Human Verification of LLM-Powered Structured Data Extraction from Image Files

Katherine Thornton[1], Kenneth Seals-Nutt[2], Mika Matsuzaki[3] and Marcel Nguemaha[4]

[1]WikiFCD Collaborative, Olympia, WA, USA
[2]WikiFCD Collaborative, New York, New York, USA
[3]Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, United States
[3]Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, United States

Abstract

Many food composition tables are published as PDF files. Data in these files are relevant to researchers looking into the change of food composition values over time, regional differences in values and other questions. While there are software tools to extract structured data from more recent versions of PDF, these tools do not support the earliest versions of PDF. We tested an LLM-based workflow for creating CSV files of structured data extracted from legacy PDF files which we converted to PNG files. We manually reviewed each value reported by the LLM to determine the suitability of this approach. We found multiple inaccurate values in the dataset extracted by the LLM. While this approach is insufficient as a stand-alone method, we discuss potential for human-in-the-loop workflows to leverage the power of LLMs to assist with data extraction from legacy versions of PDF files.

Keywords

Food Composition, Nutri-informatics, Wikibase, Wikidata, artificial intelligence, large language models

Introduction

Researchers interested in the nutritional composition of foods commonly eaten in Cameroon may need to consult the food composition tables published in 1957 and 1966. Each of these tables was published in print first, and also are available digitally in the Portable Document Format (PDF). As members of the WikiFCD community, we aim to make food composition data available in our knowledge base[1]. After adding data from a food composition table (FCT) to WikiFCD they can be combined with other data on the web more easily. Querying WikiFCD allows users to ask questions about food data across multiple food composition tables at once.

We value food composition tables that were published decades ago because they provide data useful for comparison with food composition values measured more recently. In some cases these older FCTs may be the only data source for a particular region of the world. While these older FCTs are important sources for WikiFCD, they can be challenging to work with in our software pipelines because of the fact that PDF is not a machine-readable file format.

1. Related Work

People who publish documents on the web often use the Portable Document File (PDF) [1]. Variations in the PDF format lead to differences in how data can be extracted from files [2]. Differences between versions of PDF and character sets used to encode text, among other issues present challenges for extracting data from these files [3].

Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025)

*Corresponding author.
†These authors contributed equally.
katherine.thornton@yale.edu (K. Thornton); kenneth@seals-nutt.com (K. Seals-Nutt); mmatsuz2@jhu.edu (M. Matsuzaki); mnguema1@jh.edu (M. Nguemaha)
0000-0002-4499-0451 (K. Thornton); 0000-0002-5926-9245 (K. Seals-Nutt); 0000-0002-7020-3757 (M. Matsuzaki)

[1]https://wikifcd.wikibase.cloud/wiki/Main_Page

People are exploring the use of LLMs to automate tasks [4]. Some have used LLMs for data extraction [5]. This technique works when the LLM can understand the format of the original file. We found that GPT-4-Turbo could not extract tabular data from the our PDF file. Researchers have demonstrated that a drawback of using LLMs is that they are known to provide plausible but incorrect responses, sometimes termed "hallucination" [6, 7, 8]. Our concern about the risk of hallucination of data values through this data extraction process motivated our decision to manually review each value the LLM reported.

2. WikiFCD

We created WikiFCD to offer web-based access to food composition tables from around the world[9]. WikiFCD is a knowledge base of food items and food composition data free for anyone to reuse. In order to make data from WikiFCD easier to reuse, we also map food items to FoodOn [10, 11]. We maintain mappings to food items in Wikidata which allows us to integrate our food composition data with the data in Wikidata as well as with datasets that also map their identifiers to Wikidata. This type of integration across databases opens many pathways for investigating questions of how nutritional intake interacts with topics related to human health [12].

3. WikiFCD Data Import

When FCTs are published using the CSV format, which is machine-readable, we are able to provide them to our software pipeline which is enabled by Wikidata Integrator[2]. Many researchers and practitioners use WikidataIntegrator to work with data pipelines for Wikidata and for other Wikibases [13].

When FCTs are published using the PDF format, which is not machine-readable, we need to extract the data and convert it to CSV. The PDF family of file formats has a long history and the developers of this file format have improved it in different ways between each version of the format. Software developers have created a wide variety of tools for working with different versions of PDF, with the number of available tools increasing for more recent versions of the format.

4. LLM-Powered Data Extraction

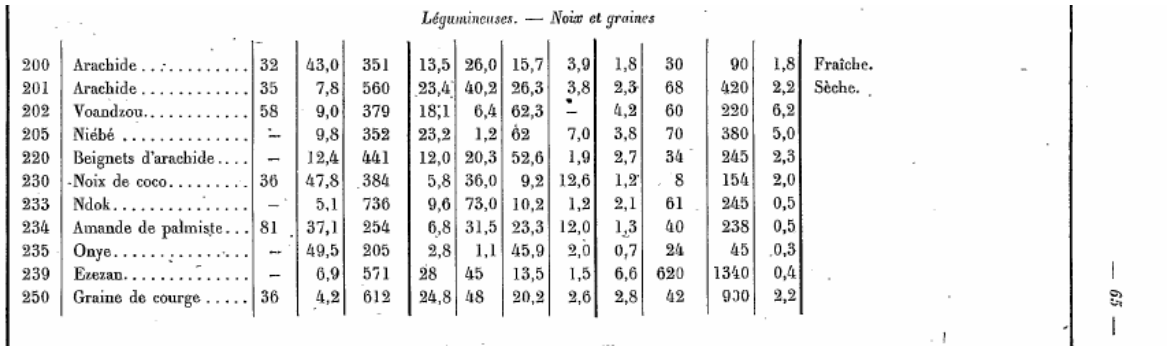
The PDF file for the 1957 Cameroon FCT was originally created on Saturday, February 10, 2001. According to the metadata for the file, the version of this PDF is 1.3. This PDF file contains a textual introduction as well as several pages of food composition values for food items presented in a table.

To extract food composition data from the scanned PDF, we initially tested several Python libraries including pyPDF[3], pdfplumber[4], and pdf2image+OCR[5]. However, we found that these libraries were unable to extract the data successfully due to the structure and quality of the source file. The PDF contained scanned images with tabular data, which lacked embedded text layers and often had faint or distorted lines, making traditional parsing unreliable for our specific PDF file.

We subsequently decided to try GPT-4-Turbo with vision capabilities via the OpenAI API to extract the tabular data. Because the food composition tables spanned only a few pages, we opted to manually capture screenshots of relevant sections and save them in the Portable Networked Graphics (PNG) format. This approach allowed for precise targeting of the visual content while minimizing noise from surrounding text or headers. This also simplified the prompting process by allowing direct alignment between the image content and the extraction instructions.

We used a simple Python workflow in which the images were read as binary objects and encoded in base64 before being passed to GPT-4-Turbo via the beta.chat.completions.parse method. We used a structured prompt to instruct the LLM to extract the tabular content of the image. We also inculed

[2]https://pypi.org/project/wikidataintegrator/
[3]https://pypi.org/project/pypdf/
[4]https://pypi.org/project/pdfplumber/
[5]https://pypi.org/project/pdf2image/



Légumineuses. — Noix et graines

200	Arachide	32	43,0	351	13,5	26,0	15,7	3,9	1,8	30	90	1,8	Fraîche.
201	Arachide	35	7,8	560	23,4	40,2	26,3	3,8	2,3	68	420	2,2	Sèche.
202	Voandzou	58	9,0	379	18,1	6,4	62,3	–	4,2	60	220	6,2	
205	Niébé	–	9,8	352	23,2	1,2	62	7,0	3,8	70	380	5,0	
220	Beignets d'arachide	–	12,4	441	12,0	20,3	52,6	1,9	2,7	34	245	2,3	
230	Noix de coco	36	47,8	384	5,8	36,0	9,2	12,6	1,2	8	154	2,0	
233	Ndok	–	5,1	736	9,6	73,0	10,2	1,2	2,1	61	245	0,5	
234	Amande de palmiste	81	37,1	254	6,8	31,5	23,3	12,0	1,3	40	238	0,5	
235	Onye	–	49,5	205	2,8	1,1	45,9	2,0	0,7	24	45	0,3	
239	Ezezan	–	6,9	571	28	45	13,5	1,5	6,6	620	1340	0,4	
250	Graine de courge	36	4,2	612	24,8	48	20,2	2,6	2,8	42	900	2,2	

Figure 1: Detail of the PDF file for the FCT for Cameroon showing the use of '-' to indicate no value for some components.

specific instructions for how to handle missing data, represented by the '-' symbol. We instructed the model to parse the output into a Pydantic[6] model (TableData) and convert it into JSON. We then converted the structured JSON to a CSV file.

In Figure 1, we see a section of the original PDF of the FCT, in which the food items labeled in French. To improve usability of this data, we instructed the language model to translate the food item labels from French to English so that we could provide English language labels for these food items in the WikiFCD system. The LLM generally produced accurate translations; however, in one instance it returned a different French label than what appeared in the original PDF file. As a result, we implemented manual verification to ensure consistency and accuracy of the final output data.

5. Human Review of Data Quality

To validate the method, we performed a manual review of the results generated by the LLM and compared the values with the numbers in the original food composition table. In total we found eighty five food items in the Cameroon FCT. For each food item there are eleven nutrients described. Of the nine hundred thirty five food composition values, the LLM provided thirteen incorrect values. The LLM also provided two inaccurate English labels.

We found that this LLM-powered approach generally worked well, except when the values in the PDF were illegible due to the low quality of the scan that cut off some numerals. We observed that for food items where the LLM only reported one or two incorrect component values, the incorrect values seemed to be duplicated from a nearby value. For example, for the food items labeled 'Courge' seen in 2, the value the LLM returned for 'Formic Insoluble matter' was incorrect. The LLM returned a value of '0.6', which may have been duplicated from the 'Ash' value of '0.6'.

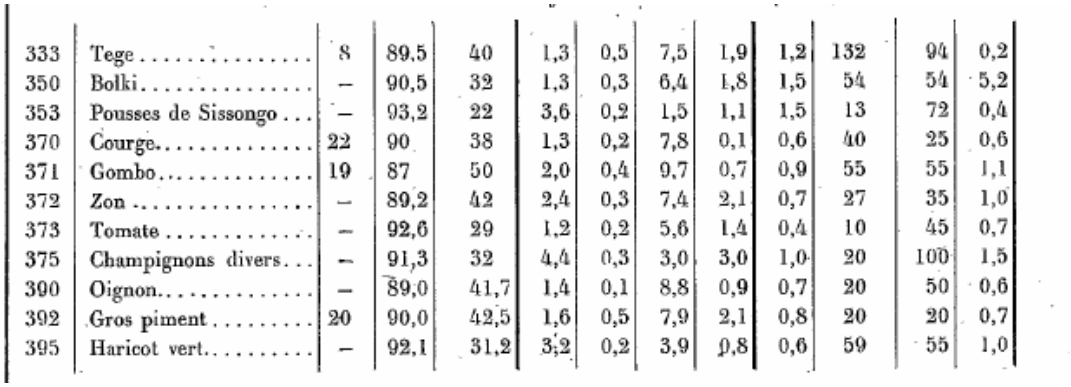
In addition to numerals, the LLM accurately reported the dash character '-' used in the original file to indicate 'no value' as seen in 1. However, there were few instances where the '-' was reported as '0' (see the food item labeled 'Canne à sucre').

When the person who created the digital scan of the print version of the FCT was holding open the pages some of the values in the table were cut off and didn't come out clearly. In Figure 3 the values for 'Avocat' are cut off, but still legible. The values for 'Barbadine' are not legible. Unsurprisingly, the LLM returned multiple incorrect values for 'Barbadine'.

6. Discussion

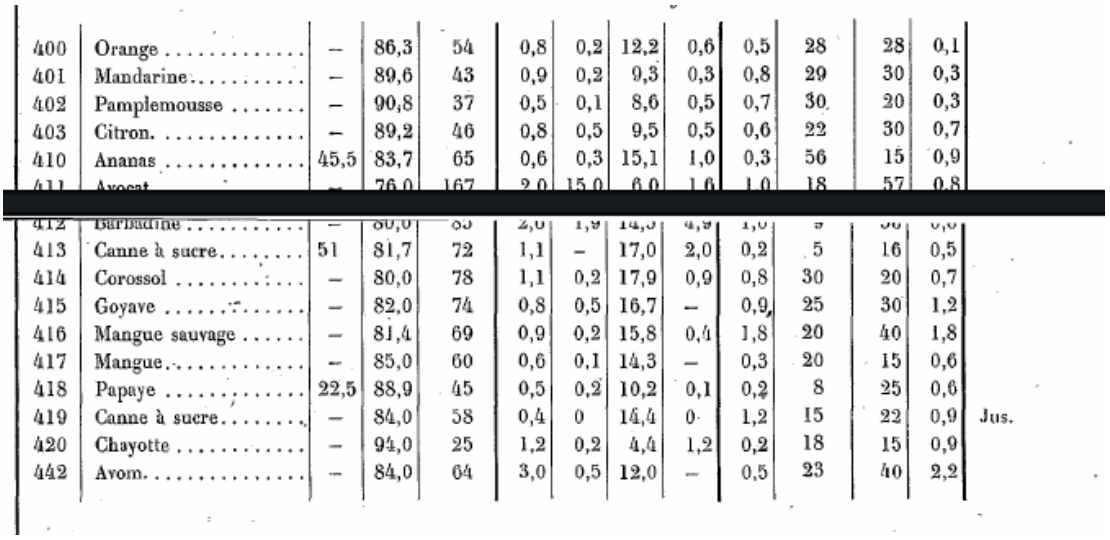
We tested this LLM-based data extraction approach to determine if this could be a viable strategy for working with data in legacy versions of PDF. The LLM performed well translating the labels of the food items and the column names into English. While the LLM reported accurate values for most nutrients

[6] https://docs.pydantic.dev/latest/



333	Tege	8	89,5	40	1,3	0,5	7,5	1,9	1,2	132	94	0,2
350	Bolki	–	90,5	32	1,3	0,3	6,4	1,8	1,5	54	54	5,2
353	Pousses de Sissongo	–	93,2	22	3,6	0,2	1,5	1,1	1,5	13	72	0,4
370	Courge	22	90	38	1,3	0,2	7,8	0,1	0,6	40	25	0,6
371	Gombo	19	87	50	2,0	0,4	9,7	0,7	0,9	55	55	1,1
372	Zon	–	89,2	42	2,4	0,3	7,4	2,1	0,7	27	35	1,0
373	Tomate	–	92,6	29	1,2	0,2	5,6	1,4	0,4	10	45	0,7
375	Champignons divers	–	91,3	32	4,4	0,3	3,0	3,0	1,0	20	100	1,5
390	Oignon	–	89,0	41,7	1,4	0,1	8,8	0,9	0,7	20	50	0,6
392	Gros piment	20	90,0	42,5	1,6	0,5	7,9	2,1	0,8	20	20	0,7
395	Haricot vert	–	92,1	31,2	3,2	0,2	3,9	0,8	0,6	59	55	1,0

Figure 2: Detail of the PDF file for the FCT for Cameroon showing the 'Courge' entry.



400	Orange	–	86,3	54	0,8	0,2	12,2	0,6	0,5	28	28	0,1
401	Mandarine	–	89,6	43	0,9	0,2	9,3	0,3	0,8	29	30	0,3
402	Pamplemousse	–	90,8	37	0,5	0,1	8,6	0,5	0,7	30	20	0,3
403	Citron	–	89,2	46	0,8	0,5	9,5	0,5	0,6	22	30	0,7
410	Ananas	45,5	83,7	65	0,6	0,3	15,1	1,0	0,3	56	15	0,9
411	Avocat	–	76,0	167	2,0	15,0	6,0	1,6	1,0	18	57	0,8
412	Barbadine	–	80,0	55	2,0	1,9	14,3	4,9	1,0	9	30	0,6
413	Canne à sucre	51	81,7	72	1,1	–	17,0	2,0	0,2	5	16	0,5
414	Corossol	–	80,0	78	1,1	0,2	17,9	0,9	0,8	30	20	0,7
415	Goyave	–	82,0	74	0,8	0,5	16,7	–	0,9	25	30	1,2
416	Mangue sauvage	–	81,4	69	0,9	0,2	15,8	0,4	1,8	20	40	1,8
417	Mangue	–	85,0	60	0,6	0,1	14,3	–	0,3	20	15	0,6
418	Papaye	22,5	88,9	45	0,5	0,2	10,2	0,1	0,2	8	25	0,6
419	Canne à sucre	–	84,0	58	0,4	0	14,4	0	1,2	15	22	0,9
420	Chayotte	–	94,0	25	1,2	0,2	4,4	1,2	0,2	18	15	0,9
442	Avom	–	84,0	64	3,0	0,5	12,0	–	0,5	23	40	2,2

Jus.

Figure 3: Detail of the PDF file showing illegible values due to scan.

for most food items, but also reported multiple inaccurate values. The number of inaccurate values the LLM reported indicates that this strategy would need to be followed by manual review of the data.

Manual review of a small dataset like that of the Cameroon FCT is time-intensive and requires precision. Comparing the time it would take a human to manually generate a spreadsheet of values from consulting the PDF file to the time it takes to manually review the data would be an interesting experiment. It is possible that it would be faster to manually create a CSV. We are aware of concerns regarding the energy consumption associated with LLM integration into software systems [14]. When we consider the resources required to generate a set of values that require manual verification it becomes more difficult to justify this approach.

We found the data in rows one through forty two to be accurate. The accuracy of the LLM declined after row forty three. In future work we would like to test additional strategies to improve the LLM's performance. One strategy we could employ would be to create one PNG image for each page within the PDF to test if providing smaller sections of the dataset to the LLM would improve performance. We could also test if structuring the prompt to work row-by-row could improve performance.

7. Conclusion

Identifying tools that can extract data from legacy versions of PDF is challenging. Due to the fact that some of the food composition tables we would like to import into WikiFCD are in legacy versions of PDF, we have an interest in automating data extraction from these files. We tested a workflow in which

we converted a PDF file to a PNG and asked GPT-4-Turbo to extract the food composition data from the image. We manually reviewed each value reported by GPT-4-Turbo for accuracy by comparison with the values in the PDF file. While we did find that the LLM reported some inaccurate values, we believe this approach is still worth consideration. With the rapid rate of improvements in performance of LLMs it is possible that their capacity for this type of data extraction could improve in the future.

Acknowledgments

We thank the Joint Food Ontology Working Group for productive discussions about FoodOn and data related to food. We thank the Wikidata community for continuing to improve the Wikidata knowledge base.

Declaration on Generative AI

Or (by using the activity taxonomy in ceur-ws.org/genai-tax.html):
We used OpenAI GPT-4-Turbo to extract text from a PDF file as the basis for the process that we then manually reviewed. We also asked OpenAI GPT-4-Turbo to translate the labels for the food items from French into English. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Allegrezza, Addressing The Problem of File Formats Obsolescence: Italian Guidelines on File Format Conversion for the Long-Term Preservation of Electronic Records, in: iPRES 2024 Papers - International Conference on Digital Preservation, 2024. Https://ipres2024.pubpub.org/pub-/oswmkgvc.

[2] A. S. Corrêa, P.-O. Zander, Unleashing tabular content to open data: A survey on pdf table extraction methods and tools, in: Proceedings of the 18th Annual International Conference on Digital Government Research, dg.o '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 54–63. URL: https://doi.org/10.1145/3085228.3085278. doi:10.1145/3085228.3085278.

[3] J. M. Ockerbloom, Archiving and preserving pdf files, RLG DigiNews 5 (2001) 1–6.

[4] Y. Shen, K. Song, X. Tan, W. Zhang, K. Ren, S. Yuan, W. Lu, D. Li, Y. Zhuang, Taskbench: Benchmarking large language models for task automation, Advances in Neural Information Processing Systems 37 (2024) 4540–4574.

[5] A. Konet, I. Thomas, G. Gartlehner, L. Kahwati, R. Hilscher, S. Kugley, K. Crotty, M. Viswanathan, R. Chew, Performance of two large language models for data extraction in evidence synthesis, Research synthesis methods 15 (2024) 818–824.

[6] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, J. Weston, Neural text generation with unlikelihood training, in: International Conference on Learning Representations, 2019.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. URL: https://doi.org/10.1145/3571730.

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. arXiv:2311.05232.

[9] K. Thornton, K. Seals-Nutt, M. Matsuzaki, Introducing wikifcd: Many food composition tables in a single knowledge base, in: CEUR Workshop Proceedings, volume 2969, CEUR-WS, 2021.

[10] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. Brinkman, W. W. Hsiao, Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration, npj Science of Food 2 (2018) 1–10.

[11] K. Thornton, K. Seals-Nutt, M. Matsuzaki, D. Damion, Reuse of the foodon ontology in a knowledge base of food composition data, Semantic Web Journal (2023).
[12] A. Farran-Codina, M. Urpí-Sardà, The power of databases in unraveling the nutrition–health connection, Nutrients 17 (2025). URL: https://www.mdpi.com/2072-6643/17/10/1725. doi:10.3390/nu17101725.
[13] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, B. M. Good, M. Griffith, O. L. Griffith, K. Hanspers, H. Hermjakob, T. S. Hudson, K. Hybiske, S. M. Keating, M. Manske, M. Mayers, D. Mietchen, E. Mitraka, A. R. Pico, T. Putman, A. Timothy, N. Queralt-Rosinach, L. M. Schriml, T. Shafee, D. Slenter, R. Stephan, K. Thornton, G. Tsueng, R. Tu, S. Ul-Hasan, E. Willighagen, C. Wu, A. I. Su, Wikidata as a knowledge graph for the life sciences, Elife 9 (2020) e52614. URL: https://doi.org/10.7554/ELIFE.52614.
[14] A. de Vries, The growing energy footprint of artificial intelligence, Joule 7 (2023) 2191–2194.