Annals of Work Exposures and Health, 2025, **00**, 1–6
https://doi.org/10.1093/annweh/wxaf081
Advance access publication date: 24 November 2025
Short Communication

The Chartered Society for Worker Health Protection
OXFORD

Implementing generative pretrained transformer models for text recognition tasks in safety data sheets

Floris Pekel[1,*], Gino Kalkman[1], Erik Lemcke[1], Robin van Stokkum[1], Anjoeka Pronk[1], Lode Godderis[2,3], Janne Goossens[2,3], Hilde De Raeve[2], Eddy Coene[2], and Eelco Kuijpers[1]

[1]TNO, Department of Risk Assessment, Prevention, Innovation and Development, Netherlands Organization for Applied Scientific Research, Princetonlaan 6, 3584CB Utrecht, The Netherlands
[2]IDEWE, External Service for Prevention and Protection at Work, Knowledge, Information and Research Department, Interleuvenlaan 58, Leuven 3001, Belgium
[3]Centre for Environment and Health, KU Leuven, Herestraat 49, Leuven 3000, Belgium

*Corresponding author: Email: floris.pekel@tno.nl

Abstract

Workplaces handling chemicals require an up-to-date and comprehensive assessment of the potential risks for their workforce. Online safety data sheets (SDSs) inventories provide adequate information to perform risk assessments. However, current practices that manually import information from SDSs into the online inventories are time-consuming, leading to delayed or inadequate risk assessments. This study presents a pipeline using large language models (LLMs) to automate the extraction and management of data from SDSs to online chemical inventories. The pipeline achieved an average accuracy of 0.83 in (close to precisely) extracting multiple variables of interest, such as company name, product name, and hazard statements, in comparison to manually extracting these variables. Overall, this pipeline illustrates the ability of LLM tools to automate SDS inventory management and thereby support the possibility to perform up-to-date risk assessments and evaluation tasks on the work floor, ultimately contributing to occupational safety.

Keywords: large language models; occupational safety; safety data sheets; text extraction.

What's Important About This Work?

This study produced a software pipeline involving large language models (LLMs) that semiautomate the transfer of safety data sheet information into online chemical inventories. While a level of manual control remains necessary, this work shows that LLMs may be able to automate information extraction, which is promising and could support occupational health risk assessments.

Introduction

Ensuring compliance with (inter)national regulatory standards is crucial for safeguarding occupational health and safety. European Union legislation mandates the use of safety data sheets (SDSs) for all occupations involved in handling chemical substances (European Comission. 2006). As new products are continuously introduced at a company level, the need for accurate, accessible, and regularly updated safety information remains important for determining chemical risks and thereby preventing occupational health outcomes (NIOSH 2009; Nayar et al. 2016; Demasi et al. 2022; Otten et al. 2022).

In practice, companies face challenges in staying up to date with the newest versions of SDSs. For larger companies, the complexity is amplified by using a wide range of chemical products, resulting in vast databases of dozens to hundreds of SDSs. The current practice of manually transferring SDS information into

Received: March 17, 2025, Accepted: October 27, 2025

online chemical inventories is time-consuming, which can lead to delays and incomplete chemical inventories, possibly hampering efficient risk assessment in the workspace. One of the main efforts to tackle these challenges surrounding risk assessment and subsequently prevention focuses on automation of data (eg SDSs) management processes, which improves efficiency and provides a comprehensive overview of chemical inventories.

While automated data management processes exist, these are often based on pattern and format recognitions (eg regular expressions), which limit their usability for unstructured data files such as written text (Bartoli et al. 2016; Suman et al. 2024). Recent advancements in large language model (LLM) technology have notably increased their natural language processing capabilities, including text extraction, summarization, and question answering (Raiaan et al. 2024; Usman Hadi et al. 2024).

The introduction of LLM frameworks like Google's BERT (Bidirectional Encoder Representations from Transformers), OpenAI's generative pretrained transformer (GPT), and Meta AI's Llama has led to numerous applications supporting occupational tasks, such as data analyses or customer support (Devlin et al. 2019; Touvron et al. 2023; Usman Hadi et al. 2024). Clear examples encompass the use of LLMs in patient services (Javaid et al. 2023), sentiment analyses and report generation in the finance sector (Araci 2019), or summarizing biomedical papers with BioBERT (Lee et al. 2020). Within the occupational risk assessment domain, the use of LLMs for extracting information from textual documents, such as SDSs, seems promising (Suman et al. 2024).

Here, we developed and evaluated a pipeline using OpenAI's GPT models (OpenAI 2024a) to extract text-specific information from SDSs and transform it into a usable format suitable for SDS inventory software. We also explore the limitations of the pipeline and discuss the potential of multiple optimization strategies.

Method and pipeline

The following variables were selected for text extraction from SDSs: company name, product name, signal word, SDS version number, release date, hazard statements (H-phrases), and physical status. Together, these variables provide the required data to identify the SDS and extract useful information for the inventory.

Performance of the pipeline was assessed by implementing a matching algorithm that compares the LLM output against the manually extracted data, which is the standard for extracting SDS data. A binary scoring system is used where, originally, a match is classified "good" if the LLM output is fully identical to the manually extracted variable. A single deviation from the manual output classifies the LLM output as "no" match. The accuracy per variable is defined as the number of correct predictions (sum true negative and true positives cases) divided by the total number of cases.

In total, the pipeline underwent 3 iterations (IT1 to IT3) to optimize the process of correctly extracting the variables. Improvements were made in (i) the selection process of the context (ie the correct chapter), (ii) adding precleaning steps on the selected context before entering it into the LLM, (iii) refining the task description of the prompt, since it influences the output of a LLM (Chen et al. 2023), and (iv) adding multiple post-process cleaning steps (further specified below) of the LLM output.

For IT1 and IT2, the same SDS database was used ($N = 470$). For IT3, a new SDS dataset ($N = 462$) was used to ensure the pipeline adjustments were not overfitted for the initial dataset. The SDS databases, and a manually indexed database, were obtained from an online SDS inventory which is managed by an Occupational Health & Safety Institute.

SDSs in PDF format were processed using the PyMuPDF library (PyMuPDF 2024) to extract text line by line and convert it to string format. Next, we use regular expressions for selecting relevant sections in the SDS for each of the variables mentioned above. Since SDSs have a standardized structure, these variables can always be found in the same sector of a document. OSHA describes the required structure of a SDS extensively, which we recommend readers to consult for a comprehensive overview (Standard and Sheets 2006). In short, each SDS comprises of 16 sections which are named: "Identification", "Hazard Identification", "Composition/Information of Ingredients", "First-Aid Measures", "Fire-Fighting Measures", "Accidental release measures", "Handling and storage", "Exposure controls/personal protection", "Physical and chemical properties", "Stability and reactivity", "Toxicological information", "Ecological information (Non-mandatory)", "Disposal considerations', "Transport information (Non-mandatory)", "Regulatory information (Non-mandatory)", and "Other information". These section names were used as markers to identify the relevant context for the LLM (Fig. 1).

For each variable, a specific prompt (ie description of the task to be executed by the LLM), paired with the relevant context (ie, the SDS section in which the variable is expected to be present), was put into the LLM. The GPT3.5-turbo (GPT3.5) and GPT4-*omni* (GPT4-o) models developed by OpenAI (Brown et al. 2020; OpenAI 2024b) were used to identify the variable of interest within the selected context.

In the first iteration (IT1), no cleaning steps of the selected context or postprocessing of the output occurred to deliberately observe the LLM's ability to provide the desired information with minimal modification.

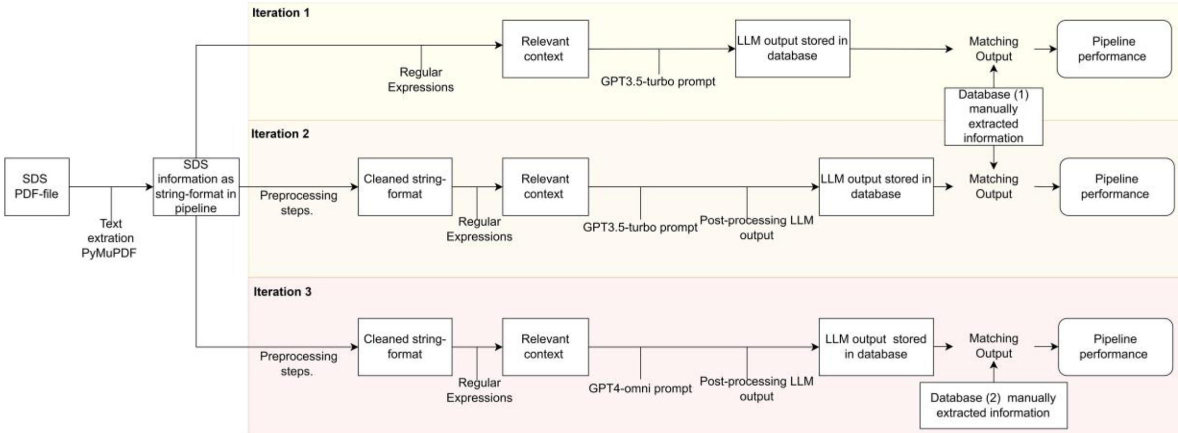

Fig. 1. Graphical representation of pipeline development and evaluation. In total, 3 iterations were performed to increase the pipeline performance. Abbreviations: GPT, generative pretrained transformer (model); PyMuPDF, a Python library that enables the import of PDFfiles. Graphical representation created using Drawio software - http://drawio.com.

In the second iteration (IT2), we added preprocessing steps that remove newline and whitespace characters from the strings, refined the prompts to obtain the output in a format that is similar to the format required in online chemical inventories (eg "provide one word instead of a sentence describing the word"). Moreover, we expanded the number of variables to be extracted from the SDS with the variable "physical status." Finally, after examining IT1 results, we added postprocessing steps and adjusted the matching algorithm, allowing slight deviations from the manual answer to be considered a "good" match. Examples for these postprocessing steps are as follows:

- Company name: If trademark symbols (eg ™ or ®) were missing or present, while the manual output showed the opposite, we determined it to be a "good" match.
- Signal words: If the LLM-output provided punctuation while the manual output did not (eg "*Danger.*" vs "*Danger*"), we deleted the punctuation mark during the postprocessing.
- Release date: Date formats can contain different punctuation marks (eg slash, hyphen, or a period). We transformed all punctuation marks to a hyphen (−), so that the matching algorithm recognized it as a "good" match.

Iteration 3 (IT3) contains a final optimization of prompt description (ie "provide output in Dutch") and contains additional variables of interest (CAS numbers and the product composition). Moreover, we applied the newer GPT4-o version, which was released during pipeline development. Prompts of iteration 3 can be found in the Supplementary material.

Results and discussion

Table 1 shows the pipeline's progression with each iteration, with the most noticeable improvement happening between IT1 and IT2 as average accuracy increased from 0.39 to 0.75, while IT3 shows slightly improved accuracy scores (average = 0.83) compared to IT2. These results show that minor adjustments in postprocessing of the LLM output and better prompt descriptions may significantly improve accuracy. In general, prompt refinement should be considered a main improvement, since it has a high impact on the provided LLM output (C. Li et al. 2021; P. Li et al. 2021; Y. Li et al. 2021; Suman et al. 2024). Identifying which factors contributed most to the increased accuracy is challenging, since they were simultaneously applied and differed between variables. The performance increase between IT2 and IT3 can largely be attributed to the implementation of the newer GPT4o model, which has been shown to outperform earlier versions on academic benchmarks (Achiam et al. 2023). However, looking at individual variables, we can see that IT3 performed worse than IT2 for identifying the company name (0.09 decrease in accuracy). This could originate from certain postprocessing steps in the pipeline optimized for the dataset used in IT1 and IT2. Moreover, potential errors in our validation dataset from iteration 3, which is manually extracted and compared, could also influence our evaluation here (Table 1).

In the third and final iteration, the list of variables was expanded with the CAS numbers and the list of ingredients of the product. The LLM outputs for IT3 showed accuracy scores of 0.87 and 0.90 for CAS numbers and ingredients list, respectively.

Table 1. Overview of pipeline performance expressed in accuracy scores for the 3 iterations for each of the variables of interest that were extracted from the safety data sheets.

Iteration	Variable	Nr. SDS	Good match	No match	Accuracy
IT1	Product name	470	262	208	0.56
	Company name		287	183	0.61
	Signal words		178	292	0.38
	Document version		284	186	0.60
	Release date		55	415	0.12
	H-phrases		30	440	0.06
Iteration	Variable	Nr. SDS	Good match	No match	Accuracy
IT2	Product name	470	356	114	0.76
	Company name		421	49	0.90
	Signal words		413	57	0.88
	Document version		363	107	0.77
	Release date		254	216	0.54
	H-phrases		333	137	0.71
	Physical status		321	149	0.68
Iteration	Variable	Nr. SDS	Good match	No match	Accuracy
IT3	Product name	462	386	76	0.84
	Company name		374	101	0.81
	Signal words		435	27	0.94
	Document version		376	86	0.83
	Release date		341	121	0.74
	H-phrases		377	85	0.89
	Physical status		367	95	0.80
	CAS numbers		404	59	0.87
	Ingredients product		416	47	0.90

Further improvements are expected to be achievable by changing the way the context is imported for the LLM. The current method for importing PDFs introduces errors, since it reads the PDF file line by line, while some information (eg release date) is stored vertically. A promising approach involves using the image-to-text capabilities of the GPT4 models or the UniTable framework (Peng et al. 2024; OpenAI 2024b), which extracts information directly from a table without first forcing it into string format.

Finally, fine-tuning is another method to improve LLM performance, whereby an LLM is trained on a topic-specific dataset. However, fine-tuning requires an elaborate dataset of input–output pairs that resemble the expected context and desired response (Raffel et al. 2020; Vatsal and Dubey 2024), which can be time-consuming to construct. This raises the question of what level of accuracy for LLM-driven data handling would be acceptable when compared to the current manual methods. One can imagine when working on highly sensitive topics (eg LLM-derived cancer prognosis (Sun et al. 2024)), the margin of accuracy should be higher than the average 0.83 reported here, making fine-tuning worth the effort. Similarly, achieving more accurate outputs for the SDSs is necessary for certain key hazard variables, such as H-phrases and CAS numbers, while a small deviation in the company name could be more acceptable. A recent study found that manually SDS indexing yielded an error rate between 5 and 10%, depending on the variable (Khan et al. 2025). This could set a target for LLM approaches to match or surpass manual extraction methods in accuracy.

Within the occupational health and safety domain, many other documents are used for which LLM methods could enhance data extraction. Given that SDSs follow a (semi-)structured format, it would be interesting to expand this work to more unstructured text formats such as exposure or incident reports (Smetana et al. 2024). Summarizing large amounts of text requires a more nuanced understanding of the context by the LLM and is an ongoing field of development (Pu et al. 2023; Chang et al. 2024; Mullick et al. 2024).

Conclusion

Overall, this pipeline functions as a support tool and shows to be well-suited for automatic extraction of less critical variables, such as company and product name, for which small deviations from the original SDS variables are tolerable. The results demonstrate that significant improvements can be achieved with relatively minor adjustments in the postprocessing and future LLM development, though perfect accuracy will remain challenging.

The rapid advancements in the field of LLM applications have the potential to significantly fasten and improve data management of documents (among which SDSs) that support efficient risk assessments, thereby leading to a safer workplace when working with hazardous materials.

Supplementary material

Supplementary material is available at *Annals of Work Exposures and Health* online.

Funding

This research was funded by the Dutch Ministry of Social Affairs and Employment.

Conflict of interest

The authors declare no conflict of interest in this study.

Data availability

Model output is available upon request. Prompts used for text extraction can be found in the Supplementary material.

References

Achiam J et al. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774v6. https://doi.org/10.48550/arXiv.2303.08774.

Araci D. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. https://doi.org/10.48550/arXiv.1908.10063. Accessed 25 November 2024.

Bartoli A, De Lorenzo A, Medvet E, Tarlao F. 2016. Inference of regular expressions for text extraction from examples. IEEE Trans Knowl Data Eng. 28:1217–1230. https://doi.org/10.1109/TKDE.2016.2515587.

Brown TB et al. 2020. Language models are few-shot learners. arXiv preprint arXiv: 2005.14165v4. https://doi.org/10.48550/arXiv.2005.14165.

Chang Y, Lo K, Goyal T, Lyyer M. 2024. Booookscore: A systematic exploration of book-length summarization in the era of Llms. In: 12th International Conference on Learning Representations (Vienna). *preprint* arXiv: 2310.00785v4. https://doi.org/10.48550/arXiv.2310.00785.

Chen Banghao, Zhang Zhaofeng, Langrené Nicolas, Zhu Shengxin. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. arXiv: 2314.14735v5. Available from: https://doi.org/10.48550/arXiv.2310.14735.

Demasi A, Elston H, Langerman N. 2022. Safety data sheets: challenges for authors, expectations for End-users. ACS Chem Health Saf. 29:369–377. https://doi.org/10.1021/acs.chas.2c00015.

Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). p. 4171-4186. *print* arXiv: 1810.04805v2. https://doi.org/10.48550/arXiv.1810.04805.

European Comission. 2006. Regulation (EC) 1907/2006 of the European Parliament and of the Council of 18 December 2006—REACH. http://eurlex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006R1907&from=en. Accessed 15 October 2024.

Javaid M, Haleem A, Singh RP. 2023. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Eval. 3:100105. https://doi.org/10.1016/j.tbench.2023.100105

Khan M, Penfield J, Suman A, Crowell S. 2025. A machine learning driven automated system to extract multiple information fields from safety data sheet documents. Heliyon. 11:e42215. https://doi.org/10.1016/J.HELIYON.2025.E42215.

Lee J et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 36:1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

Li Y et al. 2021c. Structext: structured text understanding with multi-modal transformers. Association for Computing Machinery.

Li C et al. 2021a. StructuralLM: Structural Pre-training for Form Understanding. Association for Computational Linguistics. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers):6309–6318. https://aclanthology.org/2021.acl-long.493/.

Li P et al. 2021b. SelfDoc: Self-Supervised Document Representation Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Virtual / Nashville, TN):5652–5660. arXiv preprint arXiv: 2106.03331v1.

Mullick A et al. 2024. Leveraging the power of LLMs: a fine-tuning approach for high-quality aspect-based summarization. *arXiv preprint arXiv*: 2405.02584v1. https://doi.org/10.48550/arXiv.2408.02584.

NIOSH, National Institute for Occupational Safety and Health. 2009. Qualitative risk characterization and management of occupational hazards: control banding (CB). Publication No. 2009-152.

Nayar GA et al. 2016. The efficacy of safety data sheets in informing risk based decision making: a review of the aerospace sector. J Chem Health Saf. 23:19–29. https://doi.org/10.1016/j.jchas.2015.09.002.

OpenAI. 2024a. Fine-tuning—OpenAI API. https://platform.openai.com/docs/guides/fine-tuning. Accessed date 5 December 2024.

OpenAI. 2024b. Hello GPT-4o | OpenAI. https://openai.com/index/hello-gpt-4o/. Accessed December 2024.

Otten W et al. 2022. Optimizing the benefit of REACH worker exposure assessments: ensuring meaningful health risk communication—TNO2022 R10516. TNO.

Peng S et al. 2024. UniTable: Towards a unified framework for table structure recognition via self-supervised pretraining. *arXiv preprint arXiv*: 2403.04822v2. https://doi.org/10.48550/arXiv.2403.04822.

Pu X, Gao M, Wan X. 2023. Summarization is (almost) dead. *arXiv preprint arXiv*:2309.09558v1. https://doi.org/10.48550/arXiv.2309.09558.

PyMuPDF. 2024. PyMuPDF 1.24.14. URL: https://pymupdf.readthe docs.io/en/latest/the-basics.html. Accessed: 2 October 2024.

Raffel et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 21(140):1–67. https://jmlr.org/papers/v21/20-074.html.

Raiaan MAK et al. 2024. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access. 12:26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742.

Smetana M, Salles LSD, Sukharev I. 2024. Highway construction safety analysis using large language models. Appl Sci. 14:1352. doi:https://doi.org/10.3390/app14041352.

Standard HC, Sheets SD. 2006. Hazard communication. Safety data sheets. 1200:35–40. https://doi.org/10.18356/a01cf234-en.

Suman A et al. 2024. A machine learning driven automated system for safety data sheet indexing. Sci Rep. 14:4415. https://doi.org/10.1038/s41598-024-55231-1.

Sun D et al. 2024. Outcome prediction using multi-modal information: integrating large language model-extracted clinical information and image analysis. Cancers (Basel). 16:2402. https://doi.org/10.3390/cancers16132402.

Touvron H, et al. 2023. LLaMA: open and efficient foundation language models. *arXiv preprint arXiv*:2302.13971v1. http://arxiv.org/abs/2302.13971.

Usman Hadi M, et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv preprint, TechRxiv*. https://doi.org/10.36227/techrxiv.23589741.v7.

Vatsal S, Dubey H. 2024. A survey of prompt engineering methods in large language models for different NLP tasks. *arXiv preprint arXiv*:2407.12994v2. https://doi.org/10.48550/arXiv.2407.12994.