From Documents to Data: AI for Digitizing Local Land Use Plans (*Bebauungspläne*)

Laia Domenech Burin[1], Felicitas Sommer[2], Sebastian Riznik[3], Hope Ewudor[2]

Introduction – Scope

- Local land use plans provide valuable insights into environmental development and serve as key data sources for regional statistics.
- However, accessibility is limited due to their dispersion across various online platforms in non-standardized, unstructured PDF formats.
- Efforts to address this gap include the development of a regulatory information system for real estate, currently led by the Institute for Financial Services at HLSU Switzerland.
- We introduce an open-source data extraction pipeline designed to systematically extract and analyze data from German municipal land use plans (*Bebauungspläne*).

The main goal is to **create a dataset with structured information from local land use plans**, enabling the analysis of economic land utilization, sustainability requirements, and ESG-driven location decisions.

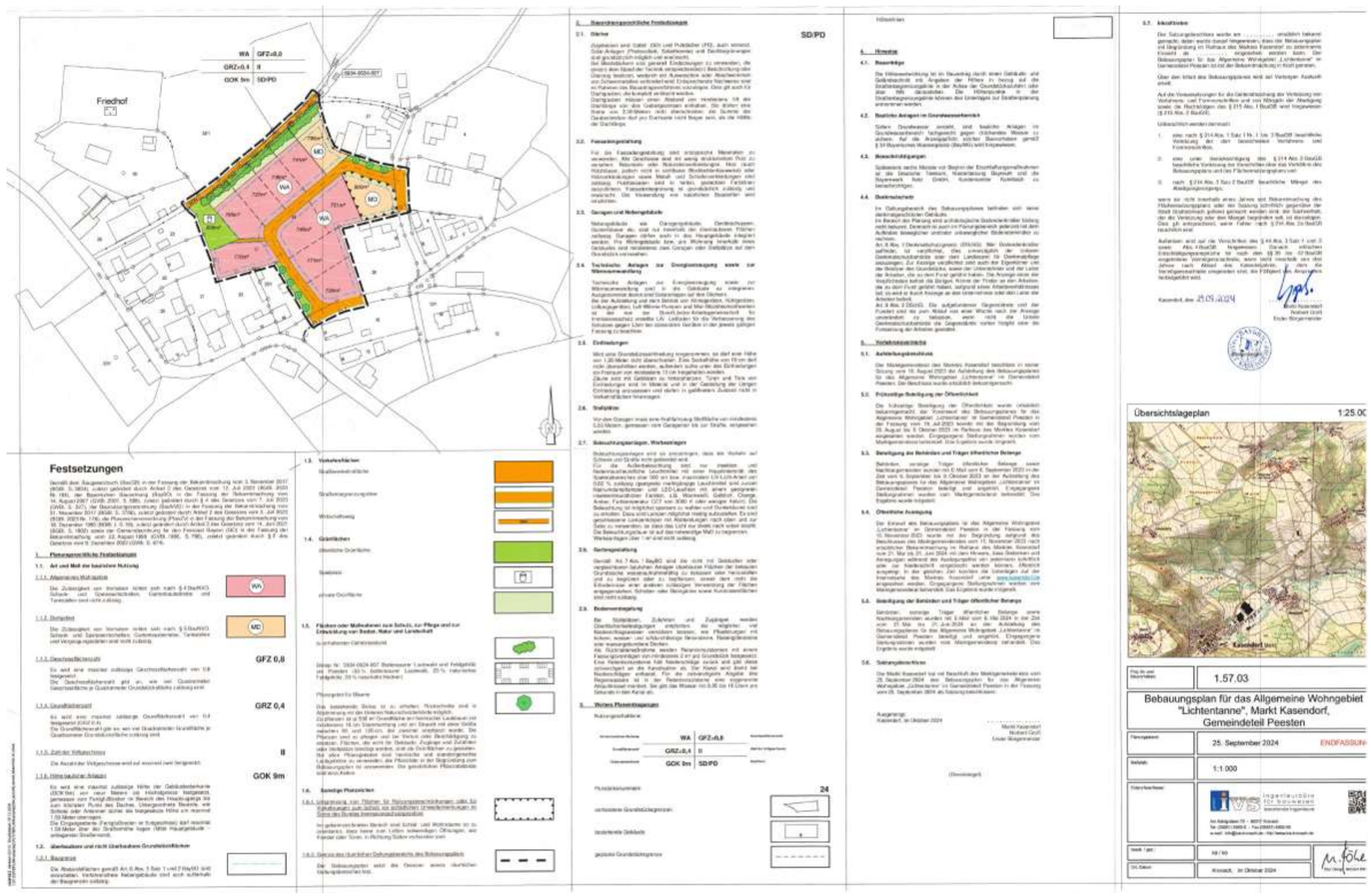


Figure 1. Example of a *Bebauungsplan*, which includes a map, Festsetzungen (stipulations), and Bauordnungsrechtliche Festsetzungen (building code regulations). Information and values may appear in any section, with some terms having multiple values.

Methodology

Building plan links were scraped from regional **Geoservices portals** like BayernAtlas.

PDF building plans were downloaded in bulk, processed, and parsed into **structured JSON**.

Large Language Model (LLM) prompts were applied to the JSON text to extract information and construct the database.

Sustainability and urban planning experts **defined key terms of interest.**

Relevant metrics were **manually annotated** by three expert annotators to serve as a benchmark for LLM performance.

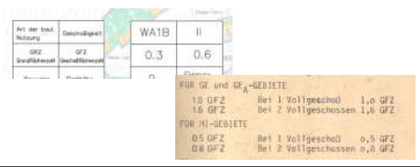
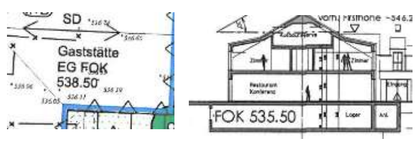
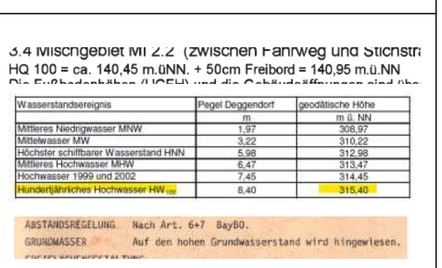
Term	Examples
GFZ (*Geschossflächenzahl*)	
GRZ (*Grundflächenzahl*)	
GOK (*Geländeoberkante*)	
EG FOK (*Erdgeschossfußbodenoberkante*)	
FOK (*Fußbodenoberkante*)	
HW100 (*Jahrhundertwasser*)	
HW10 (*10-Jährliches HW*)	
GW (*Grundwasser*)	

Table 1. Key Terms and Their Representation in Bebauungsplan Documents. The terms were grouped into three categories: land sealing (GFZ and GRZ), height (GOK, EG FOK and FOK), and flooding prevention (HW100, HW10 and Grundwasser)

Performance and Error Patterns

- The most common category across all terms was the correct absence of values in the document, indicating that the pipeline **effectively minimizes hallucinated values** (false positives).
- **Most extraction errors arised from missing values.**
- A smaller subset of errors occurred when multiple values were present in the building plan. In these cases, the model extracted only one value per metric, **leading to incomplete extractions.**
- *Grundflächzahl* and *Geschossflächenzahl* had the highest number of extractions and errors, reflecting their frequent occurrence in *Bebauungsplan* documents. However, for these terms **most documents had at least a partial or complete extraction.**

Learnings, discussion and future research

- *Bebauungsplan* documents show to be a **challenge in information extraction** even for the case of LLMs, likely due to their complex structure and the multiple ways information can appear
- The pipeline based solely on text extraction has shown **limited success** in fully capturing this information. However, it demonstrates promising results for utilizing LLMs to digitize unstructured documents, laying a foundation for further optimization.
- Future implementations to **enhance the pipeline** and the value of extraction:
  - Split document elements (map, text, legends) for **separate processing**.
  - Implement **vision models**
  - Enhance **prompt engineering** (e.g., few-shot learning).

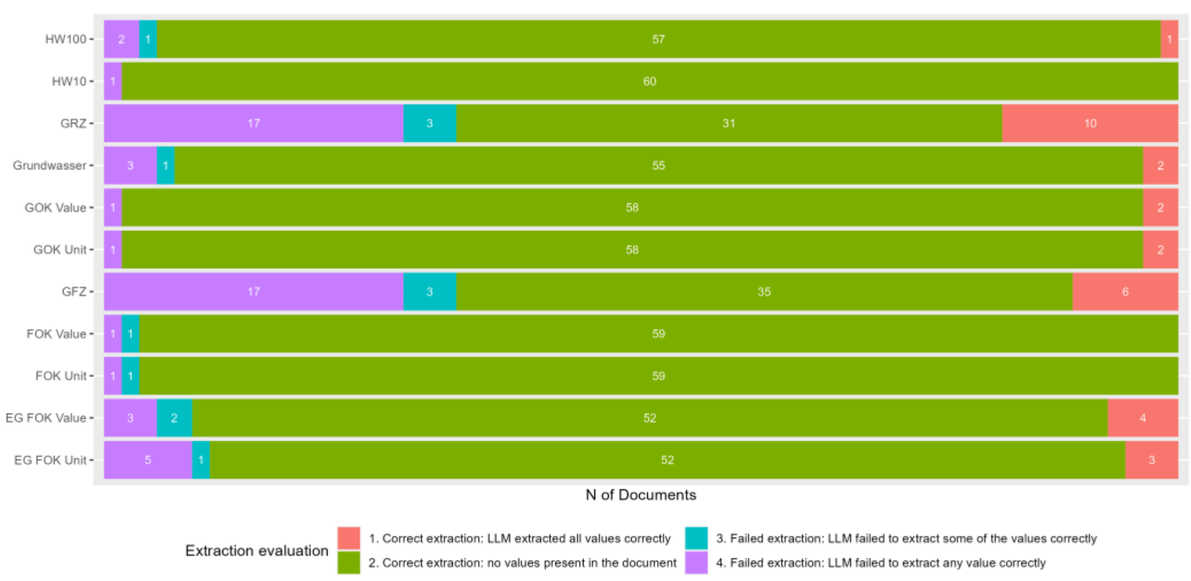


Fig. 2. Evaluation of extracted values across 61 documents, categorized into four levels: (1) fully correct extraction (pink), (2) correct absence of the term (green), (3) failed extraction (blue), and (4) partial extraction (purple). Each term could appear in multiple *Bebauungspläne*, with results referring to the total extractions per document.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

[1]Department of Statistics, LMU Munich, [2] Technische Universität München
[3] Bundeswehr University Munich

Corresponding Author:
Laia.DomenechBurin@stat.uni-muenchen.de

References
Hochschule Luzern – Informatik. (n.d.). Green Buildings: Datenbanken für nachhaltige Standortentscheide. Retrieved October 8, 2023, from https://www.hslu.ch/de-ch/hochschule-luzern/forschung/projekte/detail/?pid=6490
DSSGx Munich. (2023). Land sealing dataset and analysis for better understanding of climate change. Retrieved October 8, 2023, from https://www.dssgxmunich.org/dssgx-munich-2023/projects-fellows-2023