activities

(A1) upgrade and publication of an up-to-date version of the datasets useful for reference quality evaluation;

(A2) benchmarking of alternative open-source LLMs and architectures to improve the performance of ProVe's main evaluation engine;

(A3) conducting essential user-centric research to co-design user interfaces for reference quality evaluation; and

(A4) consolidating findings into guidance for designing meaningful transparency interactions with end users.

In what follows, we describe how these activities fit into the reference verification process illustrated in Figure 2.

For each statement and associated external reference, ProVe first verbalises the statement (A), then retrieves the text of the reference document, segmenting it into passages (B). Passages are then ranked according to their relevance to the verbalised statement (C), and the support stance of the most relevant passages with respect to the statement analysed. The overall support stance of the referenced document for the statement is computed and provided along with the evidence found (D). Finally, the results for each statement-reference pair are aggregated to give the user an indicative score of the quality of the references in the item (E). Details of each step are given below.

Verbalisation (A)

Although the verbalisation of some statements may appear simple, in practice there are many complications. Firstly, each component of a statement <*subject, predicate, object*> is provided with a set of alternative labels describing the component, and a judicious choice for a suitable label needs to be made. Secondly, there are implicit assumptions about the roles played by the *subject* and *object* in the relationship. For example, in the triple <*william, child, george*> the intended meaning is that the subject of the statement is the parent, and its object is the child. Although just conventions, these implicit assumptions need to be taken into account to produce suitable verbalisations. To generate verbalisations that are fluent and resemble natural text, ProVe employs a T5-base model [9] fine-tuned on the WebNLG 2017 dataset [10]. As part of original research done to underpin ProVe, a dataset with verbalised Wikidata statements was generated and made publicly available [7]. Given how quickly Wikidata grows and NLP techniques evolve, this dataset and associated models would benefit from an upgrade (activity A1). In particular, a more recent version of the English WebNLG dataset became available after the first fine-tuning was performed for ProVe.

Text Retrieval (B)

Layout and other structural information embedded into referenced documents make the extraction and meaningful re-combination of text non-trivial. In addition, the text itself can contain semantical constructs spread across sentences making ad-hoc segmentation not suitable for posterior semantic evaluation of the passages. ProVe employs several custom rules to transform and remove HTML markup, after which the text is divided up into segments using spaCy's sentence segmenter using the *en_core_web_lb* model [11]. Combinations of one and two sequential text segments are produced to cater for constructs such as pronominal anaphora, before they are sent for subsequent relevance evaluation and passage selection. We have identified the need to explore alternative segmentation models to tackle some particularly challenging types of references, such as those containing supportive information in complex semi-structured form. This is one of the main objectives of activity (A2).

Passage Selection (C)

Once the claim has been verbalised and the reference text segmented into passages, we then need to rank the passages to determine those that are most relevant to the claim (independently of their support stance, which is analysed later). For the passage selection, ProVe uses a pre-trained BERT transformer [12] fine-tuned on the FEVER dataset. The passages are fed to this model to give each a relevance value in the interval [-1,1]. Since some of the passages overlap (due the combinations described in the previous part), ProVe only keeps the five highest ranked passages that do not overlap and their relevant scores. The scores are used in the claim verification step below.

Claim Verification (D)

Evaluating the support stance of the referenced document for the statement as whole is performed in two stages. First, the stance of each of the most relevant passages is evaluated by a pre-trained BERT model (also fine-tuned on FEVER) for RTE yielding a probability distribution for the classes *supportive, refuting,* and *not enough information* (i.e., inconclusive). At a second stage, the relevant scores of all passages computed in step (C) along with their support stance probability distributions are aggregated to provide an overall support stance (in one of three classes) and the support degree of the document for the statement.

ProVe Score (E)

Evaluating the support of individual statements is critical, but for editors working on items an indication of how well-supported (or refuted) an item is by its external references is also very important. Since an item i can appear as the subject of many statements, each of which can have many references, some further aggregation is needed. This is done as described below.

Let the set $\{s_1,\ldots,s_n\}$ be the set of support stances for all statement-reference pairs of item i calculated as described in step (D) above, where s_i is a number in {-1,0,1}, where -1 is used for refuting, 0 for inconclusive, and 1 for supportive references.[2] The *ProVe score* for i ($PS(i)$) is calculated as follows:

$$PS(i)=\frac{1}{n}\sum_{j=1}^{n} s_i$$

It is easy to see that $PS(i)$ is a value in $[-1,1]$ with the following intended meaning. Positive values indicate that the number of supporting references surpasses the number of refuting references and negative values indicate the opposite. In either case, the proportion of references which are inconclusive brings the score closer to 0. As a result, proximity to 1 is associated with "good" quality of the references; proximity to 0 is associated with inconclusive or missing references; and proximity to -1 indicates high levels of disparity between claims and references. Of course there is room to provide more informative indicative scores, and activity (A3) will also explore this with the caveat of keeping the evaluation results intuitive.

The ProVe score of an item along with a summary of the total of references in each stance category is shown to editors when the Wikidata page for the item is loaded (see Figure 1 with widget on the left and infobox on the right). Given the nature of the MVP and the size of Wikidata, not all scores have been computed yet. However, the user can request computation and re-computation of new scores by pressing appropriate buttons. The scores allow users to factor in reference information in the prioritisation of items to edit and to perform custom analysis with the extra information

[2] For completeness, we also consider irretrievable references as inconclusive since we cannot establish their stance with respect to the claim.

provided via the web API, e.g., the progress achieved in reference quality improvement.[3]

As implicitly hinted above, a lot of information is produced and combined during the verification process, not all of which is given to the users. For example, the degree of relevance of the passages may be potentially useful for some end-user tasks; references with contradictory information about the claim could be highlighted, and the claim verification itself may be fine-tuned from user feedback. In general, what information to exchange with users, how best to do this, and how to integrate user feedback are the main objectives of activity (A3).

We will conduct user surveys to understand user interface and feature requirements from Wikidata editors and users; and to evaluate the effectiveness of ProVe and its extensions. This will be done in two stages: focused workshops, and broader online surveys. In focused workshops, we will recruit editors with whom we have had workshops for other projects (e.g. Wikidata item recommendation [13]), via the Wikidata Telegram channel, and from the current user base of ProVe. In these focused workshops we will use "think aloud" and Wizard of Oz as techniques to understand UI and functionality requirements; we will process answers to form online survey questions. Workshop recordings will be transcribed and analysed with thematic analysis. Survey responses will be analysed with descriptive statistics and significance tests.

More generally, we then expect to consolidate the findings into general guidance for the design of AI-assisted tools for Wikidata (activity (A4)).

Workplan

	Months											
	1	2	3	4	5	6	7	8	9	10	11	12
(A1)	▨	▨	▨				▨	▨	▨			
(A2)			▨	▨	▨			▨	▨	▨		
(A3)	▨	▨				▨	▨				▨	▨
(A4)										▨	▨	▨

We start with a re-training of the existing models to produce up-to-date datasets (A1) and an initial evaluation of the current interface and API to inform any future improvements (A3). Subsequently, we explore new models for segmentation and textual entailment recognition (A2). We proceed with improvements to the user interface and API (A3); develop new models/datasets as needed in response to the findings in the first phase of (A2) and evaluate its performance and usability in a second phase of (A2). Finally, we evaluate how the tool has been received by the community (A3), consolidating lessons learned into general guidance that we hope will be useful for other developers (A4).

Expected outputs

The execution of the activities of this research proposal will result in the outputs described below.

(O1) A new version of the datasets ProVe employs will be made available, in particular an updated version of WDV, the claim verbalisation dataset built for Wikidata, following its initial release on GitHub in 2022. The intended audience of this output is any researcher, user or developer who may want to develop their own specialised verbalisation and/or verification models.

[3] This process being used by an Audiology group in Brazil (see Community impact plan).

(O2) The main ProVe server will be overhauled, resulting in improved performance. In the current iteration of the server, all requests are handled by the same process. This will be streamlined as appropriate, incorporating state-of-the-art language components and featuring a de-coupling of user and maintenance requests. The resulting service will be more stable and envisaged it will be able to process references more effectively.

(O3) The user survey conducted in (A3) will guide improvements to ProVe's interfaces. We would like to explore how best to communicate to users how the AI engine arrives at its results. This includes the role played by intermediate values obtained during the quality evaluation of the references, such as the relevance of passages within references, how these are aggregated and how an overall stance is obtained. To improve the engine, we will also collect user feedback about the quality of the evaluations performed. The intended audience of this output are mainly Wikidata editors. We note that this goes beyond direct users of ProVe as a gadget, since the API also provides programmatic access.

(O4) In general, the insights described above can also be used to provide general guidance on how to improve the design of meaningful interfaces to improve the transparency of AI-assisted tools for Wikidata. These should help in the future development of other related functionality. The audience of this output is for developers and editors alike.

(O5) We will consolidate the findings in publicly available software, documentation, and research articles as appropriate, building on the results and ongoing engagement with the community (see Community Impact Plan).

Risks

In any process that combines numerical outputs of several models in sequence to produce an indicative evaluation of quality, there is a risk that we either produce something that does not match the subjective expectation of end users, or that we may overwhelm them with too much information, detracting from their main workflow. We have specifically taken this into account and designed activity (A3) to better understand user needs and mitigate potential negative outcomes.

Community Impact Plan

We have already been actively engaging with the Wikidata community. The first version of the prototype was presented in the Wiki Workshop 2024. This led to a collaboration with an Audiology group in Sao Paulo, Brazil. Indeed, Prove is listed as a resource in the project WikiProject Hearing Health, where it is being used to help track the improvement in the quality of references the research group edits. As a result of this interaction, we received feedback which was used to improve the web API to cater for the needs of the group. Once their study is completed, we will leverage any insights into new improvements and involve the group in the execution of activity (A3).

We have a Wikidata project page with instructions on how to install the prototype and volunteer for engagement with the developers. We can reach out to some of these users to better understand their needs.

In terms of past/current events, besides the Wiki Workshop 2024 already mentioned, we were invited to present the prototype in the event "Wikimedia e Ciência: potencialidades na extensão e difusão científica" (Wikimedia Brazil), in October 2024. We are presenting at the Wikimedia Hackathon and the Wiki Workshop in May 2025; and Wikidata and Research in June 2025. We will continue to engage with these events in 2026.

In summary, we already have good community communication channels upon which we can develop closer collaboration, support the activities involving user interaction, and generate impact. We have been actively involved in the Wikidata community presenting and discussing the work with researchers. We will leverage all this in the development and sharing of research results.

Evaluation

This proposal builds on an existing prototype, implementing a process whose critical components have been based on peer-reviewed research, and whose intermediate results have been appropriately evaluated [2], [6], [7]. These works provide a blueprint for the types of research evaluation that will be needed here: ease of access, relevance, model validation, RTE metrics, etc, and contain tried and tested methodologies, including statistical analysis, user perceptions captured via crowdsourced surveys, and workshop sessions with end users. In this sense, our team has a successful track record in the research, development, and evaluation aspects of the proposal.

In terms of the software components, our development process includes mechanisms to monitor coverage, performance and usage. These can be used as indirect metrics of adoption and usefulness. We cannot measure all direct benefits to the community, but we will be able to report on the evolution of the quality scores of the items submitted to ProVe for evaluation.

Users whom we have interacted with directly have found ProVe very useful as part of their editing workflows. Ultimately, we would like to have the tool increase substantially its coverage of Wikidata items, keep the analyses up to date, make a meaningful impact in the improvement of the quality of Wikidata references, and become widely adopted.

Finally, we can envisage the indirect benefit of inspiring users and developers alike to embed AI-assisted tools in the editing workflow of Wikidata items.

Budget

The budget spreadsheet can be found here.

References

[1] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, “Quality assessment for linked data: A survey,” *Semant Web*, vol. 7, no. 1, pp. 63–93, 2016.

[2] A. Piscopo, L.-A. Kaffee, C. Phethean, and E. Simperl, “Provenance Information in a Collaborative Knowledge Graph: an Evaluation of Wikidata External References,” in *International semantic web conference*, 2017, pp. 542–558.

[3] X. Wang *et al.*, “Knowledge graph quality control: A survey,” *Fundamental Research*, vol. 1, no. 5, pp. 607–626, 2021, doi: https://doi.org/10.1016/j.fmre.2021.09.003.

[4] E. McAndrew and C. Strathmann, “Quality Assurance and reliability,” Aug. 2021, *The University of Edinburgh*. [Online]. Available: https://www.ed.ac.uk/information-services/help-consultancy/is-skills/wikimedia/wikidata/quality-assurance-and-reliability

[5] G. Amaral, A. Piscopo, L.-A. Kaffee, O. Rodrigues, and E. Simperl, “Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach,” *Journal of Data and Information Quality (JDIQ)*, vol. 13, no. 4, pp. 1–35, 2021.

[6] G. Amaral, O. Rodrigues, and E. Simperl, “ProVe: A Pipeline for Automated Provenance Verification of Knowledge Graphs against Textual Sources,” *The Semantic Web Journal*, 2023, doi: https://doi.org/10.3233/SW-233467.

[7] O. and S. E. Amaral Gabriel and Rodrigues, “WDV: A Broad Data Verbalisation Dataset Built from Wikidata,” in *The Semantic Web – ISWC 2022*, A. and K. M. and P. V. and A. J. P. A. and T. H. and M. P. and P. G. and d’Amato C. Sattler Ulrike and Hogan, Ed., Cham: Springer International Publishing, 2022, pp. 556–574.

[8] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER:

a Large-scale Dataset for Fact Extraction and VERification,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 809–819. doi: 10.18653/v1/N18-1074.

[9] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

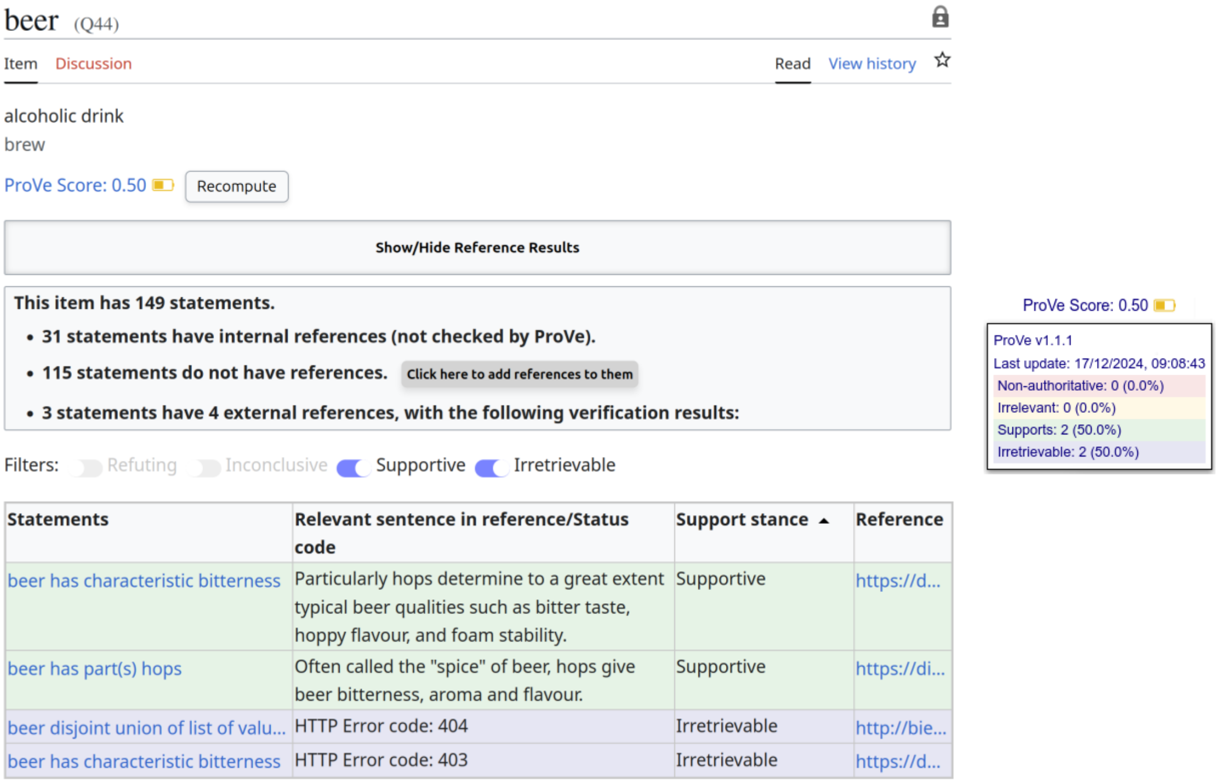
[10] C. Gardent, A. Shimorina, S. Narayan, and L. Perez-Beltrachini, “The WebNLG challenge: Generating text from RDF data,” in *Proceedings of the 10th International Conference on Natural Language Generation*, 2017, pp. 124–133.

[11] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020, doi: 10.5281/zenodo.1212303.

[12] A. Soleimani, C. Monz, and M. Worring, “BERT for Evidence Retrieval and Claim Verification,” in *European Conference on Information Retrieval*, 2020, pp. 359–366.

[13] M. and S. E. AlGhamdi Kholoud and Shi, “Learning to Recommend Items to Wikidata Editors,” in *The Semantic Web – ISWC 2021*, E. and D. S. and F. A. and D. Y. and B. P. and H. A. and D. M. and A. H. Hotho Andreas and Blomqvist, Ed., Cham: Springer International Publishing, 2021, pp. 163–181.

Figures



beer (Q44)

alcoholic drink
brew

ProVe Score: 0.50

This item has 149 statements.

- **31 statements have internal references (not checked by ProVe).**
- **115 statements do not have references.** Click here to add references to them
- **3 statements have 4 external references, with the following verification results:**

Filters: ○ Refuting ○ Inconclusive ● Supportive ● Irretrievable

Statements	Relevant sentence in reference/Status code	Support stance ▲	Reference
beer has characteristic bitterness	Particularly hops determine to a great extent typical beer qualities such as bitter taste, hoppy flavour, and foam stability.	Supportive	https://d...
beer has part(s) hops	Often called the "spice" of beer, hops give beer bitterness, aroma and flavour.	Supportive	https://di...
beer disjoint union of list of valu...	HTTP Error code: 404	Irretrievable	http://bie...
beer has characteristic bitterness	HTTP Error code: 403	Irretrievable	https://d...

ProVe Score: 0.50

ProVe v1.1.1
Last update: 17/12/2024, 09:08:43
Non-authoritative: 0 (0.0%)
Irrelevant: 0 (0.0%)
Supports: 2 (50.0%)
Irretrievable: 2 (50.0%)

Figure 1 Main user-interface for Wikidata editing and infobox

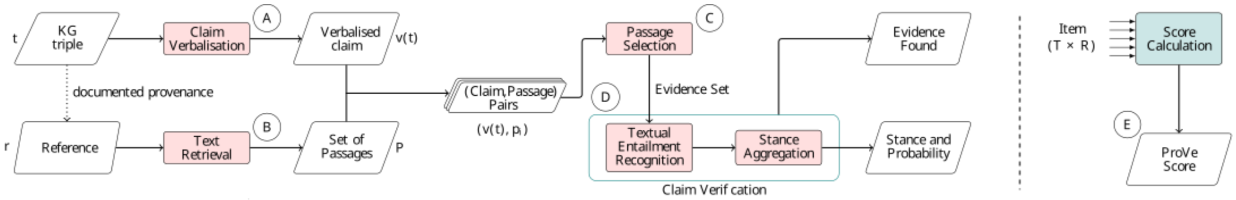


Figure 2 ProVe's main reference verification process