Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 14 Issue 03s, 2025

Claire.Ai - Hybrid NLP-Based Clause Recognition and Authentication System for insurance Policy Document Retrieval

[1]Parth Upadhye, [2]Sunil Wanjari, [3]Nakul Armarkar, [4]Khushi Choudhari, [5]Gouri Biswas

[1,2,3,4,5] Department of Computer Science and Engineering St. Vincent Pallotti College of Engineering & Technology Gavsi Manapur, Wardha Road, Nagpur – 441108, Nagpur,India

Email: [1]parthupadhye.22@stvincentngp.e du.in, [2]swanjari@stvincentngp.edu.in, [3]nakularmarkar.22@stvincentngp.edu.in , [4]khushichoudhari.22@stvincentngp.edu.in, [5]gouribiswas.22@stvincentngp.edu.in

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Conversational AI, Insurance Policies, Natural Language Processing, Document Retrieval, Policy Accessibility.

Abstract

CLAIRE.AI streamlines the process of extracting information from insurance documents by enabling users to ask questions in natural language. Since understanding policy details often requires domain expertise and can be challenging for policyholders, this system bridges that gap through an AI driven framework that interprets queries and delivers precise responses. The platform incorporates Google Gemini 2.5 Flash within a Retrieval-Augmented Generation (RAG) pipeline to manage various document types, including PDFs and Office files. It employs context-aware chunking, multilingual functionality, and vector-based retrieval to maintain accuracy and accessibility. Developed using FastAPI, the system supports asynchronous execution, token-based authentication, and input validation, ensuring both secure and efficient performance. Additionally, voice-input capability improves user interaction. Experimental results indicate that the solution provides faster and more reliable policy information compared to manual review. Users can easily clarify coverage, claims, and terms without requiring specialized knowledge.

Introduction

In the insurance industry, the ability to access accurate policy information at the right time plays a crucial role in shaping customer satisfaction, operational efficiency, and the overall outcome of claim processing. Policyholders frequently encounter situations where timely clarification of coverage, exclusions, or procedural requirements becomes essential—especially during emergencies or claim filing. However, traditional methods of retrieving information from insurance documents often prove difficult, as these documents are filled with dense legal terminology, lengthy clauses, and complex structures that are not easily understood by the average user.

This complexity creates a significant barrier for policyholders who do not possess legal or technical expertise. As a result, many individuals struggle to interpret their own policy documents, leading to confusion, delays, and dependence on customer service agents for even simple clarifications. To address this challenge, **CLAIRE.AI** has been designed as an intelligent document analysis and interpretation system. Its core purpose is to empower users with minimal insurance knowledge to easily and confidently retrieve the information they need.

CLAIRE.AI converts user queries expressed in natural, everyday language into precise insights extracted directly from policy documents. It is capable of handling complex policy structures, interpreting legal terminology, and delivering accurate responses that are easy to understand. To enhance trust and reliability, the system

incorporates hallucination detection and validation mechanisms, ensuring that the information provided is factual and consistent with the policy content.

In addition to accuracy, CLAIRE.AI emphasizes accessibility. The system supports speech-to-text input, enabling users to ask questions verbally rather than typing them. It also accommodates Hindi and multiple other languages, ensuring that people from diverse linguistic backgrounds can interact with insurance data without difficulty. As a result, even those unfamiliar with insurance documentation or legal wording can comfortably navigate their policy information. This makes the retrieval of coverage details, claim guidelines, and policy benefits not only possible, but intuitive, user-friendly, and highly efficient.

1. Significance

The significance of **CLAIRE.AI** lies in its transformational ability to bridge the gap between complex insurance documentation and the general public. Insurance policies are known for their legal language and intricate structure, which often discourages policyholders from fully understanding the details of their coverage. CLAIRE.AI directly addresses this long-standing issue by enabling users to ask questions in simple natural language and receive clear, accurate explanations derived from their policy documents.

By removing the need for specialized knowledge, the system democratizes access to insurance information. This is particularly valuable in real-life scenarios where quick decisions and clear understanding are required for example, during an accident, a hospitalization, or while verifying whether a particular treatment or damage is covered. In these situations, the ability to obtain immediate clarification can greatly reduce stress and guide policyholders to take appropriate actions.

The system becomes even more powerful through its support for **speech-to-text input**, allowing users to verbally ask questions instead of typing them. Its multilingual capability including Hindi and other regional languages ensures that people who are not fluent in English can still benefit from the platform. This inclusivity broadens the user base and ensures that insurance information is accessible to individuals across different regions, backgrounds, and literacy levels.

Overall, CLAIRE.AI not only simplifies insurance comprehension but also encourages informed decision-making. It empowers users to understand their policies thoroughly, reduces dependency on intermediaries, and enhances transparency within the insurance ecosystem. Whether the user is a policyholder, an agent, or someone seeking quick clarification, the system supports them with clarity, accuracy, and ease of use.

2. Problem Statement

Modern insurance policies have become increasingly complex, especially in comprehensive or specialized coverage plans. These documents often span multiple pages, contain numerous clauses, and rely heavily on legal terminology. Due to this complexity, ordinary policyholders face significant challenges when trying to extract relevant information, leading to misunderstandings, missed benefits, and slower claim processing.

Traditional policy analysis demands the ability to interpret legal language and navigate dense document structures—skills typically possessed only by insurance professionals. As a result, there exists a significant information gap between the content of the policy and the understanding of the policyholder.

This project aims to bridge that gap by developing an **AI-powered document analysis system** capable of interpreting natural language queries and generating precise, context-sensitive answers sourced directly from the policy document. By leveraging advanced natural language processing and modern large language models, the system ensures that users receive accurate interpretations without needing specialized knowledge. The problem addressed is therefore the difficulty policyholders face in understanding complex insurance documents, and the proposed solution is an intelligent system that simplifies information retrieval while maintaining accuracy and legal relevance.

3. Objectives

The primary objective of this project is to create a user-friendly system that enables seamless interaction with insurance policy content through natural language queries. The goal is to ensure that policyholders can understand their coverage, claim procedures, exclusions, and other policy-related information without difficulty.

The specific objectives are:

Model Development

- To develop a robust AI model using Google Gemini 2.5 Flash capable of understanding natural language queries related to insurance.
- To fine-tune or optimize the system for interpreting complex policy clauses, exclusions, and definitions.
- To ensure accuracy through validation

mechanisms and hallucination detection.
- User Interface Design
- To build an intuitive, web-based interface that supports both text and voice input for user queries.
- To integrate speech-to-text technology, allowing users to verbally ask questions.
- To support Hindi and additional languages, making the interface accessible to diverse user groups.
- Through these objectives, the project aims to create a platform that enhances insurance literacy and improves the overall user experience.

Review of Literature

Recent advancements in Artificial Intelligence (AI) and Large Language Models (LLMs) have significantly reshaped the insurance industry. Research has shown that AI-powered tools can streamline processes, reduce human workload, and improve accuracy across various insurance operations. The following studies highlight key contributions in this domain

Automation in Insurance Claims: Sivaraman et al. discuss the use of NLP to automatically extract crucial details from claims and policies. Their research shows that domain-trained models significantly reduce the time needed for manual verification while improving claim settlement accuracy.

LLMs in Risk Assessment and Fraud Detection: Kumar and Bhattacharya demonstrate how AI models detect fraudulent claims by identifying anomalies in vast datasets. Their work shows that fine-tuned LLMs outperform traditional rule-based systems, helping insurers save time and reduce financial losses.

Decision Support with Explainable LLMs: Zhang et al. emphasize the need for explainability in AI-driven insurance systems. Their frameworks allow LLMs to provide transparent reasoning behind their outputs, ensuring human oversight and regulatory compliance.

Conversational AI for Customer Engagement: Nguyen et al. highlight the use of conversational AI to engage policyholders, respond to inquiries, check claim statuses, and make recommendations. Their findings show significant improvements in customer satisfaction and operational savings.

Domain-Specific Fine-Tuning: Sun et al. showcase the benefits of fine-tuning general-purpose LLMs on insurance-domain datasets. Such models excel in tasks like claims summarization, compliance checks, and document interpretation.

AI-Driven Document Understanding: Li et al. present multimodal systems that combine text and image processing for tasks like motor or health claim verification. This approach proves especially useful where claims include supporting images or scanned documents.

Ethical and Regulatory Considerations: Pourreza and Rafiei address the ethical issues involved in using AI for insurance, including privacy, fairness, and bias reduction. They propose lightweight LLM frameworks designed to operate securely within sensitive datasets. Collectively, these studies demonstrate how AI and LLM technologies are becoming indispensable in the insurance industry. They not only speed up processes and reduce errors but also establish new standards for transparency, accuracy, and customer-centric operations. These insights lay the foundation for developing systems like CLAIRE.AI. performance in tasks like claims summarization, policy compliance checks, and document verification. Their results stress the need for specialized corpora and contextual embeddings to achieve domain accuracy.

AI-Driven Document Understanding: Li et al. investigate multi-modal AI systems combining text, images, and structured data for insurance claim verification. This approach is particularly relevant for motor and health insurance, where claim documents often include images (e.g., accident photos, medical reports). Their findings confirm that combining LLMs with vision models enhances fraud detection and accelerates claim approvals.

Ethical and Regulatory Considerations: Finally, Pourreza and Rafiei examine the ethical dimensions of AI in insurance, focusing on data privacy, bias mitigation, and compliance with financial regulations. They propose lightweight LLM frameworks optimized for sensitive data environments, balancing performance with ethical responsibility. The reviewed literature collectively demonstrates that the insurance sector is undergoing a paradigm shift through the adoption of AI and LLMs. From automated claim settlement and fraud detection to customer support and risk analysis, these technologies promise faster, more reliable, and transparent systems. However, the studies also highlight the necessity of domain-specific fine-tuning, explainability, and compliance mechanisms, which are crucial for deploying LLM-powered insurance solutions like CLAIRE.AI in real- world environments.

1. Feasibility Study

The feasibility of implementing CLAIRE.AI, an intelligent insurance document analysis system, is shaped by advancements in natural language processing (NLP), large language models (LLMs), and cloud-based computing infrastructures. A comprehensive feasibility study requires the evaluation of technical, economic, and operational aspects to determine the viability and impact of deploying such a system in the insurance domain. With the rise of state-of-the-art LLMs such as Google Gemini, OpenAI GPT, and LLaMA- based models, the practical implementation of insurance- focused document analysis has become achievable. These models are capable of interpreting complex legal language, policy clauses, and claim-related documentation, while also supporting features like hallucination detection and multi- lingual processing. Additionally, integration with cloud computing platforms allows for scalable deployment, ensuring low latency responses and high availability.

Economic Feasibility:
Implementing CLAIRE.AI demonstrates a strong potential for return on investment (ROI) by reducing operational overheads associated with manual claim verification, fraud detection, and customer service. Key benefits include cost reduction through automated policy interpretation, fraud minimization via AI-driven detection mechanisms, and cloud deployment models that minimize upfront costs.
Operational Feasibility: The user-focused design of CLAIRE.AI fits well within insurance workflows, where accuracy, transparency, and timely responses are essential. By allowing both policyholders and insurance professionals to ask questions about documents using natural language, the system reduces the need for specialized expertise and boosts operational efficiency.

2. Technical Feasibility
Evaluating the technical feasibility of CLAIRE.AI involves assessing the necessary infrastructure, tools, and resources for successful deployment. Thanks to recent advances in large language models (LLMs), natural language processing (NLP) workflows, and cloud computing, building such a system is well within reach.
Advances in Large Language Models: Modern LLMs like Google Gemini and OpenAI's GPT-4 are capable of parsng intricate insurance documents, supporting multiple languages, and generating contextually accurate answers. Techniques such as fine-tuning and transfer learning enable these models to specialize on insurance-specific data, improving their understanding of clauses, exclusions, and complex policy language**.**
Cloud Computing and Scalability: Cloud platforms such as AWS, Microsoft Azure, and Google Cloud provide scalable computational resources required for training and running these models efficiently. Technologies like containerization through Docker and orchestration with Kubernetes support scalable, fault-tolerant deployments.
Security and Compliance: To ensure regulatory compliance, the system must implement end-to-end encryption, strong authentication, and strict access controls. Comprehensive audit logs enable accountability and traceability for insurance claim decisions.

Proposed System
The proposed system is an AI-powered assistant designed specifically for the insurance sector. It understands user questions, retrieves relevant policy information, and delivers responses in natural language by utilizing Large Language Models fine-tuned on insurance data.
Key Components include:

- A web or chatbot user interface supporting natural language queries
- An NLP engine that processes and interprets queries, managing ambiguity and intent A centralized insurance knowledge base containing policies, procedures, and regulations, optimized for quick retrieval
- A module that translates queries into database searches or knowledge base lookups for accurate answer retrieval
- A response generator that produces clear, context- aware answers for ongoing conversations
- A security layer ensuring role-based access control and compliance with privacy regulations

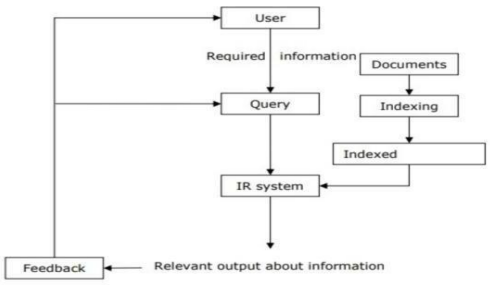

Figure 1. Proposed System Architecture

1. Architecture Overview Retrieval-Augmented Generation (RAG):
The system uses a Retrieval-Augmented Generation architecture to enhance response accuracy by combining document retrieval techniques with generative AI. Insurance policies

are divided into manageable segments that are transformed into vector embeddings and stored in a vector
database. When a query is received, the system retrieves the most semantically similar chunks from the database, ensuring that the answers generated remain relevant and grounded in the original policy text.

Fine-Tuning Large Language Models: A large language model like Google Gemini is fine-tuned on insurance-specific datasets including coverage terms, claim guidelines, exclusions, and legal jargon. This adaptation improves the model's understanding of domain-specific information and helps it generate reliable and precise answers.

Frontend (Next.js): The frontend is built using Next.js, providing an intuitive interface where users can submit queries using either text or voice. It supports document uploads in various formats including PDF and Word, and facilitates multilingual inputs to accommodate diverse users. **Backend (FastAPI):** FastAPI serves as the backend, coordinating requests from the frontend and managing the processing workflow. This includes tasks such as query parsing, retrieving relevant documents, and running model inference. FastAPI's asynchronous framework ensures the system responds quickly even during high load.

Databases (Vector and Relational): Document embeddings are stored in a vector database optimized for semantic search. A relational database is employed to handle user profiles, authentication credentials, and query logs, offering a balance between fast retrieval and secure data storage.

Coding & Implementation

Key Libraries and Frameworks

Hugging Face Transformers: A widely-used library for natural language processing tasks, offering pre-trained models and utilities to perform question answering, summarization, and text generation.

AutoModelForSeq2SeqLM: A class within the Transformers ecosystem designed for sequence-to sequence models, useful for generating text such as SQL queries or context-aware answers from insurance documents.

FastAPI: The framework powering the backend, managing API requests, orchestrating NLP workflows, database interactions, and delivering results.

Next.js: The React-based frontend framework that provides server-side rendering and dynamic interfaces.

SQLite3: A lightweight relational database selected to store structured insurance data including customer info, claims details, and policy documents.

1. Backend Implementation

The backend of CLAIRE.AI is developed using FastAPI, which facilitates fast, asynchronous handling of user queries. It connects to the Google Gemini 2.5 Flash model through a Retrieval-Augmented Generation (RAG) pipeline to process queries related to insurance policies. During startup, essential environment variables such as API keys and vector database configurations are loaded, and CORS middleware is set up to allow secure communication with the frontend. User inputs undergo validation via Pydantic models before being processed at the /ask endpoint.

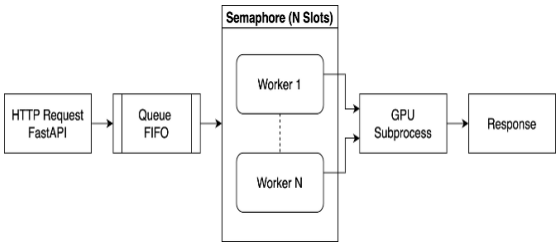

Figure 2. FastAPI Code Implementation

2. Model Training and Fine-Tuning

Fine-tuning customizes the Google Gemini model to the specific domain of insurance by training it on tailored datasets containing pairs of natural language questions and their corresponding responses. This training process helps the model learn industry-specific vocabulary, query structures, and contextual nuances.

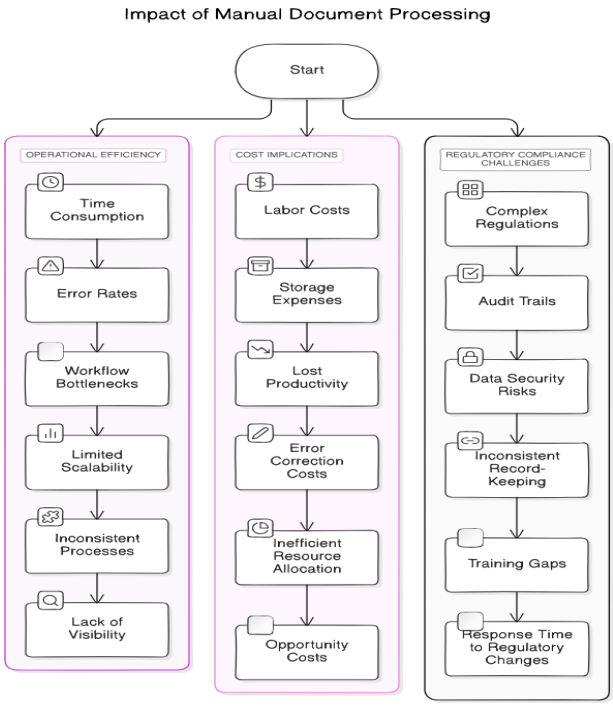

Figure 3. Model Loading Process Show Image

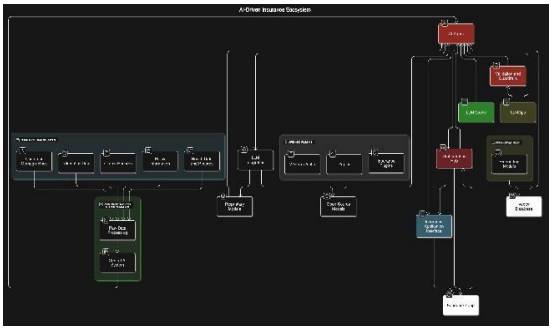

Figure 4. Model Training Implementation

3. Frontend Development

The frontend of the system is developed with Next.js, a React-based framework known for its support of server-side rendering and high-performance capabilities. It delivers a smooth and intuitive user interface where users can input queries either by typing or using speech.

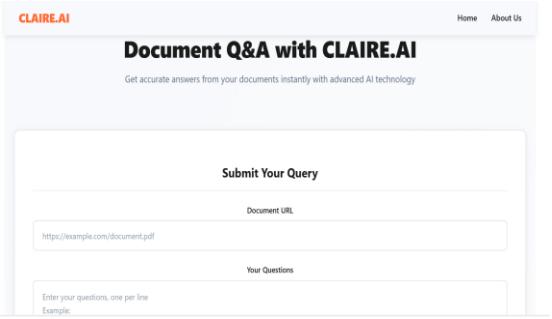

Figure 5. Frontend User Interface

4. Testing

API testing was conducted to ensure system reliability and performance. A set of natural language queries related to insurance policies, claims, and customer details was used, with each query paired with its expected output for validation.

Show Image

Testing Components:

1. Test Dataset: A set of natural language queries related to insurance policies, claims, and customer details was used. Each query was paired with its expected output for validation.

2. Functionality Testing: Core functionality was tested by entering different insurance-related queries and verifying that the generated responses executed correctly.

3. Edge Cases:

Queries with typos, ambiguous wording, and date-based requests were tested to ensure reliable handling.

5. Performance Metrics

Definitions:

- True Positives (TP): 85 (Queries correctly processed with accurate results)
- False Positives (FP): 5 (Incorrect responses yielding irrelevant results)
- False Negatives (FN): 10 (Valid queries not processed accurately)

Calculated Metrics:

Accuracy: 85/100 = 85% **Precision:** 85/90 = 94.44% **Recall:** 85/95 = 89.47% **F1 Score:** 91.73%

Technologies Used

Claire.AI utilizes a set of advanced technologies to simplify access to complex insurance information for non-technical users:

1. Retrieval-Augmented Generation (RAG): The system employs a RAG framework to improve the accuracy of answers by retrieving relevant information from insurance documents before generating responses.

2. Large Language Models (LLMs) – Gemini Model: Google's Gemini model is fine-tuned on insurance-specific datasets to understand natural language questions and generate correct outputs.

3. Frontend – Next.js: The user interface is built with Next.js, providing a responsive and dynamic platform where users can submit queries via text or voice.

4. Backend – FastAPI: FastAPI powers the backend, managing incoming requests, processing queries through the Gemini model, and interacting with the database.

5. Database – SQLite: SQLite stores insurance policies, claims, and related information with lightweight and easy-to- integrate design.

6. Speech-to-Text Functionality: To increase accessibility, Claire.AI allows users to submit voice-based queries.

7. Multilingual Support: The system accommodates multiple languages, including Hindi, to serve a diverse user base.

8. Security and Compliance: Robust security measures protect sensitive policyholder data and maintain regulatory compliance.

Conclusions And Future Scope

The development of Claire.AI marks a transformative advancement in making insurance policy information accessible to non-technical users. By enabling users to query complex insurance documents using natural language, the system bridges the gap between intricate policy details and user understanding. Leveraging Retrieval-Augmented Generation (RAG) and the Gemini model, Claire.AI accurately interprets user queries and extracts relevant information, ensuring timely and precise responses.

The system's performance has been validated through rigorous testing, demonstrating high accuracy, precision, and reliability in generating

correct outputs. This capability not only simplifies policy information retrieval but also supports informed decision-making by policyholders, reducing dependency on insurance experts and enhancing operational efficiency.

Future Scope

Enhanced Model Fine-Tuning: Continuous fine-tuning of the Gemini model with larger, domain-specific datasets can further improve accuracy and adaptability. Incorporating user feedback loops and active learning will help the system handle nuanced and complex queries more effectively.

Expanded Multilingual Support: Introducing additional languages, such as Spanish, French, and regional Indian languages, can broaden accessibility, enabling global and cross-cultural user engagement.

Integration with External Data Sources: Connecting Claire.AI with insurer databases, claim management systems, and document repositories can provide a more holistic view of policies, claims, and coverage.

Real-Time Query Processing: Adding real-time processing capabilities can enable immediate responses to user queries, enhancing efficiency in claim processing and policy verification.

Advanced Analytics and Insights: Future iterations could include dashboards, trend analysis, and predictive analytics to help users understand patterns in claims, coverage, and policy utilization.

Enhanced Security and Compliance: Implementing robust security measures, including encryption, user authentication, and adherence to insurance regulations and data protection standards, will ensure safe handling of sensitive policyholder information.

References

D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, pp. 1870–1879, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint* arXiv:1810.04805, 2018.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine reading comprehension," *Proc. Empirical Methods Natural Lang. Process.* (EMNLP), pp. 2383–2392, 2016.

W. Yang *et al.*, "End-to-end open-domain question answering with BERTserini," *Proc. NAACL*, pp. 72–77, 2019.

F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and reading: A comprehensive survey on open-domain question answering," *arXiv preprint* arXiv:2101.00774, 2021.

L. Kenton and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in neural network models for natural language processing," *J. Artif. Intell. Res.*, vol. 57, pp. 615–686, 2020.

X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Tech. Sci.*, vol. 63, no. 10, pp. 1872–1897, 2020.

S. Wang *et al.*, "R^3: Reinforced ranker-reader for open-domain question answering," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.