Assessing the Impact of CLEAR Prompt Engineering on LLM-based Term Extraction in Welsh and English*

Tomos Williams[1,*] and Sarah Cooper[2]

[1] *Language Technologies Unit, Bangor University, College Road, Bangor, Gwynedd, Wales*
[2] *School of Arts, Culture and Language, Bangor University, College Road, Bangor, Gwynedd, Wales*

Abstract

This study explores the effectiveness of the CLEAR (Concise, Logical, Explicit, Adaptive, Reflective) prompt engineering framework [1, 2] for extracting technical terms from Welsh texts using ChatGPT-5. In the contemporary digital landscape, language technologies for minoritised languages like Welsh continue to face challenges, particularly in the development of high-quality terminological resources and automated term extraction methods.

Using an experimental methodology, the study compares ChatGPT-5 outputs across four different prompting conditions: unstructured Welsh and English prompts, and structured prompts in both languages following the complete CLEAR framework. Extracted terms are evaluated against the Termiadur Addysg, a Welsh-language dictionary of educational terminology, focusing on the legal domain.

Results indicate that the use of the CLEAR framework is associated with a statistically significant increase in the proportion of terms that match entries in the Termiadur Addysg. No statistically significant difference was observed between Welsh and English prompt outputs. Qualitative analysis further suggests that CLEAR-based prompts are also associated with fewer erroneous, misspelled, and ambiguous terms than unstructured prompts.

Keywords

Prompt Engineering, CLEAR Framework, Large Language Models, Welsh Language, Minoritised Languages, Term Extraction

1. Introduction

In the contemporary digital age, language technologies have become integral to daily life. However, these technological developments have not been distributed equally across the world's languages. While language technologies for English have developed rapidly, minoritised languages like Welsh have faced significant challenges in remaining relevant in the digital age [3, 4]. One estimate published by UNESCO stated that 43% of the world's languages are endangered [5].

Recent research has shown that limited access to language technologies may contribute to risks for the long-term vitality of minoritised languages [6], as languages lacking digital resources may become less relevant to contemporary speakers [7]. Welsh is in a relatively favourable position compared to many other minority languages, due in part to strategic commitment by the Welsh Government and substantial investment in Welsh language technologies [8, 9]. Nevertheless, important gaps remain. One such gap concerns technical term extraction, a task central to the development of terminological resources, domain-specific dictionaries, and natural language processing systems [10].

The emergence of Large Language Models (LLMs) presents new opportunities to refine approaches to term extraction. In addition to the quality and size of training data, LLM performance is strongly influenced by the quality of instructions, or 'prompts' given to them [1]. As a result, Prompt Engineering (PE) has emerged as an active research area, with frameworks such as CLEAR [1, 2] proposing structured principles for producing high-quality outputs from LLMs.

* *5th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2026, June 25-26, 2026, Zadar, Croatia.*
[1,*] Corresponding author.
tom.williams@bangor.ac.uk (T. Williams); s.cooper@bangor.ac.uk (S. Cooper)

English has generally dominated in the field of Language Technology [3, 4, 5]. It is therefore important to quantify differences in LLM and PE performance in both English and minoritised languages.

Although CLEAR was originally developed for educational contexts, its potential for research tasks such as term extraction has not been widely examined, particularly in the context of minoritised languages. The framework consists of 5 attributes of an effective prompt, and are as follows: Concise (no unnecessary information), Logical (follows logical steps and is easy to follow), Explicit (the prompt must instruct the LLM as of exactly what to do), Adaptive (the prompt can be changed to produce better results), Reflective (the user must reflect on the results before adapting) [1].

This research therefore investigates whether the use of the CLEAR PE framework improves an LLM's ability to extract technical terms from Welsh texts. Specifically, the study seeks to answer two key research questions: (1) Does using the CLEAR framework increase the proportion of dictionary-validated terms compared with unstructured prompts? (2) Does prompt language (Welsh or English) influence the proportion of dictionary-validated terms produced by the model? The findings aim to contribute to ongoing discussions around the use of LLMs for terminology extraction and their integration into workflows for the development and maintenance of digital terminology resources.

2. Methodology

2.1. Data and LLM selection

The research used a corpus of past examination papers and marking schemes from the Welsh examination board WJEC, focusing on the legal domain. This domain was chosen because the Welsh Language Act 1993 requires specific public bodies to provide Welsh language services where possible, suggesting the LLM should have encountered Welsh legal training data as Chat-GPT is trained on the whole content of the internet [11]. The data comprised 21 files (.txt), 67,962 words, and 179,727 tokens, all Welsh language materials.

Tokens are units of text processed by the model, corresponding to words or subword units. ChatGPT's GPT-5 Tier 5 API was selected as the LLM due to its relatively high token rate limit (40,000,000) and context window (128,000 tokens) which allowed processing of large text size. The model was trained on web data published up to September 30, 2024 [11].

2.2. Prompt design

CLEAR was selected due to its structured, stepwise design, which allows controlled manipulation of prompt components and facilitates analysis of their effects on LLM output behaviour. Unstructured prompts were designed to imitate a non-expert user using LLM methods. Irrelevant information such as "please just what I asked for." were included to imitate normal non-expert usage of LLMs and to intentionally undermine the CLEAR 'concise' step. Welsh prompts were developed by machine translation, then were edited by the author to be a correct literal translation of the English prompt but not undermining correct Welsh syntax. CLEAR prompting is iterative in its nature due to the 'Adaptable' and 'Reflective' steps. The full CLEAR prompt therefore contained calls for, e.g., "established definitions" to improve the output from the LLM.

Four prompt conditions were developed representing two levels of CLEAR framework adherence in two languages. Firstly, unstructured prompts (Welsh and English) designed specifically not to follow CLEAR, mimicking how a non-expert user might communicate with an LLM. The study contrasts an optimised structured prompting framework (CLEAR) with minimally structured baseline prompts, rather than attempting to compare equivalently tuned or optimised prompting strategies. To develop the full CLEAR prompts, to begin with, CLE prompts were constructed to follow only the initial CLEAR components (Concise, Logical and Explicit). These prompts provided structured instructions but did not yet include the Adaptive and Reflective

elements of the framework. The full CLEAR prompts were produced by adapting the CLE prompts following reflection on their initial outputs, incorporating refinement in line with the adaptive and reflective principles of the framework.

Examples of the unstructured and final CLEAR prompts in English are provided below.

The English unstructured prompt example:

> "You are searching for Welsh terms. Find the terms in {text} that are Law terms and only give one term per line like in a shopping list. Don't give me headings or anything like that, please - just what I asked for."

The English CLEAR prompt example (final iteration)

> "You are searching for Welsh legal terms. Follow these steps:
> 1. Only consider terms in the field of Law.
> 2. Search for the technical terms in {text}.
> 3. Only include terms that are relevant to legal contexts.
> 4. Avoid general terms that do not have specific legal significance.
> 5. Only include terms that are commonly used in Welsh legal documents.
> 6. Only include terms that have established definitions in legal terminology.
> 7. If the term is in the plural form, change it to the singular form.
> 8. Consider your output and highlight any terms that do not fit the above criteria.
> 9. Removed the terms from step 8 from the list.
> 10. Only include one technical term per line.
> 11. Do not include bullet points or headings.
> 12. Do not duplicate terms in the list.
> 13. Only include the finalized list."

2.3. Evaluation methodology

Extracted terms were evaluated against the Termiadur Addysg (TA), an open-license Welsh terminological dictionary available digitally with domain-specific tags [12]. TA was selected as the resource for validation due to its status in Welsh-language education and its structured domain tagging, enabling systematic identification of law-specific terminology. Terms appearing in the TA with the appropriate domain tag were marked as 'dictionary-validated terms'. This represents a dictionary-based validation rather than annotation of all valid legal terms (see Limitations section).

Each prompt was executed 10 times to account for variation in LLM output and to enable statistical analysis. Each run was conducted independently in a fresh session, with no retained conversation context across API calls. The primary quantitative measure was the proportion of extracted terms that matched entries in the TA. This measure reflects dictionary validation rate rather than precision, as no gold-standard annotated dataset was available.

Statistical analysis employed paired t-tests to measure differences in dictionary validation rates between unstructured and CLEAR prompting conditions and between Welsh and English prompt languages.

In addition, qualitative analysis was conducted to identify patterns in error types, misspellings, and ambiguous terms (defined as not being specifically related to law but to many general fields within the education system, e.g., "Exam Number"). Error categorisation was performed through manual inspection of outputs. This analysis provides complementary insight into the output reliability beyond the dictionary validation rate.

3. Results

3.1. Number of extracted terms

Table 1 summarises the mean number of extracted terms and the mean percentage of dictionary-validated terms across all prompt conditions. A clear pattern emerges whereby the use of the CLEAR framework results in fewer extracted terms overall, but a higher proportion of dictionary-validated terms, suggesting increased selectivity in output. This indicates a shift from higher output towards greater sensitivity when structured prompt engineering following CLEAR is applied, in both Welsh and English.

Table 1
Terms extracted by LLM

Prompt	Mean number of terms extracted	Mean percentage valid terms (%)
Welsh	174.8	60.7
Welsh CLEAR	141.36	68.74
English	184.3	59.22
English CLEAR	106.9	71.12

3.2. Unstructured vs CLEAR

Paired t-tests (Table 2) showed statistically significant differences in the number of dictionary-validated terms extracted when comparing unstructured and CLEAR prompts in both Welsh and English. In both languages, CLEAR prompts produced a significantly higher proportion of dictionary-validated terms (despite the smaller overall extracted terms described in Section 3.1). These results indicate that adherence to the CLEAR framework is associated with improved term extraction quality compared to unstructured prompting.

Table 2
% Proportion of extracted terms matching Termiadur Addysg entries by prompt type

	Unstructured		CLEAR		t(9)	p	Cohen's d
	M	SD	M	SD			
Welsh	60.68	5.82	68.75	3.85	-4.0249	0.0030	-1.2728
English	59.24	4.62	71.12	6.98	-4.6316	0.0012	-1.4647

3.3. Welsh vs English

Paired t-tests (Table 3) compared the percentage of dictionary-validated terms extracted using Welsh and English prompts under both unstructured and CLEAR conditions. No statistically significant difference was observed in the percentage of dictionary-validated terms extracted by Welsh and English prompts. This was the case for both unstructured and CLEAR prompts. These findings indicate that term extraction via Welsh prompting is at a level comparable with English prompting, despite the differences in the data available for each language.

Table 3
% Proportion of extracted terms matching Termiadur Addysg entries by language condition

	Welsh		English		t(9)	p	Cohen's d
	M	SD	M	SD			
Unstructured	60.68	5.82	59.24	4.62	0.8533	0.4156	0.2698
CLEAR	68.75	3.85	71.12	6.98	-0.8918	0.3957	-0.2820

3.4. Qualitative findings

Qualitative analysis provides further details on the type of output produced in different prompting conditions. Unstructured prompts produced numerous erroneous terms, including terms not appearing in the original text, misspellings (e.g., 'Hunanynysgi' instead of 'hunanysgogi'), ambiguous terms (e.g., 'adran 12', 'Cod A'), and terms with incorrect word order with respect to the original text.

In contrast, CLEAR prompts produced progressively fewer errors. Outputs generated using the full CLEAR framework were largely free of non-existent or domain-inappropriate terms, with remaining discrepancies limited to morphological marking (singular forms of plural terms) or spacing variations. This qualitative reduction in error types suggests that CLEAR prompting not only increases the percentage of dictionary-validated terms, but also improves the overall usability of the terminology extracted.

4. Discussion

4.1. Framework effectiveness

The results indicate that following the CLEAR prompt engineering framework is associated with improved proportion of dictionary-validated candidate terms produced by the LLM. The observed increase in the proportion of dictionary-validated terms alongside reduced output volume, suggests that structured prompting encourages more refined and selected term generation.

Rather than maximising number of terms, CLEAR prompts appear to promote more selective term generation, increasing the proportion of terms that match entries in the reference dictionary. The reduction in erroneous, ambiguous and misspelled terms with CLEAR PE supports the interpretation that structured, logical prompts enable better LLM performance producing more reliable outputs. These results are consistent with previous research that structured, logical input produces more accurate output [13].

Taken together, the result suggests that CLEAR provides a practical and effective framework for term extraction tasks, and may offer benefits for the training of terminologists using LLMs as tools.

4.2. Language equivalence

The absence of statistically significant differences between Welsh and English prompt outputs suggests that prompt language alone did not strongly influence term extraction quality in this study. These findings challenge the assumptions that languages with less training data receive poorer LLM performance, at least within this specific task type.

One possible explanation is that, in the legal domain, sufficient Welsh language training data exists on the web for this LLM to perform effectively, potentially influenced by legislation such as the Welsh Language Act 1993.

An alternative interpretation is that prompt structure exerts a stronger influence on output quality than prompt language. Importantly, the findings do not suggest that Welsh poses challenges for LLMs, but rather that under-structured prompting conditions within a well-represented domain achieves outcomes comparable to English.

From a Language Technology perspective, this finding is encouraging, as it suggests that structured prompting conditions may help mitigate some of the disadvantages that minoritised languages may face in applied NLP tasks.

4.3. Limitations and future research

The study is exploratory in nature and focuses on prompt engineering effects on LLM-based term candidates rather than full automatic term extraction evaluation. As a result, this study has several limitations that also point towards directions for future research. Firstly, the focus on a single

domain (Law) with potentially atypically rich Welsh data may have influenced the results, particularly given the relatively strong presence of Welsh legal texts online. This limits the generalisability of the findings to other domains. Future studies should examine different subject areas with varying levels of Welsh-language data availability in order to assess whether the observed benefits of CLEAR prompting generalise across domains.

Secondly, the evaluation relied on publicly available examination materials and the TA as a benchmark. As these resources may have been encountered during the model's training, it is possible that prior exposure influenced the results. This introduces uncertainty regarding whether observed performance reflects prompt engineering effects or prior exposure to the evaluation resources. In addition, reliance on a single terminological resource introduces a bias towards dictionary coverage rather than comprehensive identification of valid domain terminology. Future research could address this by using controlled, newly created texts or by comparing results across different terminological resources to reduce reliance on a single source for evaluation which may not include all valid legal terms. Future work could also incorporate expert validation to produce more nuanced assessment of term validity.

Finally, although qualitative analysis identified a reduction in the erroneous outputs when using CLEAR prompts, a more fine-grained categorisation of error types was beyond the scope of this study. Future work could develop more detailed error analysis and quantitative error analyses to better characterise how prompt engineering strategies influence different types of extraction errors.

5. Conclusion

This research provides evidence that following the CLEAR prompt engineering framework can improve the selectivity of LLM-generated terms in a dictionary-validated evaluation setting based on Welsh texts. The results indicate that CLEAR-structured prompts yield a higher proportion of dictionary-validated terms than unstructured prompts, and that Welsh-language prompts produce results comparable to those obtained using English prompts.

The findings suggest that, with appropriate prompt engineering, Welsh-language users can achieve comparable performance to English in certain LLM applications, at least in domains with adequate training data. This has important implications for promoting Welsh language use in AI interactions and for developing language technology resources for minoritised languages more broadly.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-5 in order to redraft content and to translate from Welsh to English: the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. S. Lo, The art and science of prompt engineering: A new literacy in the information age, Internet Reference Services Quarterly 27 (2023) 203–210. doi:10.1080/10875301.2023.2227621.

[2] L. S. Lo, The CLEAR path: A framework for enhancing information literacy through prompt engineering, The Journal of Academic Librarianship 49 (2023). doi:10.1016/j.acalib.2023.102720.

[3] D. Prys, Preface, in: G. Watkins (Ed.) Language and technology in Wales: Volume 2, 1st ed., Bangor University, Bangor 2024, p. 5. URL: https://pure.bangor.ac.uk/ws/portalfiles/portal/75338920/Language_and_Technology_in_Wales_-_Volume_II_-_ISBN_978-1_84220-207-4.pdf.

[4] G. Watkins, Introduction, in: G. Watkins (Ed.) Language and technology in Wales: Volume 2, 1st ed., Bangor University, Bangor 2024, p. 6-7.

URL: https://pure.bangor.ac.uk/ws/portalfiles/portal/75338920/Language_and_Technology_in_Wales_-_Volume_II_-_ISBN_978-1_84220-207-4.pdf.
[5] A. Hursh, The importance of language rights in the information age, Vermont Law Review, 2015. URL: https://lawreview.vermontlaw.edu/wp-content/uploads/2020/07/Importance-of-Language-Rights.pdf.
[6] J. Żammit, Maltese risks digital extinction, The Sunday Times of Malta (2025) 22. URL: https://www.um.edu.mt/library/oar/handle/123456789/135354.
[7] A. Saxena, L. Borin (Eds.), Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology, De Gruyter Mouton, Berlin/New York, 2008. doi:10.1515/9783110197785.
[8] D. B. Jones, S. Cooper, Building intelligent digital assistants for speakers of a lesser-resourced language, in: C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, E. Wandl-Vogt (Eds.), Proceedings of the LREC 2016 Workshop "CCURL 2016. Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity", European Language Resource Association (ELRA), Portorož, Slovenia, 2016, pp. 74–79. URL: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf.
[9] Llywodraeth Cymru Welsh Government, Cyfraith Cymru, n.d. URL: https://cyfraith.llyw.cymru/.
[10] H. P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development 1 (1957) 309–317. doi:10.1147/rd.14.0309.
[11] OpenAI, OpenAI Developers, 2026, URL: https://platform.openai.com/docs/models/gpt-5.
[12] Prifysgol Bangor University, Y Termiadur Addysg, 2026, URL: https://www.termiaduraddysg.cymru/information/?lang=en.
[13] P. Törnberg, How to use large-language models for text analysis, SAGE Publications Ltd, London, 2024. doi:10.4135/9781529683707.