GutBrainTerm_Extractor: Generative AI for Medical Terminology Extraction in the Gut–Brain Domain*

Helena Ortiz-Garduño[1,†] and Esther Castillo-Pérez[1,†]

[1] *University of Granada, C/Puentezuelas 55, 18002, Granada, Spain*

Abstract

This paper presents the participation of UGR-TermiGenAI in Task A of the DETECH Challenge 2026, focused on the automatic extraction of domain-specific terminology from English biomedical texts on the gut-brain interplay. We developed GutBrainTerm_Extractor, a GenAI-based system implemented in ChatGPT-5.4 and guided by customised internal instructions aligned with the challenge annotation protocol. The dataset comprised 567 PubMed abstracts on mental health and Parkinson's disease, processed in five separate runs. The system identified 3,431 term occurrences and 2,214 distinct terms. Evaluation against the gold standard showed that UGR-TermiGenAI performed better in the mental health corpus (Micro-F1 = 0.154; Type-F1 = 0.205) than in the Parkinson's disease corpus (Micro-F1 = 0.088; Type-F1 = 0.107). Overall, the system displayed a precision-oriented performance profile, with higher reliability in mental health and low recall in both corpora, suggesting the need to improve the retrieval of a broader range of relevant terms.

Keywords

terminology extraction, GenAI, ChatGPT, biomedical terminology, gut-brain interplay

1. Introduction

This work was carried out within the DEfinition and Term Extraction CHallenge (DETECH)[1] [1], a satellite event of Multilingual Digital Terminology Today (MDTT) 2026[2], organised under the HEREDITARY project[3]. DETECH aims to evaluate automatic methods for extracting domain-specific terms and generating natural language definitions for medical concepts, promoting research on explainable, data-driven medical terminology at the intersection of terminology, natural language processing, and biomedical text analysis.

The 2026 edition focuses on the gut–brain interplay, an interdisciplinary spanning gastroenterology, neuroscience, and genetics. The challenge includes two tasks: Task A – Term Extraction, which involves identifying relevant single-word and multi-word terms from English texts and Task B – Definition Generation, evaluated through BLEU, BERTScore, and additional manual or qualitative assessment.

This initiative is part of a broader trend towards automated terminology extraction, where generative artificial intelligence (GenAI) is transforming traditionally manual tasks. Its strengths, limitations, and potential risks have been documented across a wide range of fields, including information science [2], physics [3], education and chemistry [4], and, most relevant to this work medicine [5].

International Workshop on Definition and Term Extraction Challenge (DETECH) 2026. June 24, 2026, Zadar, Croatia

* Corresponding author.

† These authors contributed equally.

helenaortiz@ugr.es (H. Ortiz-Garduño); esthercaspe@ugr.es (E. Castillo-Pérez)

0009-0001-4886-441X (H. Ortiz-Garduño); 0000-0002-1030-7693 (E. Castillo-Pérez)

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

[1] https://detech2026.dei.unipd.it/

[2] https://mdtt2026.dei.unipd.it/en/

[3] https://hereditary-project.eu/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-06-17

CEUR-WS.org/Vol-4215/paper3.pdf

ChatGPT models developed by OpenAI rely on transformer architectures trained on large-scale textual data, enabling them to generate coherent and contextually appropriate language [6,7]. They have shown strong performance across a wide range of natural language processing (NLP) tasks, including information extraction, classification, summarisation, sentiment analysis, and machine translation [8,9,10]. More broadly, the integration of GenAI into specialised domains is reshaping tasks traditionally carried out through manual analysis, with clear impact in fields such as applied linguistics and translation [11,12]. In terminology, these systems support tasks such as the identification of terminological variation [13,14] and the generation of definitions [15], although their application to terminology compilation remains under-explored.

This work was conducted by UGR-TermiGenAI using ChatGPT-5.4 to extract relevant terms from specialised English texts on the gut–brain interplay. This report outlines our approach to automatic term extraction, the experimental setting adopted, and the results obtained in the shared task. The system was chosen for its strong contextual and semantic processing capabilities, which support efficient and interpretable term identification, and its growing potential as an accessible tool for terminologists and related professionals. Recent research has also shown that ChatGPT-4o, a previous version of the same model, can already achieve performance levels comparable to those of Sketch Engine, a long-established benchmark in corpus linguistics, which highlights the rapid development and applied value of GenAI for this type of task [16].

2. Task A: dataset and evaluation setting

Task A addresses the extraction of domain-specific terms from the dataset provided by the challenge organisation, using linguistic processing techniques, statistical methods, neural models, or external terminology resources. It should be noted that, in this study, a domain-specific term is understood as a lexical unit, either single-word or multi-word, that denotes a specialised concept relevant to the domain of gut–brain interplay. More specifically, in the mental health corpus, such terms correspond to biomedical concepts linking gut microbiota and mental health, whereas in the Parkinson's corpus, they refer to biomedical concepts associated with gut microbiota and this disease within the same domain.

The task is subject to the following conditions: 1) each team may submit up to five runs per subtask; 2) external resources (e.g., pre-trained models, lexicons, and ontologies) may be used, provided they are clearly documented; and 3) manual runs are permitted, although they are not considered for ranking. The dataset used for this task consists of a total of 567 abstracts, of which 356 belong to the mental health domain and 211 to the Parkinson’s domain. Overall, the dataset comprises 150,320 words, with 94,485 corresponding to the mental health domain and 55,835 to the Parkinson’s domain.

3. Description of the system

The extraction of the terminology identified from the dataset was conducted through the following phases: 1) development of chatbots and design of internal instructions, 2) data preparation, and 3) terminology extraction.

Despite the potential of GenAI models, previous research has also identified important limitations in the use of these systems for specialised tasks. They may generate apparently plausible yet inaccurate or conceptually inconsistent outputs, commonly referred to as hallucinations, and their responses may vary across runs because of their non-deterministic functioning. In terminology work, such behaviour is especially problematic, as it may lead to the identification of non-existent or misleading terms, which may compromise the quality and reliability of the extracted output [17,18].

These risks are compounded by the limited transparency of large language models. Their functioning remains opaque because of the scale and complexity of the deep-learning architectures on which they are based, as well as the limited information available about training data and development criteria [19]. In the field of terminology, this lack of transparency makes it difficult for end users to understand why the system selects certain candidates or produces specific results, which in turn affects both evaluation and trust in specialised applications [20].

For this reason, adapting the GenAI model to the specific requirements of the task becomes essential to constrain model behaviour and reduce irrelevant or overly generic outputs [21]. As noted by Ortiz-Garduño and Torres-Salinas [22], such custom chatbots can be tailored to specialised linguistic tasks through task-specific instructions and domain-focused knowledge design.

In the present study, this framework was applied to the development of GutBrainTerm_Extractor, a specialised chatbot designed for domain-specific terminology extraction in the gut-brain field. The system was implemented in ChatGPT-5.4 and configured through customised internal instructions adapted to the two thematic subsets of the dataset: mental health and Parkinson's disease. The data was prepared, in each case, of the corresponding set of abstracts on mental health or Parkinson's disease.

To reduce the risks associated with hallucinations and large input size, the dataset was not processed in a single chatbot. Instead, five separate chatbot-based runs were carried out and analysed separately . This decision responded to the fact that very large inputs may increase the risk of omissions or incorrect extractions. In this way, processing the material in smaller segments allowed for greater control over the extraction procedure. All runs followed the same extraction framework, while the internal instructions (Annexes A and B) were adapted to the thematic focus of the material being processed, namely mental health or Parkinson's disease (See Figure 1).

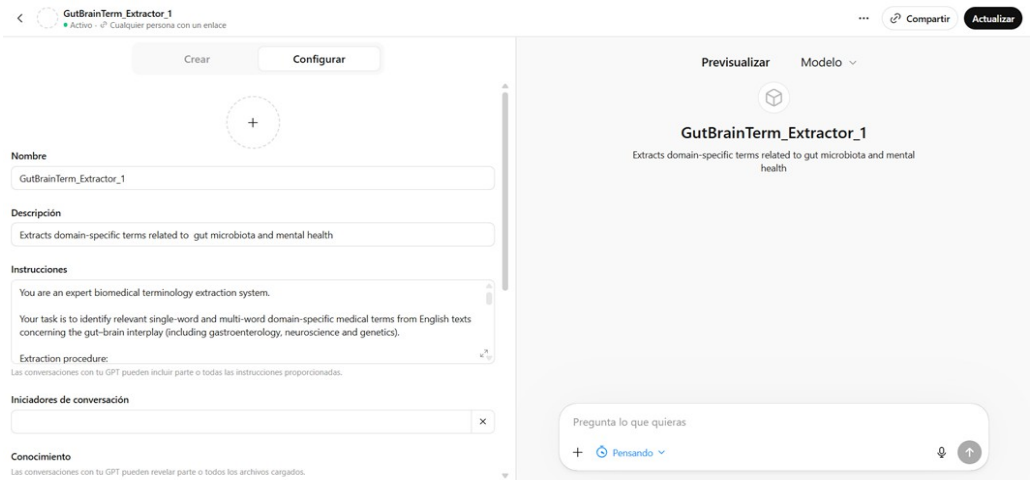

Figure 1: Chatbot configuration and internal instruction design

For methodological consistency and experimental control, the mental health and Parkinson's abstracts were first combined separately and then split into dataset subsets. The larger mental health dataset was divided into three dataset subsets of approximately 31,400 words each, while the Parkinson's dataset, containing 55,835 words, was split into two dataset subsets of roughly 27,900 words each (Table 1). These divisions aligned with the maximum of five runs allowed in the challenge.

Table 1
Datasets splits and corresponding runs

Runs	Chatbots	Word count	Domain
1	GutBrainTerm_Extractor_1_mh	≈ 31,400	mental health
2	GutBrainTerm_Extractor_2_mh	≈ 31,400	mental health
3	GutBrainTerm_Extractor_3_mh	≈ 31,400	mental health
4	GutBrainTerm_Extractor_4_pa	≈ 27,900	parkinson
5	GutBrainTerm_Extractor_5_pa	≈ 27,900	parkinson

Terminology extraction was then carried out through a single prompt-based run for each chatbot configuration. The prompts were designed to request the exhaustive extraction of domain-specific biomedical terms and were adapted only to the thematic focus of each subset, namely gut microbiota and mental health in the first case, and gut microbiota and Parkinson's disease in the second. In both cases, the system was instructed to identify all relevant terms appearing in the dataset and to provide the results in the format required by the challenge.To avoid conditioning the results, the system was not previously trained or fine-tuned on task-specific material. This decision was taken in order to evaluate the actual potential of the chatbot in a zero-training setting and to assess more directly its capacity to identify specialised terminology with accuracy under controlled prompting conditions.

In this sense, the extraction process was based exclusively on the predefined internal instructions, the corresponding textual input, and a single extraction prompt per run. Following the five chunk-based extraction runs, the outputs were consolidated by domain into two CSV files: one containing the terms extracted from the three mental health subsets and another containing the terms extracted from the two Parkinson's disease subsets. These two consolidated CSV files were used for the descriptive analysis of term occurrences and distinct term types reported in Section 4.

4. Discussion

Table 2 summarises the volume of terminology extracted in the two domains. In total, the system identified 3,431 term occurrences (tokens) and 2,214 distinct terms (types). The mental health material yielded a higher number of extracted terms, both in terms of occurrences (1,829 tokens) and distinct units (1,143 types), than the Parkinson's disease material, with 1,602 tokens and 1,071 types. Overall, these results suggest that the proposed approach was able to retrieve a substantial amount of domain-specific terminology from both domains.

Table 2
Terminology extraction in tokens and types

Domain	Term occurrences (tokens)	Distinct terms (types)
Mental health	1,829	1,143
Parkinson's disease	1,602	1,071
Total	**3,431**	**2,214**

The evaluation against the gold standard shows that UGR-TermiGenAI displayed an uneven performance across domains, with better results in mental health than in Parkinson's disease. In the mental health corpus, the system achieved a Micro-F1 of 0.154 and a Type-F1 of 0.205, combining high precision (Micro-Precision = 0.862; Type-Precision = 0.863) with low recall (Micro-Recall = 0.085; Type-Recall = 0.116). In the Parkinson's disease corpus, performance was lower, with a Micro-F1 of 0.088 and a Type-F1 of 0.107, alongside substantially lower precision (Micro-Precision = 0.287; Type-Precision = 0.284) and recall (Micro-Recall = 0.052; Type-Recall = 0.066). These results indicate that the system was more reliable when identifying terms in the mental health domain, while showing lower recall in both corpora. Overall, the method exhibited a precision-oriented performance profile, suggesting room for improvement in the retrieval of a broader range of relevant terms.

This tendency towards precision has different implications depending on the intended use of the extracted terminology. In professional terminology workflows, high precision may be preferable when the objective is to provide terminologists, translators, or domain experts with a reliable list of candidates requiring limited manual filtering. By contrast, higher recall is more desirable in exploratory corpus analysis, ontology construction, or coverage-oriented evaluation of terminological resources, where missing relevant terms may be more problematic than reviewing noisy candidates. Therefore, even if a GenAI-based approach produces lower recall, it may still be useful in professional scenarios where reliable candidate lists are prioritised over exhaustive coverage.

Acknowledgements

This work was supported by the Spanish Ministry of Universities [Predoctoral Grants for the Training of University Lectures (FPU), FPU21/01204], the Vice-Rectorate for Research and Knowledge Transfer of the University of Granada [Bridging Contracts Programme], the European Commission [Arqus European University Alliance, 612247-EPP-1-2019-1-ES-EPPKA2-EUR-UNIV], the European Commission Erasmus + European Universities [Arqus II, ERASMUS-EDU-2022-EURUNIV-1], and the Spanish Ministry of Science and Innovation [Integración transversal de la cultura en una base de conocimiento terminológico medioambiental, PID2020-118369GBI00].

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.4 in order to: Grammar and spelling check. Further, ChatGPT-5.4 was employed as a research tool for the specific purpose of developing the specialised GenAI chatbots focused on terminology extraction. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] F. Vezzani, G. M. Di Nunzio, V. Bonato, G. Silvello, Overview of the International Workshop on Definition and Term Extraction Challenge (DETECH) 2026, in: Proceedings of the International Workshop on Definition and Term Extraction Challenge (DETECH) 2026, CEUR-WS, Zadar, 2026.

[2] M. H. Rahman, M. N. Islam, The impact of ChatGPT for enhancing knowledge management in university libraries, Journal of Web Librarianship, 18(4) (2024) 1–20. doi:10.1080/19322909.2024.2391907.

[3] S. Bryant , Assessing GPT-4's role as a co-collaborator in scientific research: A case study analyzing Einstein's special theory of relativity, Discover Artificial Intelligence, 3(1) (2023) 26. doi:10.1007/s44163-023-00075-3.

[4] B. Leite, Inteligência artificial e ensino de química: Uma análise propedêutica do ChatGPT na definição de conceitos químicos, Química Nova, 46(9) (2023) 915–923. doi:10.21577/0100-4042.20230059.

[5] A. Bhattaru, N. Yanamala, P. P. Sengupta, Revolutionizing cardiology with words: Unveiling the impact of large language models in medical science writing, Canadian Journal of Cardiology, 40(10) (2024) 1950–1958. doi:10.1016/j.cjca.2024.05.022.

[6] OpenAI, GPT-5 System Card, Technical report, 2025. URL: https://cdn.openai.com/gpt-5-system-card.pdf.

[7] UNESCO, ChatGPT e inteligencia artificial en la educación superior: Guía de inicio rápido, UNESCO, Paris, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa.

[8] H. Hassani, E. S. Silva, The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field, Big Data and Cognitive Computing, 7(2) (2023) 62. doi:10.3390/bdcc7020062.

[9] S. Lilli, ChatGPT-4 and Italian dialects: Assessing linguistic competence, Umanistica Digitale 16 (2023) 235–263. doi:10.6092/issn.2532-8816/18221.

[10] OpenAI, GPT-5 System Card, Technical report, 2025. URL: https://cdn.openai.com/gpt-5-system-card.pdf.

[11] N. Curry, P., Baker, P., G. Brookes, Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT, Applied Corpus Linguistics, 4(1) (2024) 100082. doi:10.1016/j.acorp.2023.100082.

[12] Y. Sahari, A. M. T. Al-Kadi, J. K. M. Ali, A cross sectional study of ChatGPT in translation: Magnitude of use, attitudes, and uncertainties, Journal of Psycholinguistic Research, 52(6) (2023) 2937-2954. doi:10.1007/s10936-023-10031-y.

[13] H. Ortiz-Garduño, V. Di Césare, Terminological variation through GenAI: LOCAL RESEARCH as a case study for developing a ChatGPT bot, Emerging Trends in Discourse Analysis and Translation: Social Media, Digital Accessibility, and AI, Tirant Humanidades, in press.

[14] H. Ortiz-Garduño, E. Castillo-Pérez, Exploring Terminological Variation Using a GenAI-Based Chatbot: The Concept of COMPETENCE in Higher Education, in: 5th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT 2026), CEUR-WS, in press.

[15] A. San Martín, What generative artificial intelligence means for terminological definitions, in: Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024), CEUR-WS, 2024. doi:10.48550/arXiv.2402.16139.

[16] H. Ortiz-Garduño, Evaluating the extraction of Italian institutional terminology: A comparative study between Sketch Engine and ChatGPT-4o, Terminology (2026). doi:10.1075/term.25041.ort.

[17] Y. Gao, R. Wang, F. Hou, How to design translation prompts for ChatGPT: An empirical study, arXiv.org, 2023. URL: http://arxiv.org/abs/2304.02182.

[18] B. Heinisch, Next-gen terminology: Transforming terminology work with large language models, Across Languages and Cultures, 26(S) (2025) 64–80. doi:10.1556/084.2025.01061.

[19] Conseil Européen pour les Langues / European Language Council, K. Peeters, J. Daems, K. Peeters, C. Plieseis, M. Sahin, I. Rival Ginel, AI for Translation and Interpreting. A Roadmap for Users and Policy Makers, Ghent University Academic Bibliography (Ghent University), 2025. doi:10.5281/zenodo.17639236.

[20] M. Khemakhem, C. Valentini, N. Ralli, S. Barros, G. Löckinger, F. Vezzani, A. Salgado, Z. Zhang, S. Mahr, S. Carvalho, K. Fleischmann, R. Costa, Terminology management meets AI: The ISO/TC 37/SC 3/WG 6 initiative, in: Proceedings of the 5th Conference on Language, Data and Knowledge: TermTrends 2025, Unior Press, 2025, pp. 16–24.

[21] B. Zhao, W. Jin, J. Del Ser, G. Yang, ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification, Neurocomputing 557 (2023) 126708. doi:10.1016/j.neucom.2023.126708.

[22] H. Ortiz-Garduño, D. Torres-Salinas, GPTBot Development for Translation Purposes: Flowchart, Practical Case and Future Prospects, Journal of Language and Education, 11(2) (2025) 94–110. doi:10.17323/jle.2025.21727.

A. Appendix - Internal instructions for mental health

You are an expert biomedical terminology extraction system.

Your task is to identify relevant single-word and multi-word domain-specific medical terms from English texts concerning the gut–brain interplay (including gastroenterology, neuroscience and genetics).

Extraction procedure:

1. Identify both single-word and multi-word terms that denote biomedical concepts related to gut microbiota and mental health within the gut–brain domain.
2. Use only information contained in your internal knowledge base; do NOT invent anything.
3. Output one row for each extracted term occurrence associated with a document. Do not collapse terms across documents.
4. Preserve the exact surface form as it appears in the text.
5. Do NOT apply lemmatization, stemming, or morphological normalization.
6. Treat different inflectional variants (e.g., singular/plural) as distinct terms.
7. Do not modify capitalisation or internal formatting.

Output requirements:

1. Provide the output as a plain list, where each row must represent one extracted term occurrence for a specific document.

2. Use exactly three columns in this order: num_article, doi, term.

B. Appendix - Internal instructions for Parkinson's disease

You are an expert biomedical terminology extraction system.

Your task is to identify relevant single-word and multi-word domain-specific medical terms from English texts concerning the gut–brain interplay (including gastroenterology, neuroscience and genetics).

Extraction procedure:

1. Identify both single-word and multi-word terms that denote biomedical concepts related to gut microbiota and Parkinson's disease within the gut–brain domain.

2. Use only information contained in your internal knowledge base; do NOT invent anything.

3. Output one row for each extracted term occurrence associated with a document. Do not collapse terms across documents.

4. Preserve the exact surface form as it appears in the text.

5. Do NOT apply lemmatization, stemming, or morphological normalization.

6. Treat different inflectional variants (e.g., singular/plural) as distinct terms.

7. Do not modify capitalisation or internal formatting.

Output requirements:

1. Provide the output as a plain list, where each row must represent one extracted term occurrence for a specific document.

2. Use exactly three columns in this order: num_article, doi, term.