Identifying Sustainability in Public Tendering

Luca Rolshoven[1,2,*], Veton Matoshi[1], Tilia Ellendorff[3], Sarah Hostettler[1,3], Rahel Meili[1], Judith Binder[1] and Matthias Stürmer[1,2]

[1]Bern University of Applied Sciences
[2]University of Bern
[3]University of Zurich

Abstract

Public procurement serves as a significant lever for promoting sustainability, yet effectively assessing the integration of sustainability criteria within diverse and heterogeneous tender documents remains a challenge. This paper presents a Natural Language Processing (NLP) pipeline for automatically identifying sustainability criteria in Swiss public procurement documents written in German. To assess sustainability, we compiled four catalogs of official Sustainable Procurement Criteria (SPCs): three domain-specific (transport, food, furniture) and one domain-independent. Each call for tenders (CFT) document was segmented into sentences and encoded using a pre-trained sentence transformer. We then computed cosine similarity scores between each sentence and all SPCs, storing the top match from both the general and the domain-specific catalog, if applicable. While similarity scores were generally high for a majority of sentences, a preliminary manual inspection suggested that only matches with a score of 0.98 or higher tended to reflect meaningful alignment. To validate this threshold, two human experts independently reviewed 100 randomly sampled sentence-criterion pairs above this threshold. To explore whether this expert validation process could be scaled, we also prompted three different Large Language Models (LLMs) to assess the same samples, classifying each pair as a correct or incorrect match based on a majority vote. Our evaluation suggests that a similarity threshold of 0.98 is useful for reducing noise and identifying relevant sustainability criteria. LLM-based validation shows potential as a scalable alternative to human annotation, although performance varies between models. While Gemini 2.0 achieved substantial agreement with the expert judgments in terms of Fleiss' Kappa (κ = 0.754), other models demonstrated weaker alignment.

Keywords

Natural Language Processing, Sentence Similarity, LLM-as-a-Judge, Green AI, Sustainability, Public Procurement

1. Introduction

With an estimated global volume of USD 11 trillion annually [1], public procurement represents a significant lever for policy impact. In Switzerland alone, approximately CHF 41 billion is spent each year through public procurement [2]. At the time of writing, this corresponds to roughly 51 billion USD. As a result, prevailing procurement practices exert considerable influence, not only within the public sector but also across the private sector, where public entities can serve as role models [3]. Embedding sustainability considerations into public CFTs has the potential to foster more sustainable outcomes across both domains.

Recognizing this strategic potential, Switzerland has incorporated sustainability as a guiding principle in its revised procurement legislation [4], namely the Federal Act on Public Procurement (PPA) and the Intercantonal Agreement on Public Procurement (IAPP). However, assessing the effectiveness of these legal revisions remains challenging due to the heterogeneous nature of public CFTs, which often comprise diverse specifications and evaluation criteria dispersed across documents with varying formats and structures.

2nd Workshop on Green-Aware Artificial Intelligence, 28th European Conference on Artificial Intelligence (ECAI-2025), October 25-30, 2025, Bologna, Italy

*Corresponding author.
luca.rolshoven@bfh.ch (L. Rolshoven); veton.matoshi@bfh.ch (V. Matoshi); tilia.ellendorff@uzh.ch (T. Ellendorff); sarah.hostettler@bfh.ch (S. Hostettler); rahel.meili@bfh.ch (R. Meili); judith.binder@bfh.ch (J. Binder); matthias.stuermer@bfh.ch (M. Stürmer)
0009-0001-0663-9011 (L. Rolshoven); 0009-0002-6613-5701 (V. Matoshi); 0000-0002-8543-4902 (T. Ellendorff); 0009-0008-2214-2137 (S. Hostettler); 0000-0002-1185-2781 (R. Meili); 0000-0001-9038-4041 (M. Stürmer)

To address this challenge, we propose a simple NLP pipeline designed to identify and highlight sustainability-related requirements within public CFTs. From a Green AI perspective, our approach contributes to sustainability in multiple ways: (1) it enables systematic monitoring of sustainability adoption in public procurement at scale, (2) it uses lightweight sentence transformers rather than computationally expensive large models for the core matching task, and (3) it provides a foundation for automated sustainability compliance checking that could reduce manual assessment overhead.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the dataset and methodology, Section 4 details the evaluation setup with both human and automated assessments, and Section 5 concludes with a summary of findings and directions for future work.

2. Related Work

NLP techniques have been applied to identify sustainability in public procurement texts and other domains using both traditional and modern methods. Early approaches relied on keyword extraction and Elasticsearch to detect environmental, social, and economic criteria in French and Dutch tender documents from Belgium [5]. Similarly, in the context of Swiss tender documents, early efforts employed manual review and structured keyword-based approaches [6]. However, Welz and Stuermer [7] found that their automated keyword-based approach was not yet sufficiently reliable for robust, automatic monitoring of sustainable procurement activities.

Recognizing the limitations of purely keyword-driven methods, more sophisticated NLP techniques have been subsequently explored. Supervised models and embedding-based classifiers, for example, have since improved detection accuracy, particularly for classifying sustainability-related Q&A entries [8]. Chen et al. [9] applied advanced NLP techniques, including Word2Vec and Doc2Vec embeddings combined with classifiers such as SVMs and neural networks to predict Russell 1000 companies' alignment with the UN SDGs based on their CSR reports, achieving over 80% accuracy. More recently, Matoshi et al. [10] used LLM prompting to extract award criteria from Swiss CFTs and conducted a keyword-based sustainability analysis.

While previous work has explored keyword-based and supervised approaches for sustainability detection in procurement, our work specifically focuses on similarity-based matching against standardized sustainability criteria catalogs. This approach offers the advantage of leveraging official sustainable procurement recommendations while maintaining interpretability through explicit similarity scoring. Furthermore, by relying on a pre-trained, compact model and a simple similarity metric, our approach offers a scalable and accessible framework for analysis, helping to democratize the auditing of public documents without requiring extensive computational resources.

3. Data and Methodology

Sustainable Procurement Criteria Catalogs An expert with practical experience in both sustainability practices and public procurement manually compiled structured catalogs of SPCs. The SPCs were derived from authoritative sources, including the *Recommendations on Sustainable Public Procurement Criteria* issued by the Federal Office for the Environment (FOEN) [11, 12, 13, 14] and the *Green Public Procurement (GPP) Criteria* developed by the European Commission [15, 16, 17, 18], among others. These sources provide a comprehensive foundation for defining sustainability requirements for various product and service groups. To date, we have compiled criteria catalogs for the domains of **food** [11, 17, 14], **road transportation** (excluding railway) [19, 12, 15, 18, 14], and **furniture** [13, 16, 14]. In addition, a set of **general criteria** [20, 21, 22, 23, 24, 25] was compiled to capture cross-cutting sustainability principles that are not specific to a particular category of goods or services. Each domain-specific catalog is provided in the form of a separate Excel file, designed to be both human-readable and machine-readable. These files share a consistent structure to ensure interoperability and automated processing. Each entry in the catalogs contains the textual formulation of an SPC, its thematic area, a unique identifier, the source from which the SPC was derived, along with other metadata.

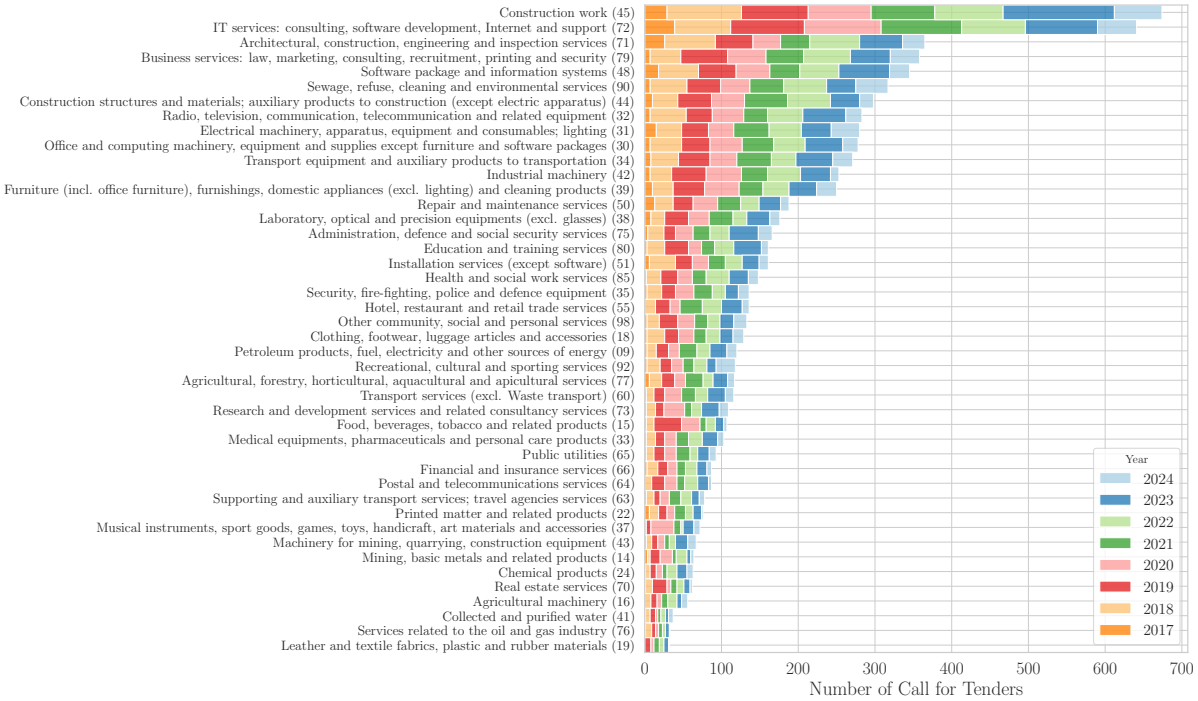

Figure 1: Number of CFTs in our sample for each of the 45 divisions. Since a CFT can span across several sectors, there may be multiple divisions assigned to a single CFT, resulting in a non-uniform distributions even when sampling CFTs uniformly across all divisions.

Document Selection All public CFTs need to be published on Simap, the official procurement platform of Switzerland.[1] This data has been processed and stored by the public procurement analytics platform IntelliProcure for several years.[2] For our research project, access has been granted to this database, which stores information about more than 60K CFT and contains over 2.2M indexed documents. These CFTs span across 45 different divisons, as indicated by the first two digits of the international Common Procurement Vocabulary (CPV) codes developed by the European Union. To make our initial baseline experiments more feasible, we selected a subset of 2.8K CFTs in German. We sampled one CFT per divison and month wherever possible between mid-2017 and mid-2024. The focus on CFTs in German allows us to evaluate the outputs of our pipeline more efficiently, but we plan to extend this approach to the other national languages of Switzerland, i.e. French and Italian. The date range was chosen such that we had CFTs both before and after the PPA and IAPP came into effect. Whenever a document is larger than 100MB, we did not consider it, since these large documents are usually construction plans or similar documents that are not relevant for our sustainability analysis. We consider the following file types: `docx`, `doc`, `xlsx`, `pptx`. Since a single CFT may span across multiple CPV codes, the distribution of our sample across divisions is not uniform, which can be seen in Figure 1.

Preprocessing First, we converted all Excel and Word documents to PDF format using the LibreOffice suite. We then extracted text content from these PDFs while preserving the page structure. For linguistic preprocessing, we applied a German spaCy model to segment the text into sentences.[3] Individual sentences were cleaned to remove unnecessary white spaces and standardize formatting. The preprocessed data is structured hierarchically, containing project metadata, document-level information (including file paths and external links), and sentence-level analysis with unique identifiers. This comprehensive preprocessing approach ensured our procurement documents were properly normalized and structured for subsequent analytical tasks.

[1] Available at: https://www.simap.ch/en
[2] Available at: https://intelliprocure.ch/
[3] spaCy model: `de_core_news_lg`

Sustainability Identification To identify the SPCs most closely aligned with individual sentences from a given CFT document, we computed sentence embeddings for both the document sentences and the SPCs using a compact German embedding model based on GBERT [26].[4] Based on the sector classification provided by the CPV code, we compared each sentence either to the combined set of sector-specific and general SPCs or, in cases where no sector-specific catalog was available, to the general catalog alone. For each catalog comparison, we retained the SPC with the highest cosine similarity score. To accelerate the computation, we parallelized the sentence pair comparisons by deploying multiple instances of the embedding model using Ray [27]. The pipeline output consists of one JSON lines file per CFT, with each line corresponding to a sentence from an attached document. Each entry contains metadata such as file path, page number, and a unique sentence identifier, along with the top-matching SPCs from each catalog and their associated similarity scores.

4. Evaluation and First Results

Threshold selection We sampled sentences with varying scores from multiple CFTs and evaluated them qualitatively. Data scientists familiar with procurement data assessed the accuracy of the SPC identification. This initial evaluation provided insight into the effectiveness of the automated detection and a preliminary understanding of the required score threshold for correct matches. Notably, we observe that high similarity scores (e.g. 0.96 or 0.97) often result in false positives. To gather meaningful feedback from sustainability experts, we aimed to provide a sample that included a substantial number of true positives or challenging false positives. Following the qualitative analysis, we set the threshold at 0.98. We aggregated all sentences from the CFTs with a similarity score >= 0.98, 1449 in total, to a SPC into a single Excel file to provide a familiar interface for the experts to use. This output is then used to solicit human feedback, ensuring that our experts review a relevant sample that helps refine the SPC detection process. The distribution of all similarity scores across the different catalogs is depicted in Figure 2.

Human Evaluation A human evaluation was performed to assess the quality of high-scoring matches. From the 1449 sentences achieving a similarity score of 0.98 or greater, a stratified random sample of 100 was selected. This sample was balanced across the four domains: transport, food, furniture, and domain-independent, with 25 sentences drawn from each. Two experts in public procurement and sustainability independently evaluated these sentences against their corresponding SPCs, labeling each match as *correct*, *incorrect*, or *unsure*.

The results revealed differences in the annotators' judgments. Annotator 1 labeled 69% of the matches as *correct*, while Annotator 2, being more lenient, labeled 80% as *correct*. Both experts agreed on the *correct* label in 65 of the 100 cases. The overall inter-annotator agreement was moderate, with a Cohen's Kappa of 0.51. However, when we excluded the 12 cases where at least one annotator was unsure (11 for Annotator 1, 2 for Annotator 2), the Kappa score on the remaining 88 samples rose to 0.69. This indicates that while defining *meaningful matches* is challenging and a source of uncertainty, the experts had substantial agreement on the clear-cut cases. Moreover, it highlights the inherent difficulty of the task and suggests that expanding annotation guidelines is necessary for future evaluations.

LLM-as-a-Judge While human evaluation remains the gold standard, it is not scalable. To explore whether we could extend this supervision signal, we employed an LLM-as-a-Judge approach [28]. Specifically, we prompted an ensemble of three different LLMs: Gemini 2.0 Flash, Llama 3.3 70B Instruct [29], and GPT-4o, to classify whether the 1449 sentence–criterion pairs with a cosine similarity score >= 0.98 constituted valid matches or false positives. Each model provided a binary judgment (*correct* or *incorrect*), and we aggregated their outputs using majority voting. The option *unsure* was excluded from this setup. We included summarized key points from annotation guidelines used in a separate

[4] Model ID on Hugging Face: `PM-AI/bi-encoder_msmarco_bert-base_german`

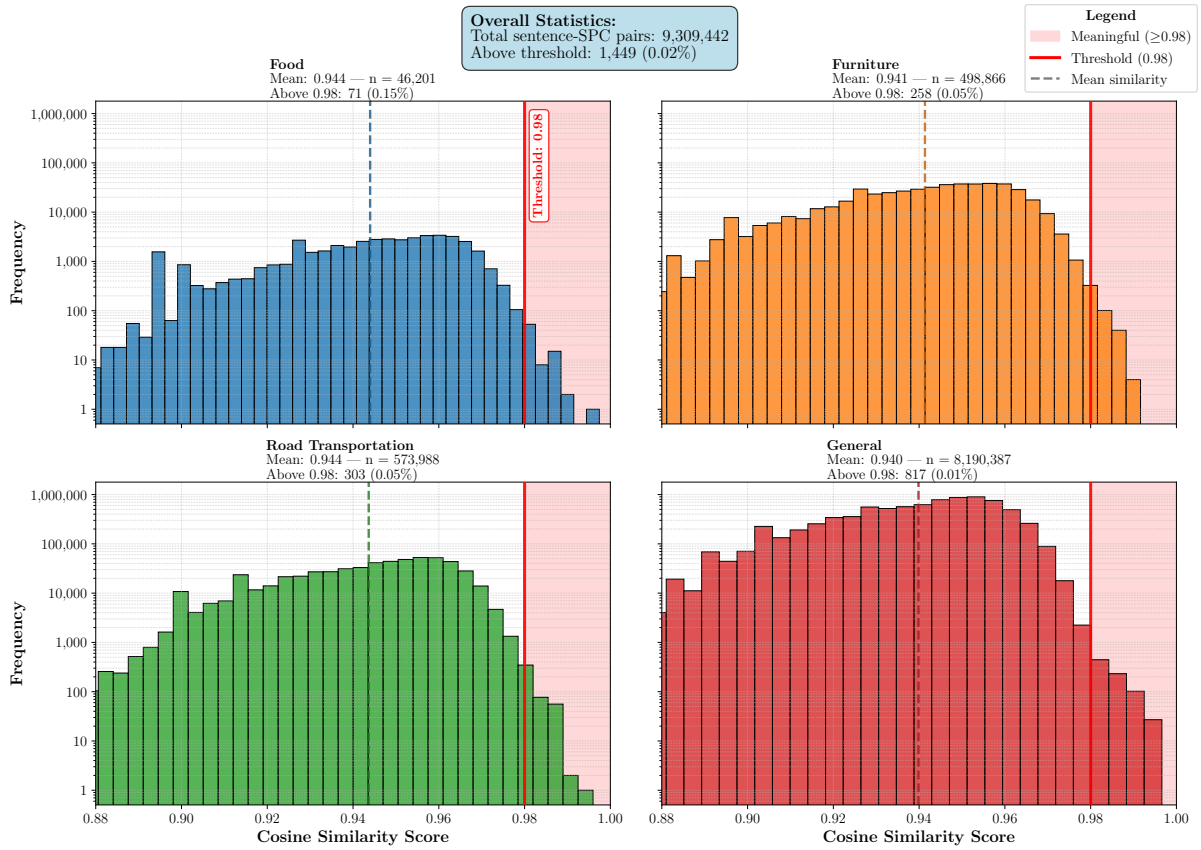

Figure 2: Distribution of the similarity scores between individual sentences and SPCs across the different catalogs. A threshold of 0.98 was chosen to drastically reduce the false positive rate of the matches.

expert-based sustainability annotation experiment in the German system prompt. A translated version of these instructions can be seen in Prompt 1.

To assess the alignment between the LLM judgments and human evaluation, we computed Fleiss' Kappa [30] on the subset of 88 sentence–criterion pairs that have been labeled by the human experts as either *correct* or *incorrect*. When considering the majority vote of the LLM ensemble as a single rater, the agreement with the human annotations yielded a moderate Fleiss' Kappa of 0.444. Treating each LLM as an individual rater, the score dropped to 0.366, suggesting variability in model agreement with the human judgments.

To better understand these differences, we further examined each model's agreement with the two expert raters individually. The results varied substantially: GPT-4o exhibited the weakest alignment with a Kappa of 0.172, while Llama 3.3 70B Instruct matched the ensemble score of 0.444. The strongest alignment was observed with Gemini 2.0 Flash, which achieved a substantial Kappa of 0.754. These findings underscore the importance of carefully selecting and evaluating models when adopting the LLM-as-a-Judge framework. The superior alignment of Gemini 2.0 Flash with human evaluations may reflect differences in instruction-following capabilities or a closer fit to this type of binary classification judgment, though further analysis is needed to confirm this hypothesis.

We also computed classification metrics for each LLM and the ensemble model. To ensure reliability, we restricted the evaluation to samples where the two annotators agreed and excluded cases in which both were uncertain. The resulting test set comprised 79 samples, of which 65 were labeled as *correct* and 14 as *incorrect*. Table 1 reports performance with respect to the positive class (correct). All models achieved perfect precision, but GPT-4o showed markedly low recall. By contrast, Gemini 2.0 Flash outperformed all other classifiers across all metrics, which is consistent with the observed inter-rater reliability and suggests it is a strong candidate for future experiments.

```
You are an AI model tasked with determining whether a single sentence from a call for tender
↪  expresses the same essential content as a given sustainability criterion.

Evaluate according to the following rules:

1. Core content is sufficient
A match is present if the sentence conveys the central claim of the criterion, even if the wording
↪  differs.

2. Flexible handling of numbers
- Numerical differences are acceptable as long as they do not significantly weaken the intended
↪  objective.
- A typical guideline is a similar order of magnitude (e.g., 70% ↔ 80%, “within a few hours” ↔
↪  “within six hours”) or a stricter requirement in the sentence.
- If a value deviates so strongly that the level of requirement is clearly lowered (e.g., 30% ↔ 80%
↪  or 48h ↔ 6h), it is not a match.

3. Reference sentences
Phrases like “see annex/appendix. . . ” without substantive content are not a match.

4. Legal minimum requirements / self-declarations
Statements about legal minimum requirements or self-declarations can be considered matches, as long
↪  as the criterion addresses exactly that obligation.

5. Incomplete sentences
If a sentence is incomplete, base your judgment on the given fragment only and do not assume how it
↪  might continue.

Respond strictly in the following format:

{
    "reasoning": "<Justification in no more than 4 sentences>",
    "decision": "<yes | no>"
}
```

Prompt 1: English translation of the German system prompt that was used during the LLM-as-a-Judge evaluation. The prompt requires the model to generate a reasoning trace before the final decision. This design is intended to (1) condition the decision on the rationale, which may steer the model's logits toward outcomes more consistent with the ground-truth labels, and (2) improve explainability, as decisions are grounded in the rationale rather than provided with only a post-hoc justification.

Table 1
Classification metrics for different LLM classifiers on a test set of 79 annotated sentence–criterion pairs. Reported values refer to the positive class *correct*, i.e., sentence–criterion pairs identified as true matches.

Classifier	Accuracy	Precision	Recall	F1-Score
GPT-4o	0.456	**1.0**	0.338	0.506
Llama 3.3 70B Instruct	0.722	**1.0**	0.662	0.796
Gemini 2.0 Flash	**0.975**	**1.0**	**0.969**	**0.984**
Ensemble	0.722	**1.0**	0.662	0.796

5. Conclusion and Outlook

In this work, we addressed the challenge of identifying SPCs in CFTs across multiple domains. Given the multi-faceted nature of sustainability, we compiled unified sustainability catalogs based on official sustainability recommendations, which served as a comprehensive knowledge base for SPC detection in CFTs. Our evaluation revealed that sentence transformers can produce meaningful results for SPC identification, though only when achieving similarity scores higher than 0.98, as confirmed by human evaluation. While leveraging LLMs-as-Judge with Gemini 2.0 Flash demonstrated high agreement with human annotators, further investigation is required to fully harness the potential of LLMs for effective SPC detection.

Future work will focus on three main directions. First, we will refine the annotation guidelines to address existing ambiguities, aiming to improve both inter-annotator agreement and the accuracy of the matching algorithm. Second, we plan to enhance the system itself by exploring hybrid approaches, combining efficient similarity search techniques for candidate retrieval with a more robust classification mechanism powered by an LLM. Third, we aim to broaden the scope of our work by compiling SPC catalogs for additional domains and extending our pipeline to support the other Swiss national languages, French and Italian, thereby increasing the societal impact of our research. Moreover, we aim to explore alternative heuristics for identifying relevant matches without manual inspection, for example by employing an LLM judge to evaluate top-ranked matches until a set number of *incorrect* classifications occur, based on the assumption that lower-ranked documents are even less relevant.

Acknowledgments

This research was funded by the Swiss National Science Foundation (SNSF) [10000100][5].

Declaration on Generative AI

In preparing this work, the authors employed GPT-4o, Claude 4 Sonnet, and Gemini 2.5 Pro, Llama 4 Maverick for paraphrasing, rewording, and text translation. We also accepted some of the Writeful assistant's suggestions on Overleaf whenever we saw fit. Some sources were converted from plain text into correct BibTeX entries using GPT-4o. All outputs were carefully reviewed and edited by the authors, who take full responsibility for the final content of the publication.

References

[1] M. Fazekas, J. R. Blum, Improving public procurement outcomes, Policy research working paper 2 (2021) 2–3.

[2] Federal Council, Botschaft zur Totalrevision des Bundesgesetzes über das öffentliche Beschaffungswesen (BBl 2017 1851) [Communication on the Complete Revision of the Federal Act on Public Procurement], 2017. URL: https://www.fedlex.admin.ch/eli/fga/2017/417/de.

[3] S. D. Sönnichsen, J. Clement, Review of green and sustainable public procurement: Towards circular public procurement, Journal of cleaner production 245 (2020) 118901.

[4] M. Steiner, D. Klingler, Switzerland · the revised swiss public procurement law: More quality and sustainability, European Procurement & Public Private Partnership Law Review 18 (2023) 87–93. doi:10.21552/eppl/2023/1/12.

[5] J. J. Grandia, P. M. Kruyen, Assessing the implementation of sustainable public procurement using quantitative text-analysis tools: A large-scale analysis of belgian public procurement notices, Journal of Purchasing and Supply Management 26 (2020). doi:10.1016/j.pursup.2020.100627.

[6] T. Welz, M. Stuermer, Sustainability of ict hardware procurement in switzerland: A status-quo analysis of the public procurement sector, in: Proceedings of the 7th International Conference on ICT for Sustainability, ICT4S2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 158–169. URL: https://doi.org/10.1145/3401335.3401352. doi:10.1145/3401335.3401352.

[7] T. Welz, M. Stuermer, 179 monitoring sustainable public procurement behaviour–demand-side analysis of public tenders in switzerland, Proceeding Paper: 20th European Round Table on Sustainable Consumption and Production (2021).

[8] Y. Torres-Berru, V. F. Lopez-Batista, L. C. Zhingre, A data mining approach to detecting bias and favoritism in public procurement, Intelligent Automation and Soft Computing 36 (2023) 3501–3516. doi:10.32604/iasc.2023.035367.

[5] More information at: https://data.snf.ch/grants/grant/10000100

[9] M. Chen, G. Mussalli, A. Amel-Zadeh, M. O. Weinberg, Nlp for sdgs: Measuring corporate alignment with the sustainable development goals, The Journal of Impact and ESG Investing 2 (2022) 61–81. URL: https://business.columbia.edu/faculty/research/nlp-sdgs-measuring-corporate-alignment-sustainable-development-goals.

[10] V. Matoshi, L. Rolshoven, M. Stürmer, Zero-shot award criteria extraction via large language models from german procurement data from switzerland, in: C. Corsin, C. Mark, W. Albert, M. Claudiu, M. Elisabeth, Z. Lucas (Eds.), Proceedings of the 9th edition of the Swiss Text Analytics Conference, Association for Computational Linguistics, 2024, pp. 113–133. URL: https://aclanthology.org/2024.swisstext-1.10.

[11] Bundesamt für Umwelt (BAFU), Empfehlungen für die nachhaltige öffentliche Beschaffung im Bereich Ernährung: Lebensmittel, Kücheneinrichtungen, Reinigungsmittel und Dienstleistungen der Gemeinschaftsgastronomie [Recommendations for Sustainable Public Procurement in the Field of Nutrition: Food, Kitchen Equipment, Cleaning Products and Catering Services], 2020. URL: https://woeb.swiss/de/documents/empfehlungen-fuer-die-nachhaltige-oeffentliche-beschaffung-im-bereich-ernaehrung.

[12] Bundesamt für Umwelt (BAFU), Personenwagen und leichte Nutzfahrzeuge. Empfehlungen und Kriterien für die öffentliche Beschaffung. Toolbox Nachhaltige Beschaffung Schweiz [Passenger Cars and Light Commercial Vehicles: Recommendations and Criteria for Public Procurement], 2021. URL: https://woeb.swiss/de/documents/merkblatt-personenwagen-und-leichte-nutzfahrzeuge-toolbox-teil-c.

[13] Bundesamt für Umwelt (BAFU), Möbel. Empfehlungen und Kriterien für die öffentliche Beschaffung. Toolbox Nachhaltige Öffentliche Beschaffung [Furniture: Recommendations and Criteria for Public Procurement], 2023. URL: https://woeb.swiss/de/documents/merkblatt-moebel-toolbox-teil-c.

[14] Bundesamt für Umwelt (BAFU), Relevanzmatrix [Relevance Matrix], 2020. URL: https://www.bafu.admin.ch/bafu/de/home/themen/wirtschaft-konsum/oekologische-oeffentliche-beschaffung/relevanzmatrix.html.

[15] European Commission, Green Public Procurement Criteria for Road Transport. German Version 1.1, 2018. URL: https://circabc.europa.eu/ui/group/44278090-3fae-4515-bcc2-44fd57c1d0d1/library/2f708829-3ad1-4215-bd07-48ffe87b989e/details.

[16] European Commission, Green Public Procurement Criteria: Furniture. German Version 1.0, 2018. URL: https://circabc.europa.eu/ui/group/44278090-3fae-4515-bcc2-44fd57c1d0d1/library/5f19d063-d5ce-4f76-8200-1b1d072c1334/details.

[17] European Commission, Eu gpp Criteria for food, catering services and vending machines. German Version 1.0, 2023. URL: https://circabc.europa.eu/ui/group/44278090-3fae-4515-bcc2-44fd57c1d0d1/library/150b4fbe-53b7-4f98-b216-e0fa3f01c7fe/details.

[18] European Commission, Green Public Procurement Criteria for Road Lighting and Traffic Signals. German Version 1.0, 2023. URL: https://circabc.europa.eu/ui/group/44278090-3fae-4515-bcc2-44fd57c1d0d1/library/8db3c9ce-f7a6-45a1-a0dd-6a06ad97999f/details.

[19] Bundesamt für Umwelt (BAFU), Busse und Kommunalfahrzeuge. Empfehlungen und Kriterien für die öffentliche Beschaffung. Toolbox Nachhaltige Beschaffung Schweiz [Buses and Municipal Vehicles: Recommendations and Criteria for Public Procurement], 2023. URL: https://woeb.swiss/de/documents/merkblatt-busse-und-kommunalfahrzeuge-toolbox-teil-c.

[20] International Organization for Standardization, Iso 20400:2017 Sustainable Procurement – Guidance. Reviewed and confirmed in 2023, 2017. URL: https://www.iso.org/standard/63026.html.

[21] Geschäftsstelle der Beschaffungskonferenz des Bundes (BKB), Nachhaltige Beschaffung [Sustainable Procurement], 2021. URL: https://woeb.swiss/de/documents/nachhaltige-beschaffung-empfehlungen-fuer-die-beschaffungsstellen-des-bundes.

[22] R. Koch, L. Biehl, Die soziale nachhaltigkeit im öffentlichen beschaffungswesen: Wie können städte und gemeinden die ziele der sozialen nachhaltigkeit und des fairen handels bei öffentlichen beschaffungen umsetzen? [Social Sustainability in Public Procurement], 2025. URL: https://arbor.bfh.ch/handle/arbor/44968. doi:https://doi.org/10.24451/dspace/11711.

[23] Beschaffungskonferenz des Bundes (BKB), Faktenblatt Lebenszykluskosten. Begriffsklärung und

Einsatzmöglichkeiten bei öffentlichen Beschaffungen von Gütern und Dienstleistungen [Lifecycle Costing Factsheet: Definitions and Applications in Public Procurement], 2023. URL: https://www.bkb.admin.ch/de/hilfsmittel#Faktenbl%C3%A4tter.

[24] Beschaffungskonferenz des Bundes (BKB), Monitoring Umsetzung Beschaffungsstrategie der Bundesverwaltung [Monitoring the Implementation of the Federal administration's Procurement Strategy], 2024. URL: https://www.bkb.admin.ch/de/beschaffungsstrategie.

[25] Bundesamt für Umwelt (BAFU), Verschiedene Toolboxen Nachhaltige Beschaffung Schweiz [Various Toolboxes for Sustainable Procurement Switzerland], 2023. URL: https://woeb.swiss/de/toolbox.

[26] B. Chan, S. Schweter, T. Möller, German's next language model, arXiv preprint arXiv:2010.10906 (2020).

[27] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, et al., Ray: A distributed framework for emerging {AI} applications, in: 13th USENIX symposium on operating systems design and implementation (OSDI 18), 2018, pp. 561–577.

[28] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2023) 46595–46623.

[29] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[30] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.