DEUTSCHE BUNDESBANK
EUROSYSTEM

Automating Company Data Validation with Multimodal Large Language Models: Integrating Text, Geospatial, and Image Data for Sector Classification

Technical Report 2026-02

Gabriela Alves Werb[1,2]
Patrick Felka[1]
Gabriel Thiem[1]
Susanne Walter[1]
Ece Yalcin-Roder[1]
Arda Yüksel[3]

[1] Deutsche Bundesbank, Research Data and Service Centre
[2] Frankfurt University of Applied Sciences, Business Information Systems
[3] TU Darmstadt, UKP

Disclaimer: The views expressed in this technical report are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Research Data and Service Centre

Abstract

The Bundesbank's statistics department provides high-quality company data by consolidating information from multiple, often inconsistent sources. Therefore, a significant challenge lies in resolving contradictions and imputing missing data in sectoral and regional metadata for large volumes of company data, a process that is partially automated but still heavily reliant on resource-intensive manual checks. This study explores the use of multimodal Large Language Models (MLLMs) to automate data validation by integrating unstructured text (such as company websites and commercial register entries) with geospatial and image data (e.g., satellite and overflight imagery, cartographic data, and 3D building models) related to company real estate. By proposing a system architecture for combining and processing geospatial data from images, texts, and digital twin data within a multimodal fact-checking approach, we aim to demonstrate the potential of multimodal fact-checking to enhance data quality and integrity. We conduct several experiments by alternating between different input sources, MLLMs used, and implementing multiple iterations to check for robustness of MLLM responses. The established pipeline can also be used for other cases such as missing data imputation or the extraction of company building features.

Acknowledgement: *We gratefully acknowledge the valuable insights provided by our colleagues in the Central Department for Reference and Foreign Trade Data Collection at the Bundesbank. Their expertise and experience in master data validation was a valuable contribution to the evaluation of our pipeline. Furthermore, this work would not have been possible without the active support of our dear colleagues Georg Steinbuß and Konstantin Körner in preparing the data. We are deeply grateful to them for their help. Furthermore, we thank Ivan Habernal for his invaluable conceptual input and for engaging in insightful discussions that greatly enriched our ideas. We also would like to express our sincere gratitude to Patrick Knöfel, Theresa Herbst and Tim Kaiser from the Federal Agency for Cartography and Geodesy for their significant contributions in facilitating the data exchange and their support in the Record Linkage. Their uncomplicated and friendly support made the process seamless and efficient and was a significant contribution to our project. Ultimately, we would like to thank Stefan Bender for reviewing the manuscript, his motivational words and for the helpful suggestions that significantly improved our paper.*

Usage of AI: *AI assistants were used in this work to assist with writing by correcting grammar and code, prompt optimization and debugging.*

Keywords: Data Quality; Multimodal Learning; Artificial Intelligence; Large Language Models; Natural Language Processing; Unstructured Data; Official Statistics; Geospatial Data; Aerial Images; Big Data; CityGML

Citation: Alves Werb, G., Felka, P., Thiem, G., Walter, S., Yalcin-Roder, E. & Yüksel, A. (2026). Automating Company Data Validation with Multimodal Large Language Models: Integrating Text, Geospatial, and Image Data for Sector Classification, Technical Report 2026-02. Deutsche Bundesbank, Research Data and Service Centre.

Contents

1 Introduction

The Statistics Department of the Bundesbank provides high-quality company data to analysts, researchers, government institutions, and the public. To achieve this, the Bundesbank collects and processes data from multiple sources for statistical purposes. A significant portion of this data is secondary, meaning it was not originally collected for the final statistical objective. Although the data is available in a structured format, it exhibits inconsistencies and disorganization, particularly regarding contextual metadata such as sectoral and regional information.

A critical step in data processing is the consolidation of information pertaining to the same entity from multiple sources, which often contain contradictory information (Christen, 2012). This process is partially automated through the application of majority voting (adopting the information confirmed by most sources) or prioritizing data sources (assigning the highest credibility to information from the most authoritative source). Nevertheless, a substantial proportion of contradictions require extensive manual verification, which is both resource-intensive and time-consuming (Batini and Scannapieco, 2016; Christen, 2012).

Given the extensive scope of the company master data, which may encompass millions of entities at the national level, the volume of data makes manual quality checks both costly and challenging to implement. This issue is relevant for any data source, public or private, that integrates information from multiple origins. Consequently, this challenge is common to all producers of high-quality data that inform policy decisions, such as environmental monitoring agencies, statistical bureaus, public health organizations, and economic research institutes (Redman, 1998).

The integration of multimodal large language models (MLLMs) with geospatial data, as well as semi-structured and unstructured textual data from companies, offers a sustainable long-term solution to replace manual checks (Y. Zheng, Lin, Chen, et al., 2023). MLLMs enable efficient processing of information from images in combination with textual company data (Li, Li, Xie, et al., 2023). Therefore, the primary objective of this study is to design a system architecture for a proof of concept demonstrating how MLLMs can be used to validate contextual company information using high-resolution satellite imagery, cartographic data, and digital twin data (i.e., 3D representations of the company buildings). The findings of this study are generalizable to other contexts that utilize geospatial data as input for MLLMs, particularly in applications involving spatial reasoning, where the integration of diverse data modalities is essential for informed decision-making and analysis. Given the multitude of attributes in the company master data, this study focuses on the fact-checking of those company attributes that contribute to the aggregated macroeconomic indicators published by the Bundesbank.

For statistics depicting the economic structure, the economic sector and the location of the companies are the two primary grouping variables for aggregation. The accuracy of the sectoral and regional classification is crucial for meaningful analyses and valid conclusions (Eurostat, 2013).

Analyses based on the economic sector or region illustrate the development of markets, the industrial structure, regional strength and regional disparities. An incorrect sector code may also indicate other incorrect company data (e.g., address, number of employees, etc.). Sector validation serves as our starting point for further validation of the company master data.

The validation of company information is a routine task in official statistics involving company data (UNECE, 2015). Therefore, the potential solutions developed in this study are also applicable to company master data in contexts beyond the Bundesbank.

2 Research Question

Considering the notable effort required to process and validate the data, there is a significant need for more efficient and universally applicable solutions.

Visual information about companies is currently ubiquitous and easy to access. For example, companies place images of their products, production facilities or headquarters buildings on their webpages and social media profiles. Furthermore, satellite images with high resolution are available from space agencies, such as NASA or ESA or from public authorities such as U.S. Geological Survey (USGS) or the European Environment Agency (EEA). It is possible to combine this visual information with classifications and descriptions of buildings (e.g., whether it is an industrial facility or an office building, volume, number of floors, surface area) from digital twin data or mapping platforms such as OpenStreetMap. These data offer a new possibility for gaining knowledge about a company's economic activity.

By combining data from different formats containing several representations of a company's activities, which might include text descriptions and images about its products, production facilities and office locations, we expect to learn valuable information to determine its economic sector more accurately. Therefore, the goal of this study is to assess the potential and feasibility of leveraging images and state-of-the-art multimodal learning methods by designing a system architecture to extract information about companies' activities and validate or impute their economic sectors using geospatial data.

From the perspective of NLP research, we can frame the problem as multimodal fact verification. However, unlike existing attempts to fact or claim verification in news, politics, or social media domains (Yao, Shah, Sun, Cho, and Huang, 2023), the nature of semi-structured textual data in this specialized business-related domain poses unique challenges. We will thus explore (1) to which extent can we employ transfer learning from text-based fact verification datasets (Nakamura, Levy, and Wang, 2019), (2) whether it is beneficial to combine images and text in large pretrained generative transformers like T5 and assess their adaptability to this problem (Pradeep, Ma, Nogueira, and Lin, 2020), and (3) explore the new task of generating explainable fact verification for multimodal data (Atanasova, 2024). Furthermore, as reproducibility of results is a major pillar for scientific analyses, we conduct robustness checks to analyse the variability in the prediction results of the MLLMs (Perkins and Roe, 2024). The findings of this study can potentially be applied to similar use cases in other public institutions, leveraging multimodal data such as satellite images and textual information to validate company details.

3 State of the Art

Past studies have relied on advanced methods to identify urban areas, buildings, forests, and other natural resources from satellite images (Sirko et al., 2021; Tam, Toan, Cong, and Hung, 2022; Yang, Qin, Grussenmeyer, and Koehl, 2018). Gebru et al. (Gebru et al., 2017) have used Google Street View to estimate the electoral behavior based on the demographic makeup of voters' neighborhoods. Popular methods include U-Net (Ronneberger, Fischer, and Brox, 2015) and related adaptations for semantic segmentation. However, many applications involve only a single data modality, such as text or images. More recent studies build on multiple modalities to improve the results of data analysis, enhance accuracy, and provide a more comprehensive understanding. Common applications include the combined usage of images and text, video, and audio, among others (Bernardi et al., 2016; Hu and Flaxman, 2018; Kiela, Bhooshan, Firooz, Perez, and Testuggine, 2019; Miller, Howard, Adams, Schwan, and Slater, 2020; Uzkent et al., 2019).

Multimodal fact verification has mostly been focused on social media or politics (Akhtar et al., 2023; Geng, Kementchedjhieva, Nakov, and Gurevych, 2024; Kotonya and Toni, 2020a; Yao et al., 2023; Zlatkova, Nakov, and Koychev, 2019). The recent COVID crisis has also brought attention to the verification of health-related claims (Kotonya and Toni, 2020b), however, the domain source remains news or fact-checking websites (Nakov et al., 2021). Very recent research on LLMs for fact-checking has expanded beyond fake news detection to include data validation, such as verifying large structured datasets (Theologitis, Dammu, Shah, and Suciu, 2026), scientific knowledge (Vladika and Matthes, 2023), and time-series data (Strong and Vlachos, 2025). Moreover, researchers have recently published studies that investigate the application of MLLM in conjunction with geospatial data (Mai et al., 2022, 2023; e.g., Roberts, Lüddecke, Sheikh, Han, and Albanie, 2024). These include vision papers that explore the promises and challenges of developing a foundation model for geospatial artificial intelligence (GeoAI) (Mai et al., 2022, 2023). But also, initial studies solve various specific geo-based tasks using MLLMs, such as predicting a country based on an image and text input (Roberts et al., 2024). Overall, research in this area is still in its infancy, but is attracting increasing interest in different fields, offering the potential for substantial improvements in processing geospatial data (J. Wu, Gan, Chao, and Philip, 2024). Regarding the quality of verification, other recent studies (Berk Atıl et al., 2025; Yuan et al., 2025) have also focused on analysing the consistency and reproducibility of MLLM responses across several complex tasks with constant configuration. Dependent on the task, they find a high variability in model responses, while logical deduction belongs to the task that shows the lowest variability (especially for GPT4o) (Berk Atıl et al., 2025).

4 Potential Solution and Challenges

Since the research question comprises manifold heterogeneous sub-questions, various solutions are conceivable. One relevant aspect of sector validation is related to companies in the manufacturing industry. Depending on the specialization of the company (whether small industry inputs are produced or large labor and capital-intensive goods), a company must maintain respective production facilities. The non-existence of a production facility (that represents the economic activity described by the company's economic sector classification) raises doubts about the correctness of the sector classification or other information (like company address, etc.). To bring together all the necessary information about a focal company it is necessary to combine the description of the economic sector of a company (e.g., from financial reports, commercial register entries or other sources), with geoinformation about the company's location, digital twin data, and satellite images (Goldblatt, Heilmann, and Vaizman, 2020; Kyriakos and Vavalis, 2023) (e.g., from Sentinel-2[1)] or Landsat[2)]).

Given the heterogeneity and volume of these data sources, the first step—extracting, linking and consolidating all relevant information from the mentioned sources— represents one of the greatest challenges (Fusco and Aversano, 2020; Koukaras, 2025), particularly as it involves not only identifying the exact location of a company but also determining the spatial extent of its premises. Furthermore, it is highly unlikely that big production companies with many employees share the same production facility. A production facility rarely consists of a single building; instead, it often comprises several facilities with distinct functions, each of which can provide valuable information about the economic activity of the company.

This is one of the main challenges that our architecture must address in order to contribute to our overarching goal and reduce the need for manual verification of an entity's master data.

Moreover, it is important to consider that not all companies provide the address of their production facility. Instead, they often list their administrative address or the address of their tax attorney, which may be located in a residential or commercial area and not necessarily near the production site. To reduce this uncertainty, we must consider the geolocation of the company's local unit (the establishment where employees are registered) and, if necessary, verify multiple addresses to accurately identify the economic sector (U.S. Census Bureau, 2026).

1 https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2
2 https://science.nasa.gov/mission/landsat/

5 Data Sources

The table below lists the data we use to evaluate the architecture presented in this study. In general, our architecture is capable of integrating satellite/overflight imagery with digital twin data stored in spatial formats such as CityGML (Kutzner, Chaturvedi, and Kolbe, 2020), along with semi-structured and unstructured text data extracted from descriptions of company activity from commercial registers, or similar sources. The table 1 below provides an overview of the data we use, where these data come from, and the multimodality of the data source.

Table 1: Data Sources and Format

Data Source	Data Name	Format
Deutsche Bundesbank	RIAD (Company Master Data to be validated)[a]	Table
Federal Agency for Cartography and Geodesy (GeoBasis-DE-BKG2024)	Building Function Data from digital twin (3D-Building models, LoD1/LoD2)[b]	CityGML (XML-based vector data)
Federal Agency for Cartography and Geodesy	Digital Orthophotos (DOP20), high resolution overflight images[c]	Geotiff (raster data)
OpenStreetMap	Nominatim, Larger polygons around the company property	Various formats available (e.g., shapefiles, XML)
Handelsregister Portal	Company Register Data[d]	Table (containing unstructured strings for the company activity description)
German statistical offices	Handbook for NACE/ WZ2008 sectors including descriptions[e]	pdf (unstructured text)

[a] https://www.verwaltungsdaten-informationsplattform.de/register/24#Allgemeines
[b] https://gdz.bkg.bund.de/index.php/default/3d-gebaudemodelle-lod1-deutschland-lod1-de.html
[c] https://gdz.bkg.bund.de/index.php/default/digitale-orthophotos-bodenauflosung-20-cm-dop20.html
[d] https://www.handelsregister.de/rp_web/welcome.xhtml
[e] https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-3100100089004-aktuell.pdf?__blob=publicationFile

The dataset to be quality-checked is the central company master data in the Bundesbank: The German part of Register of Institutions and Affiliates Database (RIAD). RIAD is an internal database that serves as the single source of truth for company master data. It is fed by various administrative and commercial data sources, each of which contains its own truth on the companies' economic activities and must be consolidated into one truth. This data is enriched with geospatial data, such as public satellite and overflight images, digital twin data containing building characteristics, map data from OpenStreetMap (OSM) (Haklay and Weber, 2008), as well as plain text extracted from unstructured textual resources such as descriptions of economic activity from commercial register entries.

To obtain instructions on how to interpret the extracted information and decide on an economic sector (WZ2008/ NACE) allocation, we also extract the guidelines from the official handbook of WZ2008.

For the proof of concept with a smaller sample, the training data is generated by manually verifying a random sample of integrated company data across different modalities (e.g., linked satellite images with company master data, digital twin data, OSM and unstructured texts from commercial register entries).

6 Model Architecture

To address the presented problem, we developed a potential system model architecture to encompass and link all data sources relevant to us. Drawing on recent advancements in this field (Wang, Zhuang, and Wu, 2024; S. Wu, Fei, Qu, Ji, and Chua, 2023), we present our architectural approach in Figure 1.

The processing of multimodal data within an LLM necessitates the preparation of data to align with the specifications of its input layer (Song et al., 2025). Based on the company master data to be analyzed, the company's address is first transferred to a Geo-Feature-Extraction Engine. This engine is designed to extract the required information from various datasets and integrate the geospatial data with the company-specific information.

Within the Geo-Feature Extraction Engine, a geocoder is first used to convert the company addresses provided into geographical coordinates (longitude and latitude). However, the output of the geocoder only describes a single point in a coordinate system, which could, for example, represent a building at this point. As the coordinates only represent a point, but we are interested in an area and the objects located on it - such as a company property and the facilities on it - the coordinates determined must first be compared with the property boundaries. Based on these property boundaries, which are represented by geographical polygons, data can be extracted from other sources and assigned to the respective address.

By integrating company data, such as company registration information, with geographic data sourced from various platforms, the input data is generated. To maximize the potential of MLLMs in this specific domain and task (Rostam, R., and Kertész, 2025), additional economic sector descriptions are also incorporated into the model.

The predictions generated by the MLLMs, utilizing prompt engineering (Liu et al., 2026; Schulhoff et al., 2024) and economic sector descriptions, will be compared against observed economic sector data and validated economic sector classifications. This comparative analysis aims to assess the accuracy and reliability of the model's predictions.

Figure 2 illustrates this procedure using an address as an example. First, the geo-coder is used to derive a point from the address (represented by the arrow in the image), which describes the actual administrative building of the company. In the next step, data is extracted from various sources based on the property boundaries associated with the extracted point. This includes data from the crowdsourced mapping platform OpenStreetMap, which mainly contains objects with meta information such as features and categories. This meta information is semi-structured and consists of both standardized tags and descriptive, unstructured continuous text. In addition, 3D objects are extracted from the digital twin data that are located on the property and describe facilities like buildings on it. Finally, the company premises are also extracted from the overflight images based on the property markings and prepared for the next steps.

The data obtained from the geo-feature extraction engine, which is presented in the form of text, vector data and images, is transferred to the next processing layers. The different modalities are first harmonized in the multimodal encoding layer. This layer is responsible for converting heterogeneous input formats into a standardized, model-compatible structure. After harmonization, the

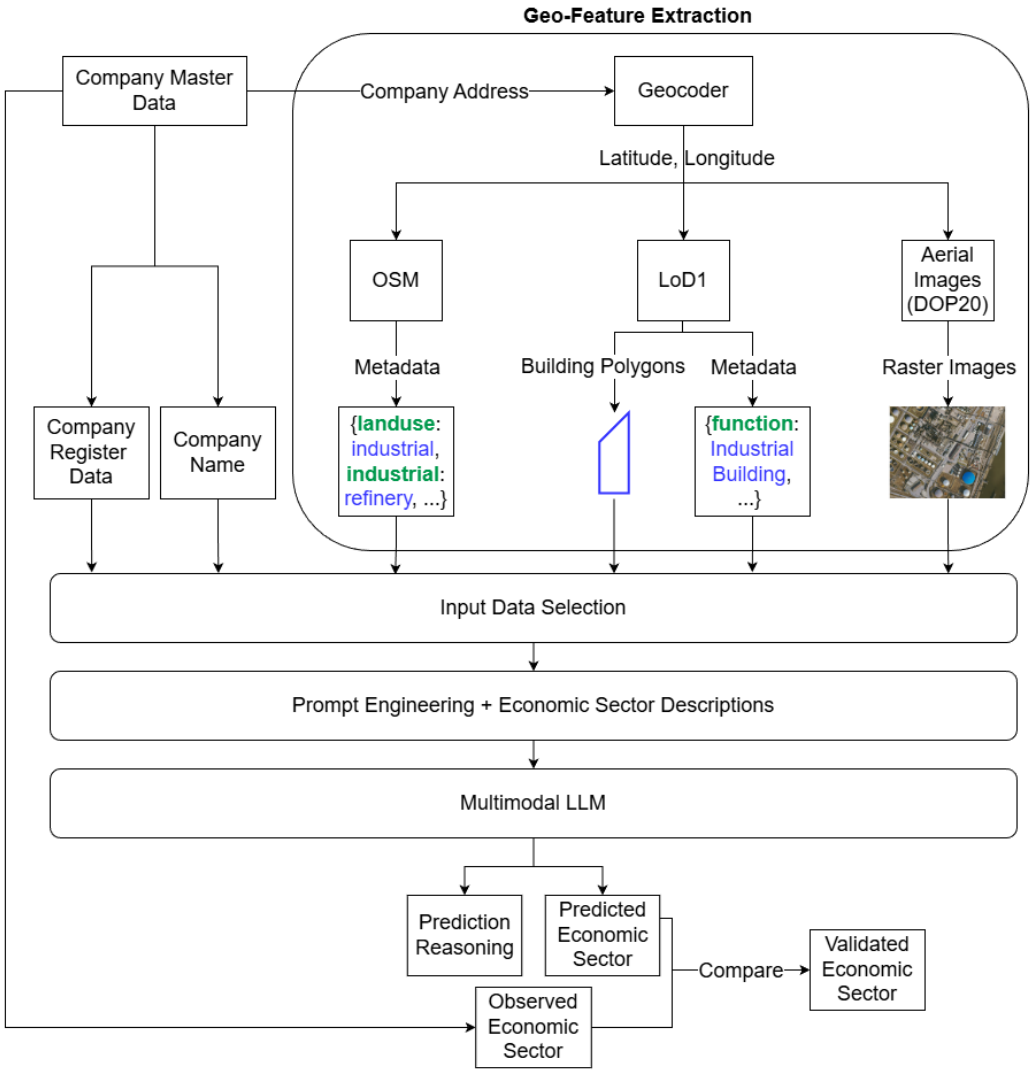


Figure 1: Proposed model architecture

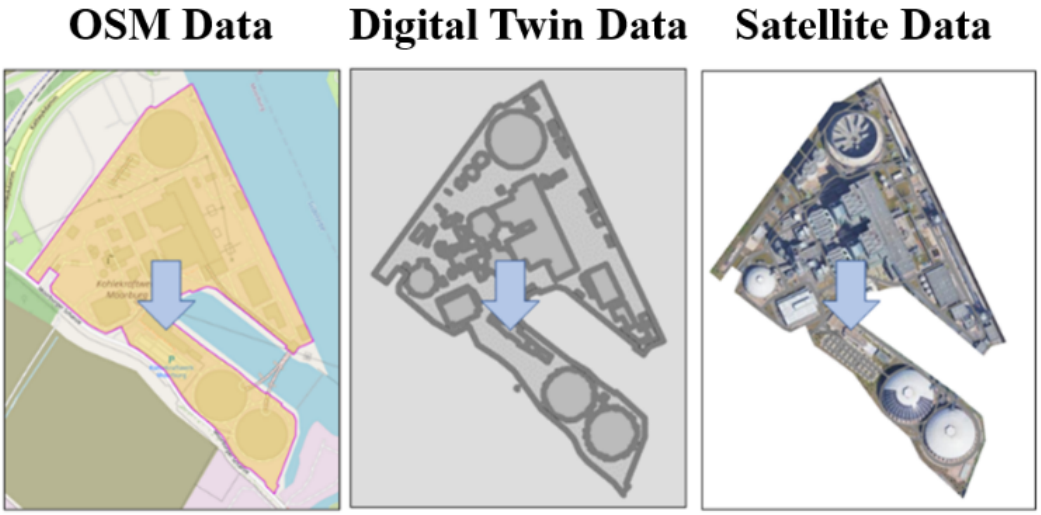


Figure 2: Extraction of geo-features based on the address

data is further transformed and processed in the input projector layer to optimally adapt it to the requirements of the underlying Large Language Model (Song et al., 2025).

By combining company data (e.g., economic sector classification and other metadata) with geographic information from various sources, a multimodal understanding of the context is established. This approach can be applied not only to our specific use case but also to other scenarios where information can be validated using geospatial data (S. Du et al., 2025; Skaug and Nojoumian, 2025).

7 Experimental Setup

Sample

As we aim to evaluate the potential of MLLMs to support human experts in data validation tasks, we need to find a configuration that offers the highest accuracy (compared to a human annotator) (Huang and Zhang, 2024; Wang, Kim, Rahman, Mitra, and Miao, 2024). Therefore, we run several experimental designs and compare the accuracy of all approaches. Specifically, this setup helps evaluate the effectiveness of various input configurations for classifying economic sectors ("Klassifikation der Wirtschaftszweige, Ausgabe 2008[3)]"– referred to as WZ) using an MLLM. Insights from our initial experiments highlighted two key findings: (1) WZ verification is challenging when relying solely on overflight images, indicating the necessity for additional information sources; and (2) incorporating WZ descriptions from the official handbook into the system prompt provides valuable context that improves model predictions. The experiment utilized internal company data from 2021, which included company names, and installation coordinates. Additionally, DOP20 aerial images from 2020 and LOD1 CityGML data containing polygons (with installation coordinates) and building function information were employed. Finally, company information taken from the Commercial Register was also used as an input signal (when available – 373 data points). In our setup, only the higher-level WZ groups (1-digit) are considered, each corresponding to a letter code [A-U]. The sample was restricted to data points with validated addresses to ensure data quality (so we could infer that any data mistake is related to a sector misclassification). A legal form filter was applied to remove personally identifiable information. The final sample consisted of a balanced set of 835 data points, with approximately 50 data points per available WZ code in the Master Data. We use this evenly distributed sample over WZ codes to understand the sector-specific performance of the model prediction.

Prompt Design and Model Interaction

To facilitate accurate predictions of economic activity, the experiment employed a structured prompt system consisting of a system prompt, a user prompt, and the subsequent MLLM response. The prompts were designed to provide the model with relevant context and instructions (Liu et al., 2026; Schulhoff et al., 2024), ensuring consistency and interpretability in the outputs.

■ **System Prompt**:

The system prompt establishes the model's role and provides essential background information (Liu et al., 2026), including the list of possible economic activities (WZ 2008) and their descriptions from the handbook. The prompt is as follows:

"You are an expert in analyzing company data who can precisely predict the economic activity of companies based on an overflight image. You have a list (WZ 2008) of economic activities in Germany to choose from. Here is the list of activities with their descriptions: <WZ 2008 list>. No economic activity outside of this list can be predicted. You are given an overflight image and

3 The WZ is the harmonized and standardized economic classification scheme for the German Official Statistics and administrative data (https://www.klassifikationsserver.de/klassService/thyme/variant/wz2008?lang=EN)

a company name so you can predict which economic activity from the list is MOST LIKELY to correspond to the image and name. Here is your task:"

■ **User Prompt**:

The user prompt provides the specific input for each prediction, including the company name and instructions regarding the interpretation of the image and polygon. The prompt is as follows:

"The company name is {company_name}. What economic activity do you think could correspond to this company? Consider the red polygon in the image to be a building where the company is located. Also consider the surroundings as clues if needed. Remember to only choose from the WZ 2008 list. If you are unsure, answer with your most likely prediction. ALWAYS start your answer with one of the possible upper case letter codes for the activities. Remember to not start with any special formatting so we can parse the first char. After that, ALWAYS provide your reasoning step by step."

■ **MLLM Response**:

The model's response consists of a prediction and a step-by-step explanation. An example response is provided below:

Prediction: E, Water supply; sewerage, waste management and remediation activities

Reasoning: *E The most likely economic activity corresponding to the company "company_name" is "E Water supply; sewerage, waste management and remediation activities." The reasoning is as follows:

1. **Company Name**: The name "company_name" directly translates to "Wastewater Association," which strongly suggests involvement in wastewater management or treatment.
2. **Overflight Image**: The image shows infrastructure typical of a wastewater treatment plant, including circular clarifiers, sedimentation tanks, and other facilities associated with water treatment processes.
3. **Red Polygon Location**: The building within the red polygon is part of a larger complex that appears to be dedicated to wastewater treatment, as evidenced by the surrounding tanks and equipment.
4. **Surroundings**: The facility is located in a relatively open area, which is common for wastewater treatment plants due to the need for space and environmental considerations.*

This prompt structure ensures that the model's predictions are both interpretable and traceable, allowing for systematic evaluation of its reasoning and accuracy (Huang and Zhang, 2024). When evaluating other input modalities such as information from the Commercial Register, the structure remains the same, with a few small changes indicating the new input configuration.

Model and Experimental Configuration

The large language models used for this experiment were GPT-4o, GPT-5 from OpenAI[4)], and the open-weighted models Mistral Large 3[5)] and Qwen3-VL-8B-Instruct[6)] (Bai et al., 2025; Hurst et al., 2024; Mistral AI, 2026; A. Singh et al., 2025).

Four distinct setups were tested to assess the impact of different input combinations:

- **Setup 1**: Company name only.
- **Setup 2**: Company name, aerial image, and building polygon.
- **Setup 3**: Company name and description from the commercial register.
- **Setup 4**: Company name, description from the commercial register, aerial image, and building polygon.

This structured approach enabled a systematic comparison of model performance across varying levels of input complexity and information richness.

Evaluation

■ Accuracy

Accuracy of the MLLM response is evaluated by comparing the manually assigned sector classification to the predicted sector classification by the MLLM. The accuracy metrics can be read as the percentage of correctly classified economic sectors by the MLLM.

■ Confidence

When evaluating the quality of the MLLM response, we are also extracting the logarithmic probability (logprob) of the first response token (which is due to our prompt design always the letter of the WZ2008-section). We can interpret these logprobs as the confidence of the model in its decision and final response Z. Zheng, Feng, Li, Knoll, and Feng (2025). This confidence measure, as defined, can be compared with the accuracy of the response and yields an indication of how trustworthy the WZ-predictions of the models are. In doing so, we can assess whether there is a rule to follow (an optimal confidence threshold) to detect high quality predictions. If the model confidence shows a high correlation with accuracy, our pipeline might also generate reliable results for the imputation of missing data, where confidence of the response can replace accuracy as a quality metric. Furthermore, high confidence could be used as a quality flag for the predicted WZ. WZ predictions exceeding a sufficient confidence score could directly substitute human quality checks.

■ Consistency

After the assessment of the optimal experiment setup, we have run four iterations with the best

4 https://openai.com/
5 https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512
6 https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

performing configuration (Setup 3) with stable input sources and prompt configuration to analyze the consistency of the MLLM response (Liang et al., 2022). In detail, we assess the variability in the accuracy of the MLLM response.

8 Results

The accuracy of the classification model varied across the different experimental setups and language models. In particular, GPT-4o, Mistral Large 3 and Qwen3-VL achieved their highest accuracy values in Setup 3 (Company name and commercial register description), with 65.7%, 67.9% and 61.2% respectively. However, when utilizing GPT-5, Setup 4 (including geospatial data) yielded the highest accuracy, reaching 72%. These results indicate that the inclusion of both the commercial register description, aerial image, and building polygon as input features enhances the model's predictive performance, especially when leveraging the more advanced GPT-5 model. However, the company name alone as sole input source already yielded a minimum accuracy of almost 60 %. We interpret this predictive power of the company name as the richness of information that is already contained in the training data of the MLLM. These findings are in line with (Yüksel et al., 2026), who ran the same experiments with similar public data and open and closed source models.

The comparison between the different models further demonstrates the impact of model size, architecture and input complexity on classification accuracy, similarly to (Wei et al., 2022). While GPT-4o and Mistral Large 3 performed best with textual information from the commercial register, GPT-5 benefited from the combination of multiple data sources, achieving superior results in the most comprehensive setup. Due to computational constraints, Qwen3-VL was only evaluated on the textual setups (1 and 3). It achieved the worst results in the model comparison but is still competitive, especially considering its much smaller parameter count.

It is important to note that Mistral Large 3, also an open-weight model, achieves comparable results with both commercial models, pointing to the general robustness of the method regardless of MLLM architecture.

After comparing specific models, we were interested in the heterogeneity of the model response with respect to different economic sectors. The input source might yield stronger clues depending on the underlying economic activity of the company. We also wanted to elaborate on whether our design exhibits consistent performance across several economic activities. Figure 4 presents the accuracy of the WZ-prediction by the MLLM over the individual economic sectors (WZ2008,

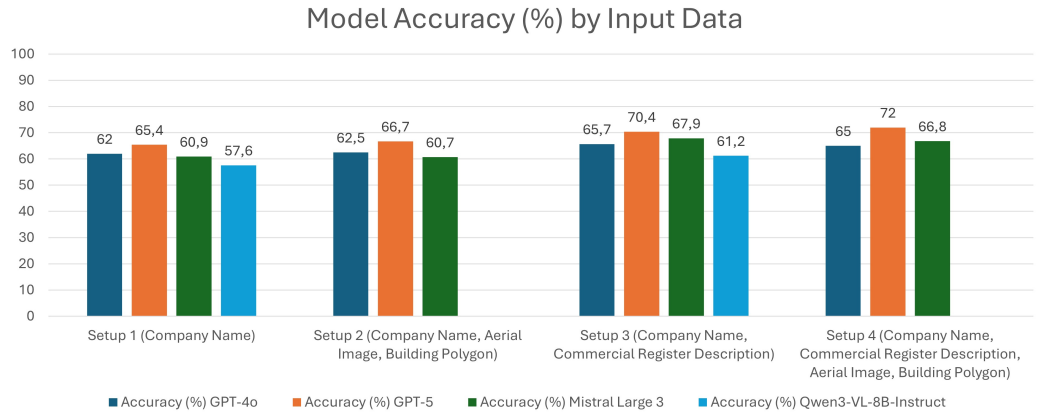


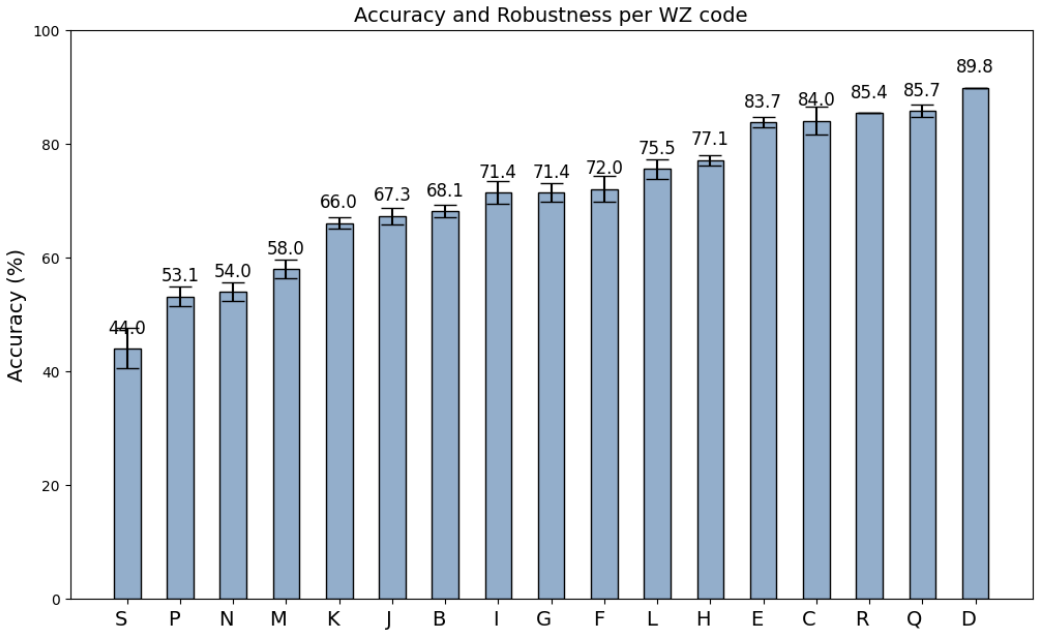
Figure 3: Comparison of model performance comparing proprietary with open models and varying input data

1-digit). Furthermore, it contains the results from the consistency checks of the four iterations.

The bars represent the accuracy of the WZ prediction for each WZ section for experiment setup 3. Clear differences can be seen between the individual sections, with the model using the WZ manual, the company name, and the business purpose from the commercial register as input. For example, for "S: Other service activities" (Sonstige DL), only 44% of the WZ codes are predicted correctly, whereas for "D:Electricity, gas, steam and air conditioning supply," it is almost 90% (the largest bar). The distributions at the top indicate the variance of the results. Here, we tested how reproducible the results are across four test runs. It is clear that, depending on the section, the results are relatively robust and thus reproducible (e.g., "D: Electricity, gas, steam and air conditioning supply" and also "B: Mining and quarrying"), but for other sections, such as "S: Other service activities", the results fluctuate significantly, indicating a lack of confidence in the model predictions (Geng, Cai, et al., 2024; Lee et al., 2024). This is not surprising as these are also the sectors with the lowest accuracy. However, accuracy and consistency do not seem to necessarily be correlated, as sectors with high accuracy in prediction do not always have the lowest variability in model responses.

To further understand the erroneous predictions, we were interested in the misclassification pattern as compared to the true economic sector validated by the human expert. Figure 5 provides an overview of this misclassification in the confusion matrix. We ran experiment setup 4 here as it yields the best overall results. The confusion matrix illustrates the performance across various economic sectors. The columns contain the WZ predicted by the MLLM, the rows contain the true WZ assigned by the human expert. In particular, the manufacturing sector ("C"), energy supply ("D"), water supply ("E"), human health and social work activities ("Q"), and arts, entertainment, and recreation ("R") exhibit the highest number of true positives. However, in some cases the model falsely predicted manufacturing ("C") while the true economic activity was in "G" – Wholesale and retail or "F"-Construction. This phenomenon can also be observed when classifications are performed manually as trade and manufacturing are often closely related and are both conducted in the same company. Conversely, other service activities ("S"), education ("P"), and administrative and support service activities ("N") demonstrate the lowest number of true positives which is also reflected in their low overall accuracy.

Another topic of our study was the potential of the MLLMs to decrease the amount of costly manual data quality checks. Therefore, we need an indication of the quality of the MLLMs prediction. Besides the accuracy over all predictions, we analysed the confidence in the model response for the single prediction and compared it with the accuracy of the prediction. If accuracy and confidence go hand in hand, MLLM predictions with high confidence can be seen as trustworthy and can substitute manual checkups, otherwise, some other form of AI driven solution is needed, such as human-AI collaboration (Vats et al., 2024). Figure 6 contains GPT-4o results and illustrates the relationship between the model's confidence (derived from the first predicted token's log probability, since it represents the class prediction) and its corresponding accuracy across different categories. Contrary to common expectations, the data reveals that increasing the confidence threshold does not consistently lead to higher accuracy, which was also shown by (A. K. Singh, Devkota, Lamichhane, Dhakal, and Dhakal, 2023). In fact, for two specific categories namely, "Professional, scientific and technical activities" - M and "Administrative and support service activities" - N. There is a clear trend where accuracy decreases as the model's confidence increases. It raises the important question of whether there exists an optimal level of model uncertainty or con-



WZ Code Legend

S: Other service activities
P: Education
N: Administrative and support service activities
M: Professional, scientific and technical activities
K: Financial and insurance activities
J: Information and communication
B: Mining and quarrying
I: Accommodation and food service activities
G: Wholesale and retail trade; repair of motor vehicles and motorcycles
F: Construction
L: Real estate activities
H: Transportation and storage
E: Water supply; sewerage, waste management and remediation activities
C: Manufacturing
R: Arts, entertainment and recreation
Q: Human health and social work activities
D: Electricity, gas, steam and air conditioning supply

Figure 4: Accuracy over different WZ codes (1-digit) and variability of responses across multiple iterations (Experiment Setup 3, GPT-5)

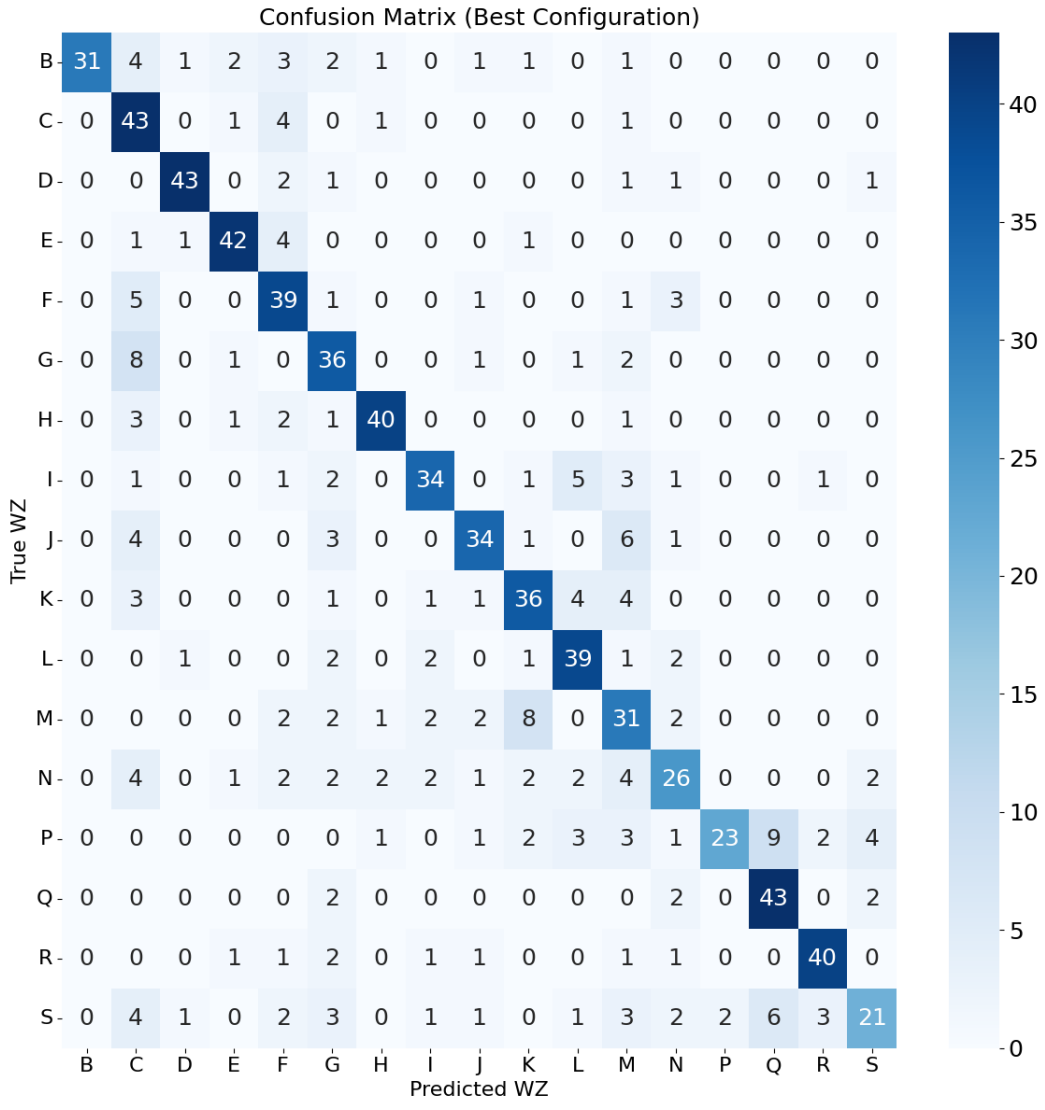


Figure 5: Confusion Matrix (for Experiment Setup 4, GPT-5) across WZ sections (1-digit)

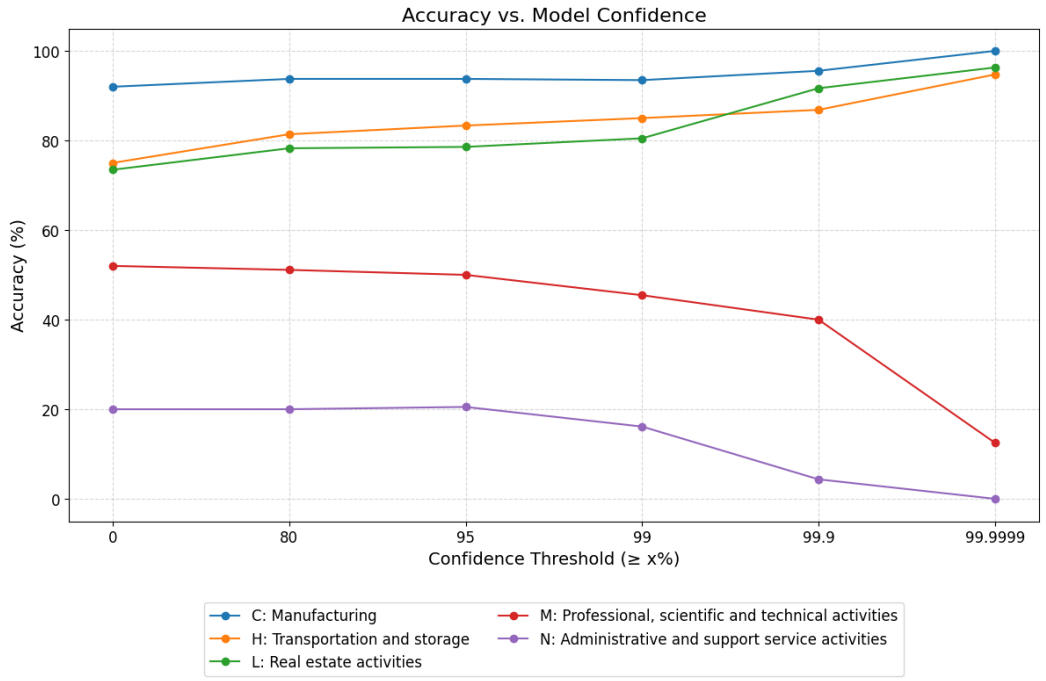


Figure 6: Accuracy of prediction against model confidence (Experiment Setup 2, GPT-4o) for selected WZ sections (1-digit)

fidence at which predictions should be reviewed by a human expert. Identifying such a threshold could enhance decision-making processes by ensuring that only predictions with an appropriate balance of confidence and accuracy are automated, while more ambiguous cases are subjected to human oversight. It is important to note that the output token log probabilities are not available for newer reasoning models such as GPT-5, so there is no direct way to derive the model's confidence from its response.

9 Conclusion and Outlook

The aim of our study was to evaluate the potential of MLLMs to support human experts in resource-intensive manual data quality checks. We propose a pipeline in which a multitude of information types, including novel unstructured geospatial data, is fed into a MLLM, which in turn proposes the economic sector of a company (WZ2008/ NACE) based on this information.

Our current findings indicate that MLLMs perform notably well in classifying companies whose economic activities are externally observable, such as those with factories, refineries, or water treatment facilities – because their physical infrastructure provides strong visual in-formation. In contrast, human validation remains necessary for complex cases involving heavily diversified companies or companies in the service sector. In such complex scenarios, the model often provides plausible reasoning that supports the decisions of human experts. Moreover, the WZ2008/NACE codes predicted by the MLLM can help resolve uncertainties faced by human experts when decisions are ambiguous and may even help reconcile inconsistencies between differing expert opinions. Furthermore, as the genAI models continue to improve and evolve to become more adept at utilizing geospatial information derived from images of company facilities, we find that the performance for the cases with strong visual information has improved (with GPT5).

With respect to the consistency of the responses, we find that predictions are not robust across several iterations with fixed prompts, depending on the predicted economic sector.

Additionally, when looking at the confidence scores, our results do not support the assumption of a relationship between accuracy of the prediction and confidence of the models as measured by token probabilities.

The pipeline we developed can be generalized to other use cases, such as the classification of company buildings (e.g., factories, multi-family residences) and the extraction of additional building features (e.g., sustainable building characteristics, installation of solar panels, and other adaptations in buildings over time). In a nutshell, geospatial data from satellites and spatial and city planning (e.g., digital twin data) offer a fruitful novel data source to enrich company data for data producers in several contexts, whether it be data quality management or new research questions such as sustainability when assessing the value of the real estate endangered by physical risks (floods or wildfires).

Our study is subject to several limitations, which provide paths for further research endeavours. First, the use of OpenAI's closed-source models entails significant costs and restricts transparency. A test with larger open models would benefit the generalizability and practical utility of our results. Second, reproducibility and model confidence are affected by the inherent variability in GPT-5's predictions. Novel concepts for a more objective confidence score for MLLM responses would support the evaluation of the trustworthiness of the predictions (Z. Zheng et al., 2025) and the potential of their practical usage. Approaches that use a "LLM-as-a-judge" (L. Zheng et al., 2023) or competition of different MLLMs (Y. Du, Li, Torralba, Tenenbaum, and Mordatch, 2023) are promising in this research field. Finally, limited access to external data sources, such as the Company Register Data, constrains the scope and robustness of our analysis.

References

Akhtar, M., Schlichtkrull, M., Guo, Z., Cocarascu, O., Simperl, E., and Vlachos, A. (2023). Multimodal automated fact-checking: A survey. *arXiv Preprint arXiv:2305.13507*.

Atanasova, P. (2024). Generating fact checking explanations. In *Accountable and explainable methods for complex reasoning over text* (pp. 83–103). Springer.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., and Zhu, K. (2025). Qwen3-vl technical report. *arXiv Preprint arXiv:2511.21631*.

Batini, C., and Scannapieco, M. (2016). *Data and information quality: Dimensions, principles and techniques* (pp. 3, 41–42, 181–182). Cham, Switzerland: Springer.

Berk Atıl, S., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., … Baldwin, B. (2025). Non-determinism of "deterministic" LLM system settings in hosted environments. *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, 135–148. Mumbai, India: Association for Computational Linguistics.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., … Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, *55*, 409–442.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection* (pp. 13–15, 143–146). Berlin, Germany: Springer.

Du, S., Zhang, Y., Zhu, L., Wang, S., Liu, Z., Zhou, H., and Du, S. (2025). A multimodal data fusion framework for urban functional zone mapping based on local-global information enhancement and adaptive spatial units: From standard datasets to global city validation. *International Journal of Applied Earth Observation and Geoinformation*, *144*, 104912.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate*. Retrieved from https://arxiv.org/abs/2305.14325

Eurostat. (2013). *European business statistics: Methodology—statistical classification of economic activities in the european community (NACE rev. 2)* (pp. 13, 21). Luxembourg: Publications Office of the European Union.

Fusco, G., and Aversano, L. (2020). An approach for semantic integration of heterogeneous data sources. *PeerJ Computer Science*, *6*, e254. https://doi.org/10.7717/peerj-cs.254

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113.

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., and Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6577–6595.

Geng, J., Kementchedjhieva, Y., Nakov, P., and Gurevych, I. (2024). Multimodal large language models to support real-world fact-checking. *arXiv Preprint arXiv:2403.03627*.

Goldblatt, R., Heilmann, K., and Vaizman, Y. (2020). Can medium-resolution satellite imagery measure economic activity at small geographies? Evidence from landsat in vietnam. *The World Bank Economic Review*, *34*(3), 635–653.

Haklay, M., and Weber, P. (2008). OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Hu, A., and Flaxman, S. (2018). Multimodal sentiment analysis to explore the structure of emotions. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discov-*

ery & Data Mining, 350–358. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3219819.3219853

Huang, J., and Zhang, J. (2024). A survey on evaluation of multimodal large language models. *arXiv Preprint arXiv:2408.15769*.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., and Kivlichan, I. (2024). Gpt-4o system card. *arXiv Preprint arXiv:2410.21276*.

Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. (2019). Supervised multimodal bitransformers for classifying images and text. *arXiv Preprint arXiv:1909.02950*.

Kotonya, N., and Toni, F. (2020a). Explainable automated fact-checking for public health claims. *arXiv Preprint arXiv:2010.09926*.

Kotonya, N., and Toni, F. (2020b). Explainable automated fact-checking: A survey. *arXiv Preprint arXiv:2011.03870*.

Koukaras, P. (2025). Data integration and storage strategies in heterogeneous analytical systems: Architectures, methods, and interoperability challenges. *Information*, *16*(11), 932. https://doi.org/10.3390/info16110932

Kutzner, T., Chaturvedi, K., and Kolbe, T. H. (2020). *OGC city geography markup language (CityGML) part 1: Conceptual model standard (OGC 20-010)*. Open Geospatial Consortium.

Kyriakos, C., and Vavalis, M. (2023). Business intelligence through machine learning from satellite remote sensing data. *Future Internet*, *15*(11), 355. https://doi.org/10.3390/fi15110355

Lee, M., Kim, K., Kim, T., and Park, S. (2024). Selective generation for controllable language models. *Advances in Neural Information Processing Systems*, *37*, 50494–50527.

Li, C., Li, X., Xie, Y., et al. (2023). Kosmos-2: Grounding multimodal large language models to the world. *arXiv Preprint arXiv:2306.14824*, 1, 5.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., and Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv Preprint arXiv:2211.09110*.

Liu, Y. Y., Zheng, Z., Zhang, F., Feng, J. C., Fu, Y. Y., Zhai, J. D., and Du, X. Y. (2026). A comprehensive taxonomy of prompt engineering techniques for large language models. *Frontiers of Computer Science*, *20*(3), 2003601.

Mai, G., Cundy, C., Choi, K., Hu, Y., Lao, N., and Ermon, S. (2022). Towards a foundation model for geospatial artificial intelligence (vision paper). *Proceedings of …*, 1–4.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., … Hu, Y. (2023). On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv Preprint arXiv:2304.06798*.

Miller, S. J., Howard, J., Adams, P., Schwan, M., and Slater, R. (2020). Multi-modal classification using images and text. *SMU Data Science Review*, *3*(3), 6.

Mistral AI. (2026). *Introducing mistral 3*. https://mistral.ai/news/mistral-3.

Nakamura, K., Levy, S., and Wang, W. Y. (2019). R/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv Preprint arXiv:1911.03854*.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., … Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. *arXiv Preprint arXiv:2103.07769*.

Perkins, M., and Roe, J. (2024). The use of generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning & Teaching*, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.22

Pradeep, R., Ma, X., Nogueira, R., and Lin, J. (2020). Scientific claim verification with VerT5erini. *arXiv Preprint arXiv:2010.11930*.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications*

of the ACM, *41*(2), 79–82.

Roberts, J., Lüddecke, T., Sheikh, R., Han, K., and Albanie, S. (2024). Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. *Proceedings of ...*, 554–563.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer.

Rostam, K., R., Z., and Kertész, G. (2025). Advances in pre-trained language models for domain-specific text classification: A systematic review. *ACM Transactions on Intelligent Systems and Technology*, *16*(6), 1–41.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., and Resnik, P. (2024). The prompt report: A systematic survey of prompt engineering techniques. *arXiv Preprint arXiv:2406.06608*.

Singh, A. K., Devkota, S., Lamichhane, B., Dhakal, U., and Dhakal, C. (2023). The confidence-competence gap in large language models: A cognitive study. *arXiv Preprint arXiv:2309.16145*.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., and Song, F. (2025). Openai gpt-5 system card. *arXiv Preprint arXiv:2601.03267*.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., … Quinn, J. (2021). Continental-scale building detection from high resolution satellite imagery. *arXiv Preprint arXiv:2107.12283*.

Skaug, L., and Nojoumian, M. (2025). A multimodal artificial intelligence framework for intelligent geospatial data validation and correction. *Inventions*, *10*(4), 59.

Song, S., Li, X., Li, S., Zhao, S., Yu, J., Ma, J., and Wang, M. (2025). How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, *37*(9), 5311–5329.

Strong, M., and Vlachos, A. (2025). TSVer: A benchmark for fact verification against time-series evidence. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 29906–29926). Suzhou, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1519

Tam, N. T., Toan, N. T., Cong, P. T., and Hung, N. Q. V. (2022). *Mapping and monitoring water areas with satellite images and deep learning*. https://doi.org/10.13140/RG.2.2.22869.70886

Theologitis, M., Dammu, P. P. S., Shah, C., and Suciu, D. (2026). *ClaimDB: A fact verification benchmark over large structured data*. Retrieved from https://arxiv.org/abs/2601.14698

UNECE. (2015). *Guidelines on statistical business registers* (pp. 69–70). Geneva, Switzerland: United Nations Economic Commission for Europe.

U.S. Census Bureau. (2026). *Statistics of u.s. Businesses (SUSB): methodology*. https://www.census.gov/programs-surveys/susb/technical-documentation/methodology.html.

Uzkent, B., Sheehan, E., Meng, C., Tang, Z., Burke, M., Lobell, D., and Ermon, S. (2019). Learning to interpret satellite images in global scale using wikipedia. *arXiv Preprint arXiv:1905.02506*.

Vats, V., Nizam, M. B., Liu, M., Wang, Z., Ho, R., Prasad, M. S., and Davis, J. (2024). A survey on human-AI collaboration with large foundation models. *arXiv Preprint arXiv:2403.04931*.

Vladika, J., and Matthes, F. (2023). Scientific fact-checking: A survey of resources and approaches. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 6215–6230). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.387

Wang, X., Kim, H., Rahman, S., Mitra, K., and Miao, Z. (2024). Human-llm collaborative annotation through effective verification of llm labels. *Proceedings of the 2024 CHI Conference on Human*

Factors in Computing Systems, 1–21.

Wang, X., Zhuang, B., and Wu, Q. (2024). Modaverse: Efficiently transforming modalities with llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26606–26616. Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/CVPR52733.2024.02512

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., and Fedus, W. (2022). Emergent abilities of large language models. *arXiv Preprint arXiv:2206.07682*.

Wu, J., Gan, W., Chao, H.-C., and Philip, S. Y. (2024). Geospatial big data: Survey and challenges. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. (2023). Next-gpt: Any-to-any multimodal llm. *arXiv Preprint arXiv:2309.05519*.

Yang, X., Qin, Q., Grussenmeyer, P., and Koehl, M. (2018). Urban surface water body detection with suppressed built-up noise based on water indices from sentinel-2 MSI imagery. *Remote Sensing of Environment*, *219*, 259–270.

Yao, B. M., Shah, A., Sun, L., Cho, J.-H., and Huang, L. (2023). End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2733–2743. ACM. https://doi.org/10.1145/3539618.3591879

Yuan, J., Li, H., Ding, X., Xie, W., Li, Y. J., Zhao, W., et al.others. (2025). Understanding and mitigating numerical sources of nondeterminism in LLM inference. *Proceedings of the 39th Annual Conference on Neural Information Processing Systems*.

Yüksel, A., Thiem, G., Walter, S., Felka, P., Werb, G. A., and Habernal, I. (2026). *MONETA: Multimodal industry classification through geographic information with multi agent systems*. Retrieved from https://arxiv.org/abs/2604.07956

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … Stoica, I. (2023). *Judging LLM-as-a-judge with MT-bench and chatbot arena*. Retrieved from https://arxiv.org/abs/2306.05685

Zheng, Y., Lin, Y., Chen, Y., et al. (2023). The rise and potential of large language model based agents: A survey. *arXiv Preprint arXiv:2309.07864*, 2, 17–18.

Zheng, Z., Feng, Q., Li, H., Knoll, A., and Feng, J. (2025). *Evaluating uncertainty-based failure detection for closed-loop LLM planners*. Retrieved from https://arxiv.org/abs/2406.00430

Zlatkova, D., Nakov, P., and Koychev, I. (2019). Fact-checking meets fauxtography: Verifying claims about images. *arXiv Preprint arXiv:1908.11722*.