

Extracting information on public water shortages using news media and LLMs: a pilot study in Sweden

Jeanne Fernandez^{1,2} Shorouq Zahra^{1,3} Mariana Madruga de Brito⁴ Johanna Mård^{1,2}

¹Uppsala University ²Centre of Natural Hazards and Disaster Science

³RISE Research Institutes of Sweden ⁴Helmholtz Centre for Environmental Research - UFZ

Abstract

Slow-onset events such as droughts lack a clearly defined onset, and systematic databases do not capture their impacts at fine scales. News reports can help fill this gap by providing information on how drought-driven water shortages affect communities and how authorities respond. Yet, extracting structured information from text remains challenging. Here, we use news articles and official public communications on water-saving recommendations and water-use restrictions in Sweden as a case study to benchmark LLMs for drought-related information extraction. We compare the proprietary GPT-5 series against three large open-access models (Llama3.3, Deepseek, and GPT-OSS) in extracting variables such as the date, location, and causes of water-saving measures. Using prompt engineering, structured outputs, and minimal post-processing, we evaluate performance using a test set of 210 manually annotated articles. The GPT-5 model series outperforms all open-access models across all variables. We find that, without task-specific fine-tuning, open-source models are sometimes outperformed by a keyword-based baseline, suggesting that general-purpose capabilities alone are insufficient for reliably extracting structured data. These findings underscore the challenges of monitoring water-shortage risks based on texts, particularly when extracting complex variables.

1 Introduction

Drought is a slow-onset event expected to become more frequent in many regions due to climate change (Chiang et al., 2021), requiring novel approaches to monitor its progress (Wens et al., 2026). While meteorological and hydrological indicators can characterize the physical hazard, information on impacts and response measures is fragmented across unstructured text sources. This challenge is pronounced in regions considered water-rich, such as Scandinavia, where the emerging risk of water

shortage is under-studied (Lindqvist et al., 2021). However, the 2018 European-scale drought marked a turning point in awareness of water scarcity, receiving widespread media coverage. In Sweden, news outlets and municipal websites disseminated information on water shortages, including water-saving recommendations and restrictions on non-essential water use. Although these measures are among the most common responses to water scarcity in Sweden (Deutschbein et al., 2023), they are not systematically recorded in regional or national databases. Thus, leveraging news aggregator datasets and text-mining approaches to record such measures over space and time can provide insights on societal responses to water scarcity and help document the reliability and resilience of drinking water systems in a changing climate.

However, extracting useful information from news requires more than identifying publication dates and municipality names. One must separate saving recommendations from mandatory restrictions, identify whether measures affect an entire municipality or only parts of it, and extract precise start and end dates for these measures. Also, to support downstream analyses, information extraction methods must identify the cause of such shortages, distinguishing physical drivers (e.g., precipitation deficits and increased evapotranspiration associated with high temperatures) from demand-driven factors (e.g., high water consumption).

A key challenge then becomes the journalistic style of news texts, making it difficult for conventional natural language processing (NLP) approaches to reliably assess the relationships among a location, a date, a measure, and causes. This is particularly problematic for texts with information about multiple co-occurring or sequential measures and covering multiple locations. In addition, the high semantic similarity in the formulation of mandatory water-use restrictions and water-saving recommendations makes it difficult to characterize

the type of measure in place. Both measures may contain terms associated with water conservation. Finally, characterizing article relevance for the proposed analysis requires context understanding and cannot be achieved with simple classification methods.

Advances in large language models (LLMs) have opened new opportunities to extract structured information about climate events from texts. Their in-context understanding abilities are valuable for such tasks (Dong et al., 2024). Several studies have applied LLMs to build climate hazard databases from text data. For instance, Li et al. (2024) and Vela-Tambo et al. (2025) find that LLMs are promising tools for climate-related data extraction from reports and news articles, even when performing cross-lingual information extraction in the case of the latter.

Nonetheless, LLM performance on domain-specific tasks and on extracting subtle signals of drought risk, such as water-shortage measures, has not been evaluated. Moreover, open-access models are cheaper and facilitate replicability as they can be run locally, often under non-restrictive licenses. Despite being trained for a wide range of general tasks, their baseline performance on structured data extraction relative to proprietary models and keyword-based NLP methods is unclear.

We propose an LLM-based method to extract data on public water supply shortage risks in Sweden from news coverage of water savings and water-use restrictions (Figure 1). The overarching goal is to identify risk hotspots for water supply, i.e., places facing frequent and long-lasting restrictions, as well as the reasons why restrictions are applied. Specifically, the resulting spatiotemporal database can help identify patterns in municipal water-management responses to droughts and understand how these escalate from warnings to water-use restrictions (Figure 2). In combination with additional datasets (e.g., drought indicators, municipal water source characteristics), the database can contribute to a more holistic understanding of water shortage risks by linking drought hazards, water-system vulnerabilities, and implemented responses. This, in turn, can support drought preparedness and provide relevant insights for water managers and planners.

We benchmark the GPT-5 series proprietary models against three open-source models (Llama 3.3, Deepseek, GTP-OSS) and one hand-crafted rule-based baseline in extracting structured infor-

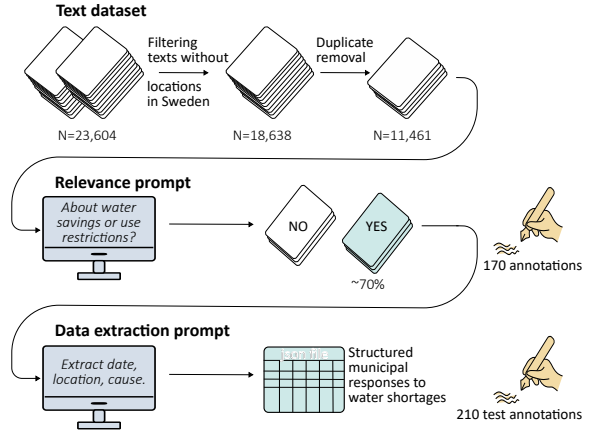


Figure 1: Data extraction pipeline and its evaluation

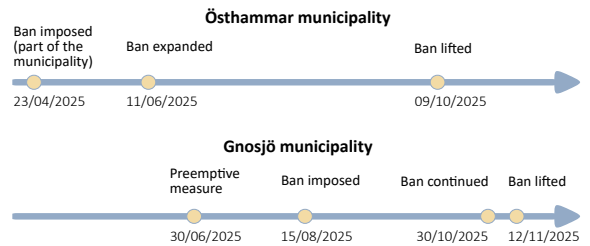


Figure 2: Example reconstruction of a timeline of measures for Östhammar and Gnosjö municipalities in 2025, based on information extracted from news articles with GPT-5.5.

mation from unstructured texts. The experimental setup relies on prompt engineering and limited post-processing to evaluate model performance under minimally tuned conditions. All prompts, code, and metadata are made available on GitHub.¹

2 Related work

News-based analysis of drought and water scarcity is a growing research field, showing that news can offer extensive, yet still underutilized, information on this hydrological extreme. Existing NLP applications have focused on detecting and classifying drought impacts into broad classes (Sodoge et al., 2023; Vela-Tambo et al., 2025). Studies have also linked trends in news coverage of droughts to domestic water use patterns, focusing on communication volume and sentiment analysis (Lee et al., 2026). More recently, embedding-based semantic similarity methods have been used to identify fuzzy concepts, such as drought-related groundwater conflicts, in news articles (Sodoge et al., 2026).

¹https://github.com/jeannefdz/swedish_water_restrictions

More broadly, LLMs are increasingly used in climate research to extract structured information from unstructured text due to their contextual understanding and few-shot learning capabilities (Brown et al., 2020; Vaswani et al., 2017). Applications include the construction of climate impact databases (Li et al., 2024; Vela-Tambo et al., 2025; de Brito et al., 2026), monitoring societal impacts of droughts via social media (Wang et al., 2025), hazard fact-sheets and knowledge graphs (Ronco et al., 2026), or evidence of gaps in the regional coverage of hazards in scientific literature (Carvalho et al., 2025). We extend these applications by extracting reports on municipal drinking water saving recommendations and use restrictions from Swedish media texts, providing a structured dataset of drought response measures.

Despite the success of LLMs on some of these tasks, several factors influence their performance. Larger commercial models often outperform open-access models, although open-access models can match the performance of proprietary LLMs on some tasks (Li et al., 2024). Performance also depends on prompt design, with prompt length and detail needing to be adapted to model size (Vela-Tambo et al., 2025). Additionally, both Vela-Tambo et al. (2025) and Li et al. (2024) find that LLMs struggle to produce structurally valid outputs, with performance depending on model size and the effectiveness of supporting pre- and post-processing techniques such as schema validation.

3 Text corpus

The data used is based on a corpus of Swedish texts on municipal water-use restrictions (also referred to as “watering bans”) and water-saving recommendations. The choice of data source is justified by ease of access, stability over time, and spatiotemporal coverage. While municipal websites represent an authoritative source regarding current restrictions, webscraping requires specific extraction scripts for each webpage structure and these structures evolve over time. In addition, older website versions containing historical restrictions may be archived inconsistently, making it challenging to retrieve data on past measures.

We retrieved articles² from Swedish sources

²In this paper, we use “article” as a generic term for records retrieved from the Retriever media archive. Most records originate from printed and online news media but the corpus also includes municipal communications and press-releases.

in the Retriever media archive³, a database with 4,574 Swedish language media sources⁴, including printed press, web, and TV/Radio. We collected both editorial and non-editorial news articles and official public communications (e.g., press releases and public advisories) published between January 1st, 2000 and December 31st, 2025, using keyword searches of titles and headlines. The search terms, informed by similar searches (Quesnel and Ajami, 2017) and expert knowledge, comprise four separate queries (Appendix A, Table 6). These queries were developed by testing alternative combinations of synonyms and expressions for watering bans and water savings and checking the effect of additional keywords on the number of articles returned (Appendix A). This strategy maximized recall, resulting in a corpus of 23,604 texts.

The subsequent pre-processing steps involved cleaning the texts for more efficient NLP processing, removing articles that mentioned no location in Sweden, and clustering duplicates (Figure 1). To reflect municipal-level management of public water supply in Sweden, we only included articles that mentioned either a municipality (i.e., local administrative unit level 2 (LAU 2)) or a statistically defined urban area (“*Tätort*” in Swedish). An urban area, as defined by Statistics Sweden, is a continuous built-up area with at least 200 inhabitants that meets specific distance criteria between buildings (Statistikmyndigheten SCB, 2023). We excluded articles that only referred to larger or smaller administrative units (e.g., regions, counties, districts, neighborhoods) as these introduced noise in the dataset. We used regular expressions (regex) patterns to search the articles’ main text for municipality and urban area names, including their lexical variations (see Appendix A for details). This simple filtering was found to perform better than more advanced named entity recognition methods for Swedish locations, based on manual checking of some articles. In this step, 4,966 articles were removed.

Duplicate and near-duplicate articles were identified using Term-Frequency Inverse Document Frequency (TF-IDF) and cosine similarities. This approach weighs the relative importance of each term in the corpus, giving a higher weight to terms that are frequent in a document but rare in the overall corpus. It ensures that the most informative

³<https://retriever.se>

⁴Retriever search interface, March 2026

words are given more weight and can capture partial matches between paraphrased journalistic texts (Rajaraman and Ullman, 2011). To build the TF-IDF matrix, we used 5- to 8-character n-grams. The cut-off value for defining duplicates was set to 0.8 (1 representing full cosine similarity) as many texts in our corpus are near duplicates with identical text bodies but paraphrased titles, leads, captions, or closing paragraphs. The choice of a high cut-off value is consistent with other duplicate-detection studies (de Brito et al., 2025; Madureira et al., 2026) and was validated by examining the distributions of cosine similarities (Appendix A, Figure 4) and duplicate clusters for a subset of 300 articles. The final corpus contains 11,461 unique texts. A quick sampling of these texts showed that most were relevant to our analysis. Thus, instead of further pre-filtering the dataset, we relied on the LLMs to determine article relevance (see Appendix B).

4 Extraction pipeline and annotations

We focus on two types of response measures to water shortages: water-saving recommendations (“*spara vatten*” in Swedish; i.e., “save water”) and water-use restrictions (“*bevattningsförbud*” in Swedish; i.e., “watering bans”). These measures, which signal potential water stress, are communicated to the general public by or on behalf of municipalities and water utilities. We distinguish the measure types to capture the progression from water savings to bans being implemented, continued, expanded spatially, lifted, or avoided (Figure 2). While watering bans require compliance between a clear start and end date (i.e., when imposed and lifted), water savings are more general recommendations that do not follow a strict timeline but may have a measure date.

4.1 Output schema

We extract information on the measure type, dates, location(s), and their cause(s), as shown in Table 1. Each news article may contain information about multiple measures at different locations, but each data entry must correspond to one measure and one location. For each field, in addition to format requirements, constraints may apply, limiting a model’s responses to predefined categories or locations within Sweden, or specifying the conditions under which a field should be filled or left blank.

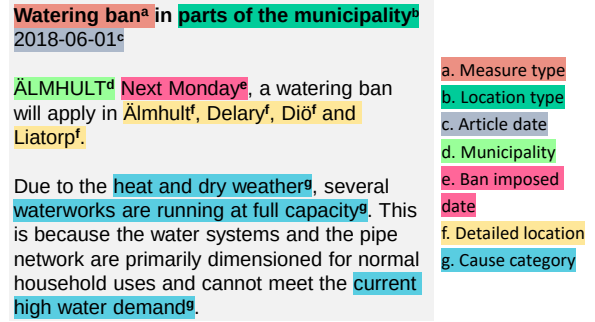


Figure 3: Example article (summarized) highlighting text segments that may inform different extraction fields.

4.2 Gold standard annotation

The annotation process was divided into two steps: first, for article relevance and second, for data extraction (Figure 1). To evaluate how well models could recognize relevant texts, a set of 170 articles was randomly sampled, stratified by year, with at least 2 and up to 15 articles per year. An article is relevant if it discusses a newly introduced, ongoing, or lifted watering ban, or communication that encourages the public to save water due to drought conditions or water supply disruptions. As non-relevant, we consider communication on water supply that does not explicitly encourage the public to take action about water use: debates, opinions, water saving tips, economic or technical innovations, and measures taken at the regional-level that do not directly concern citizens. In the sample, 123 articles (~70%) were found relevant.

To create a gold standard for data extraction, we expanded the initial sample of 123 relevant articles to 267, representing approximately 2% of the cleaned corpus. The additional articles were also sampled randomly and stratified by year to ensure temporal representation. These articles were randomly divided into two sets: 57 for model development and 210 for testing, with an even representation of dataset years in each set. Annotation of the development set was performed by two annotators to establish clear annotation guidelines, and conflicts were resolved by discussion among the two annotators until consensus was reached on all items. The test set was annotated by a single annotator with the same guideline instructions provided to the LLMs; the second annotator verified cases rated as complex by the first annotator. An example article and its annotations are shown in Figure 3.

Field	Description	Constraints	Format*
article date	Article publication date.	Must be a valid date.	date***
municipality	Municipality concerned by the measure.	LAU 2, corresponding to municipalities in Sweden. Only one municipality may be selected.	string
detailed location	Affected urban areas in a municipality.	Referenced urban area(s) in Sweden. Multiple locations allowed.	string list***
location type	Measure scope.	Categorical: entire municipality or part of the municipality. Only one allowed.	string
measure type	Type of measure applied.	Categorical: ban imposed, ban lifted, pre-emptive measure**, ban continues, ban expanded, or ban avoided. Only one allowed.	string
measure date	Date a measure is announced.	Does not apply to "ban imposed" or "ban lifted" measures.	date***
ban imposed date	Date a ban is imposed.	Only filled if measure type is "ban imposed".	date***
ban lifted date	Date a ban is lifted.	Only filled if measure type is "ban lifted".	date***
cause category	Indicates the cause of the measure.	Categorical: lack of rain, high water consumption, warm temperatures, low surface water levels, low groundwater levels, low water levels (unspecified), water repairs, water shortage, power outage, precaution, other, and NA (not available). Must be at least one.	string list

Table 1: Descriptions of extracted fields, including definitions, constraints, and expected formats. (*All dates are in YYYY-MM-DD format. **Preemptive measures refer to water saving recommendations. ***Nullable fields.)

4.3 Prompt development and LLM configuration

The extraction pipeline consists of two main prompts, one for assessing article relevance, returning a binary output (relevant/non-relevant) and the other for data extraction, with separate prompts for each extracted field (Table 1). While all articles are in Swedish, prompts were written in English, and all outputs, except location names, were returned in English, in accordance with the higher performance of LLMs with English-language prompts (Fu et al., 2022; Vela-Tambo et al., 2025). The articles were not translated as Swedish is not usually considered a low-resource language. Although the English language often dominates most open-source and commercial LLM training datasets, the availability of Swedish digital resources and the multilingual capabilities of LLMs suggest that Swedish texts can be processed reliably (Borin et al., 2023). This assumption was tested through model performance evaluation on the relevance and extraction prompts.

For the data extraction prompt, we applied a role-prompting strategy by instructing the model to act as an accurate and precise system, fluent in both Swedish and English. We also used a zero-shot prompting strategy so as not to bias the tested models toward a few examples and to reduce costs.

In addition, we relied on explicit variable and category names to reduce ambiguity. Although we aimed for concise prompts, we prioritized comprehensive descriptions over prompt brevity given the relatively large context sizes of all tested models. See Appendix C for a prompt example.

For the relevance classification task, we considered Llama 3.3 (70B) and GPT-5.2⁵. For data extraction, we used LLMs from different model families, including the GPT-5 series (GPT-5.2, GPT-5.4, GPT-5.4-mini, and GPT-5.5) (OpenAI, 2026) and the largest model size of three open-access model families: Llama 3.3 (70B) (Grattafiori et al., 2024), GPT-OSS (120B) (OpenAI et al., 2025), and Deepseek R1 (70B) (Guo et al., 2025).

To benchmark performance, all models received identical prompts. We also set reasoning levels for the GPT-5 models to “none” which improves efficiency and reduces costs as higher reasoning levels require more tokens.⁶ For open-access models, we test performance at temperature zero (i.e., deterministic outputs via greedy decoding). While temperature manipulates the token distribution, directly affecting the choice of the next token, reasoning is unique to the GPT-5 series and reflects the

⁵GPT-5.2 was the latest available GPT-5 model at the time.

⁶<https://developers.openai.com/api/docs/guides/reasoning>

model’s capacity to use “internal reasoning” tokens to solve complex tasks.

We performed additional experiments to study the effects of increasing temperature (for open-access models) and reasoning effort (for the GPT-5 series) on data extraction accuracy. We tested the following temperatures: 0.25, 0.5, 0.75, 1.0, and 1.5. To account for stochasticity, we averaged error scores of three separate runs. For the GPT-5 models, we run a single experiment for each reasoning level (see Appendix D for details).

The final relevance and data extraction prompts result from an iterative refinement process using GPT-5.2 and GPT-5.5 to maximize performance on the development set. To support prompt development and facilitate error analysis, the models were required to return justifications, text excerpts, and summaries, in addition to the extracted information. Moreover, the models were configured to prioritize output consistency and stability in JSON format. This format also facilitates the evaluation of model outputs against annotations. While the OpenAI API⁷ already offers the option for structured JSON output, open-access models do not – therefore, we perform those experiments via the Ollama API.⁸

4.4 Baseline data extraction

To assess the benefits of contextual and semantic understanding in extracting unstructured data, we implemented a rule-based extraction baseline that uses lexical matching and heuristic rules optimized on the development set. Similar to the LLM-based extraction, this deterministic pipeline processes the articles and builds a structured table of measure records, with one entry per measure-location pair.

Location fields were extracted using dictionary matching against predefined localities in Sweden and common lexical variations. The spatial scope of a measure (i.e., the location type, either full or partial coverage of a municipality) is inferred from the presence of sub-municipality locations in the text or specific textual cues. Measure types and their causes were categorized using regex based on keywords and expressions commonly found in Swedish news. Temporal data was extracted using regex and named entity recognition, with relative dates resolved using the article publication date.

We linked measures, locations, and dates based

on their proximity in the text. Each measure is associated with a short context window of surrounding sentences, which defines the search scope for locations and dates. While this approach reduces false associations, it may miss information distributed across multiple paragraphs.

5 LLM performance evaluation

For the binary classification task (i.e., categorizing articles as relevant or non-relevant), we used standard evaluation metrics: overall accuracy, precision, recall, and F1 score.

For the data extraction task, distance-based metrics were used to quantify how closely model outputs match the gold annotations. We calculated an aggregated difference metric for each field, representing the distance between the gold standard and the system outputs, ranging from 0 (perfect match) to 1 (complete mismatch), and averaged it over all measures extracted from an article. Missing information was penalized with a “null penalty”, setting the error score to 1 for the missing field. These metrics were calculated differently depending on the field data type (Table 2). We supplemented the metrics used by Li et al. (2024) with an additional date-distance metric with a 28-day cutoff, so that the model receives a maximum error score if the date is incorrect by more than 4 weeks. We also derived an overall performance score by aggregating all the field-specific scores.

A single article can yield more than one entry (in our case, 1.6 entries per article on average in the test set), and an LLM may return these measures in any order; therefore, we matched entries found by the LLMs to the gold annotation for each article using three fields: the measure type, the location type, and the municipality. If the LLM misses an entire entry or finds an extra one, it receives the “null penalty” for each field in the extra or missing row, which can significantly increase the error rate. To distinguish errors due to missing or extra entries from errors due to inaccurate outputs, we first calculated a main metric (labeled “All”) that accounts for complete mismatch penalization. Then, we repeated the data extraction evaluation by excluding missing or extra entries, thereby eliminating null-penalty effects arising from these additional or omitted measures. This alternative metric is labeled “Match”.

⁷<https://developers.openai.com/api/docs/guides/structured-outputs>

⁸<https://docs.ollama.com/capabilities/structured-outputs>

Format	Description	Expression
String	Exact location name matching.	$d_{\text{str}}(a, r) := \begin{cases} 0, & \text{if } a = r \\ 1, & \text{otherwise} \end{cases} \quad (1)$
Dates	Here, Δ_{days} is the absolute difference in days between the dates a and r . If Δ_{days} is longer than 28 days (4 weeks), it is considered incorrect.	$d_{\text{date}}(a, r) := \begin{cases} \frac{\Delta_{\text{days}}}{28}, & \text{if } \Delta_{\text{days}} < 28 \\ 1, & \text{otherwise} \end{cases} \quad (2)$
Categorical sets	Jaccard distance between two sets of pre-defined categories or location names that must match exactly.	$d_{\text{set}}(a, r) := 1 - \frac{ a \cap r }{ a \cup r } \quad (3)$

Table 2: Difference metrics used for evaluating extracted fields, where a is the annotated value and r is the retrieved value. Difference metrics are in the range $[0, 1]$, where $\downarrow$ lower is better.

5.1 Relatedness classification results

Table 3 outlines recall, precision, accuracy and F1-scores for GPT-5.2 compared to Llama3.3. GPT-5.2 exhibits a slight bias towards over-inclusion (due to producing more false positives). The false positives are articles discussing water saving campaigns and tips to save water; opinions, questions or complaints about water savings; or the effects of water savings on overall consumption. Error analysis shows that while they relate to the overall topic, they do not cover decisions about water-saving measures. The false negatives also comprise borderline cases where messaging about water savings appears, but it is difficult to determine whether it applies to the local- or regional-scale. Llama3.3, on the other hand, produced more false positives and false negatives. The false positives can be traced to the inclusion of context articles on water resources, water supply disruptions, non-municipal measures, or technical and economic initiatives to save water. A third of the false negatives occur due to difficulty in isolating relevant content from irrelevant information, or, in a case specific to Llama3.3, failure to recognize texts as articles, despite prompt instructions that all documents should be treated as such (see Appendix B). However, most false negatives correspond to misclassifying the text, even when the summary of the article content is accurate.

Metric	GPT-5.2 (no reasoning)	Llama3.3:70b (T=0)
Accuracy	0.91	0.60
Recall	0.96	0.72
Precision	0.91	0.73
F1 Score	0.94	0.72

Table 3: LLM performance on relevance classification.

5.2 Data extraction results

Table 4 presents the results of the data extraction by field for the test dataset (along with average scores, coverage, runtime, and costs) for GPT-5.5, the three open-access models, and the baseline extraction (see Appendix E, Table 7 for development set results). Similar average scores on the development and test sets indicate a stable performance for all model families, suggesting that the prompts are sufficiently robust for this use-case. The results that account for all entries (columns “All” in Table 4) show that GPT-5.5 outperforms all open-access models by over 0.3 points. The three open-access models present similar average scores, but for many variables Llama3.3 and Deepseek perform better than GPT-OSS, with lower field-specific error rates. The extraction results are, on average and for several fields, better with the baseline NLP than with open-access models. Specifically, all open-access models show error rates above the baseline error rate for detailed locations, measure type, and cause category, showing that generative AI does not automatically offer advantages for complex extraction tasks. The variables with the greatest error rates are measure type, measure date, and cause category for GPT-5.5. The open-access models struggle with almost all variables except dates of lifted bans.

When considering only matched pairs of entries (error scores in the “Match” columns in Table 4), we see lower error rates, suggesting that missed or extra entries largely influence the overall score. The baseline and all models find fewer entries on average per article compared to the gold-standard test annotations, except for GPT-OSS, which returns more entries. Note that models are permitted to return no entries if they fail to find any relevant

Metric / Set	GPT-5.5		Llama3.3 (70B)		Deepseek R1 (70B)		GPT-OSS (120B)		Baseline	
	All	Match	All	Match	All	Match	All	Match	All	Match
article date	0.13	0.00	0.33	0.01	0.36	0.01	0.55	0.04	0.35	0.00
municipality	0.17	0.04	0.52	0.29	0.58	0.34	0.73	0.42	0.69	0.52
detailed location	0.19	0.07	0.88	0.82	0.86	0.78	0.92	0.84	0.67	0.50
location type	0.16	0.03	0.61	0.42	0.60	0.37	0.77	0.51	0.75	0.62
measure type	0.27	0.16	0.76	0.65	0.70	0.53	0.86	0.71	0.69	0.53
measure date	0.30	0.26	0.54	0.54	0.53	0.56	0.50	0.61	0.35	0.32
ban imposed date	0.11	0.10	0.44	0.53	0.40	0.52	0.40	0.47	0.30	0.34
ban lifted date	0.03	0.02	0.09	0.11	0.11	0.14	0.07	0.09	0.11	0.13
cause category	0.32	0.22	0.88	0.83	0.80	0.68	0.89	0.76	0.70	0.54
avg. score	0.20	0.10	0.57	0.47	0.55	0.44	0.64	0.49	0.51	0.39
avg. score (dev)	0.22	0.10	0.56	0.47	0.53	0.43	0.62	0.50	0.54	0.45
coverage	98.57%		100%		100%		100%		100%	
avg. entries per article	1.47		1.18		1.01		1.9		1.25	
runtime (min)	60		62		63		42		0.5	
approximate cost (USD)	10		4		4		5		0	

Table 4: Comparison of extraction performance across open-access models and GPT-5.5 on the **test set**, for both “All” entries versus for an alternative “Match” test set that excludes superfluous or missing entries. Temperatures and reasoning levels are set to 0.0 and “none”, respectively. Coverage is the percentage of articles for which the models produced at least one entry. The average score on the development set (dev) is also provided for comparison with the test-set score. The lowest error rates are shown in **boldface**.

information in an article. This only occurs for a few articles with GPT-5.5.

Upon inspecting the outputs, we find that GPT-5 errors can be attributed to prompt interpretation. Date errors often stem from the extent to which the model or the annotators extrapolated dates from the article’s date, as authorized by the prompts. In terms of measure type, a frequent error is treating a continuing watering ban as a newly imposed ban, which can be explained by vague formulations (e.g., “there is a ban”) in some of the articles. Differences in the detailed locations arise because the model extracts locations such as “west of the municipality”, whereas annotators focused only on named locations.

The cause category field is problematic for all GPT-5 models, which can be attributed to its high subjectivity. While most causes are well captured, the lower performance is due to frequent additions to or omissions from the cause category. For example, a common error is the omission of the category “lack of rain”, broadly described as dry conditions in the prompts. This may reflect a relatively inclusive annotation approach, which consistently classified references to dry summers as “lack of rain”. The model otherwise over-reports the categories “water shortage”, “high water consumption”, and “other”.

Since the cause category is a short multi-label field, using Jaccard distance emphasizes the error

score arising from mismatches in the number of reported categories. For this reason, we also include precision, recall, and F1 scores for each cause label in Table 5, for both GPT-5.5 and Llama 3.3. The scores confirm that GPT-5.5 is relatively strong at predicting most labels, distinguishing explicit labels (e.g., “power outage”, “water repairs”) and fine-grained cause categories (e.g., “low groundwater levels”, “water shortage”). However, the model performs poorly for broad categories such as “precaution” and “other” which are not as clearly associated with linguistic indicators. (The same scores for the development set are reported in Table 8, Appendix E.)

For the open-access models, performance is lower across all fields than with GPT-5.5, except for the ban-lifted date. This cannot be explained by interpretation issues. Indeed, the models have difficulty assigning event dates to the correct columns (e.g., measure date versus ban date) and sometimes fail to convert textual relative dates to the desired YYYY-MM-DD format. Open-access models also struggle with categorical variables (i.e., measure types, location types, and causes), which require mapping textual data to predefined labels. For the cause field, for instance, Llama 3.3 defaults to the single label “water shortage” for close to 70% of the returned entries and substantially under-reports labels, predicting only 0.73 labels per entry on average for the test set, compared to 1.71 in the gold

		GPT-5.5			Llama 3.3			
	Label	Precision	Recall	F1	Precision	Recall	F1	Support
0	High water consumption	0.779	0.918	0.843	0.714	0.068	0.125	73
1	Lack of rain	0.833	0.650	0.730	0.000	0.000	0.000	100
2	Low groundwater levels	0.846	0.819	0.832	0.875	0.149	0.255	94
3	Low surface water levels	0.692	0.643	0.667	0.000	0.000	0.000	14
4	Low water levels unspecified	0.550	0.688	0.611	0.286	0.125	0.174	16
5	NA	0.868	0.575	0.692	0.250	0.012	0.024	80
6	Other	0.200	0.750	0.316	0.000	0.000	0.000	4
7	Power outage	1.000	1.000	1.000	0.800	1.000	0.889	4
8	Precaution	0.474	0.474	0.474	0.000	0.000	0.000	19
9	Warm temperature	0.870	0.839	0.855	0.000	0.000	0.000	56
10	Water shortage	0.667	0.784	0.721	0.161	0.588	0.253	51
11	Water repairs	0.909	0.882	0.896	0.917	0.324	0.478	34
12	Waterworks at capacity	0.816	0.851	0.833	0.000	0.000	0.000	47

Table 5: Precision, recall, and F1 scores for multi-label classification of the cause category for the **test set**. The support column represents the number of examples available in the gold annotations for each label.

annotations and 1.69 predicted by GPT-5.5. As a result, evaluation scores are very poor for most cause categories (Table 5). In addition, when several municipalities are mentioned in a text, open-access models sometimes report them as a single entry rather than as separate entries, as instructed. Finally, the outputs tend to mix Swedish and English, despite instructions to respond in English. For these reasons, the baseline extraction method, which enforces a structured output format and relies on clearly defined rules, achieves a lower overall error rate.

To provide further insights into the identified error types, examples of text excerpts, gold-standard annotations, and model extractions are shown in Tables 11 (for GPT-5.5) and 12 (for Llama 3.3) in Appendix F.

6 Discussion and conclusion

This study benchmarks proprietary LLM and open-access LLM performances against baseline NLP methods for the extraction of complex relational data on a specialized topic: the communication of measures on water use during drought and water supply disruptions. The motivation for the study was to assess the extraction performance without extensive prompt refinement or post-processing, evaluating models under comparable lightly engineered conditions.

As discussed in similar studies (Li et al., 2024; Castro et al., 2024), the performance of proprietary models surpasses that of open-access models, including larger ones, in terms of stability and capacity to generate structured outputs. The limited

effects of model temperature and reasoning levels confirm that model size, rather than configuration, is a key factor determining performance.

We find that prompt-engineering alone does not produce better outputs compared to hand-crafted (hence time-consuming) NLP approaches, even for the largest open-access models available from three main LLM families. In this study, prompt optimization may have benefited GPT-5 models, since we did not revise the prompts for the open-access models. However, we suppose that the experimental setup with limited fine-tuning, training, and post-processing explains the lower performance of the open-access models. Specifically, these multilingual models may not have been trained on sufficient amounts of Swedish text. Therefore, fine-tuning on Swedish news articles, more refined guidelines for inferring the cause category, and post-processing could improve the quality of the outputs. A practical implication is that although their access is free, a considerable time investment may be required to achieve satisfactory performance with general-purpose open-access models, creating a trade-off among time, costs, and environmental impacts. To extract structured, precise, and accurate data in this setup, and without any fine-tuning, we have to rely on proprietary models from providers like OpenAI, which is costly. In addition, the extraction process may be difficult to replicate due to model deprecation and the need to refine the prompts for each new model release.

Limitations

We recognize limitations in the development of the gold standard, as inter-annotator agreement was not measured. For the development set, two annotators collaborated to iteratively refine the annotations through joint discussions, which leaves room for bias. For the test set, annotations were performed by a single annotator, while the second annotator reviewed edge cases and proposed revisions where necessary. In addition, both annotators speak Swedish as a second language. This may influence the interpretation of certain journalistic expressions.

The study also presents limitations associated with the experimental setup. First, the prompts were developed for GPT-5.2 and 5.5 and were not revised for other LLMs, which favors the larger models. In addition, limited post-processing and no fine-tuning were performed for the open-access models, steps that would have likely increased model performance. Finally, we were not able to test models extensively trained on Nordic languages due to a low maximum token limit (i.e., context length) and instead relied on multilingual models predominantly trained on English (and Chinese in the case of DeepSeek). Thus, further testing is needed before concluding that open-access models are not suited to complex data extraction for a niche topic.

When applied to the entire dataset, the outputs of the proposed extraction method will be affected by the extent of media coverage of water-use restrictions and the quality of the reported information. The Retriever database includes local news sources, which ensures broad coverage of localized events across Sweden. However, the lack of reporting on a municipality cannot be equated to the absence of a watering ban. Therefore, triangulation approaches combining news media extraction and surveys could also be explored.

Ethics statement

The paper does not raise any ethical issues. The news articles were accessed via a media archive made available through university library agreements, and, in accordance with the copyright and archive data use requirements, no texts were shared outside the research group.

Acknowledgments

J.F. and J.M. acknowledge funding from FORMAS under the project ‘Leap in Swedish freshwater monitoring for water supply risk mitigation’ (LEAP4WATER, project ID:2022-02148_Formas). J.F. further acknowledges support from CNDS for collaboration with the UFZ. S.Z. acknowledges the ICE Datacenter at RISE Research Institutes of Sweden for providing resources to conduct experiments on open-access models, as well as for providing OpenAI Platform credits to use the OpenAI API. M.M.d.B. discloses funding from the Volkswagen Foundation through the CASCADE project, funded under the Earth Systems Sciences initiative, as well as the “Quantification of the effects of adaptation measures on spatiotemporal changes in flood and drought risk (EXTREME-ADAPT)”, an InnoPool project funded by the Helmholtz Association.

References

- Lars Borin, Rickard Domeij, Jens Edlund, and Markus Forsberg. 2023. *Language Report Swedish*, pages 219–222. Springer International Publishing, Cham.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. *Language models are Few-Shot learners*.
- Tais Maria Nunes Carvalho, Andreas Niekler, Christian Kuhlicke, Jakob Zscheischler, and Mariana Madruga de Brito. 2025. Global synthesis of peer-reviewed articles reveals blind spots in climate impacts research.
- Andry Castro, João Pinto, Luís Reino, Pavel Pipek, and César Capinha. 2024. *Large language models overcome the challenges of unstructured text data in ecology*. *Ecological Informatics*, 82:102742.
- Felicia Chiang, Omid Mazdiasni, and Amir AghaKouchak. 2021. Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature communications*, 12(1):2754.
- Mariana Madruga de Brito, Brielen Madureira, Taís Maria Nunes Carvalho, Damien Delforge, Aglaé Jézéquel, Murathan Kurfalı, Ni Li, Gabriele Messori, Joakim Nivre, Barbara Pernici, Niko Speybroeck, Stefano Terzi, Wim Thiery, Bram Valkenburg, Jingxian Wang, Shorouq Zahra, Jakob Zscheischler, and Jan Sodoge. 2026. *Assessing socio-economic climate impacts from text data*. *Preprint*, arXiv:2605.20793.

- Mariana Madruga de Brito, Jan Sodge, Heidi Kreibich, and Christian Kuhlicke. 2025. Comprehensive assessment of flood socioeconomic impacts through text-mining. *Water Resources Research*, 61(1):e2024WR037813.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. [Polyglot prompt: Multilingual multitask prompt training](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Eunmi Lee, Seunghui Choi, Sangeun Lee, Newsha K. Ajami, Hyeoncheol Yoon, Xiaogang He, and Jonghun Kam. 2026. [Spatiotemporal and economic impacts of media on water conservation during drought: An explainable ai approach](#). *Water Resources Research*, 62(6).
- Ni Li, Shorouq Zahra, Mariana Brito, Clare Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, Gabriele Messori, and Joakim Nivre. 2024. [Using LLMs to build a database of climate extreme impacts](#). In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 93–110, Bangkok, Thailand. Association for Computational Linguistics.
- Andreas Nicolaidis Lindqvist, Rickard Fornell, Thomas Prade, Linda Tufvesson, Sammar Khalil, and Birgit Kopainsky. 2021. Human-water dynamics and their role for seasonal water scarcity – a case study. *Water Resour. Manage.*, 35(10):3043–3061.
- Brielen Madureira, Mariana Madruga de Brito, and Andreas Niekler. 2026. Retrieving floods without floodlights: Topic models as binary classifiers for extreme climate events in german news. *arXiv preprint arXiv:2605.03450*.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. [gpt-oss-120b gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- OpenAI. 2026. [Gpt-5.5 system card - openai deployment safety hub](#).
- Kimberly J Quesnel and Newsha K Ajami. 2017. Changes in water consumption linked to heavy news media coverage of extreme climatic events. *Sci. Adv.*, 3(10):e1700784.
- Anand Rajaraman and Jeffrey David Ullman. 2011. *Data Mining*, page 1–17. Cambridge University Press.
- Michele Ronco, Luca Bandelli, Lorenzo Bertolini, Sergio Consoli, Damien Delforge, Alessio Spadaro, Marco Verile, and Christina Corbane. 2026. Disaster storylines and knowledge graphs from global news with large language models and Retrieval-Augmented generation. *Sci. Data*, 13(1).
- Jan Sodge, Christian Kuhlicke, and Mariana Madruga de Brito. 2023. Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather Clim. Extrem.*, 41(100574):100574.
- Jan Sodge, Christian Kuhlicke, Giuliano Di Baldassarre, Jan H Fleckenstein, Pia Ebeling, and Mariana Madruga de Brito. 2026. Emerging hotspots of groundwater conflicts during droughts in germany. *iScience*, 29(7).
- Statistikmyndigheten SCB. 2023. [Kvalitetsdeklaration tätorter](#).
- Claudia Teutschbein, Elise Jonsson, Andrijana Todorović, Faranak Tootoonchi, Elin Stenfors, and Thomas Grabs. 2023. [Future drought propagation through the water-energy-food-ecosystem nexus – a nordic perspective](#). *Journal of Hydrology*, 617:128963.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#).
- Javier Vela-Tambo, Jorge Gracia, and Fernando Dominguez-Castro. 2025. [CienLLM: Generative Climate-Impact Extraction from News Articles with Autoregressive LLMs](#).
- Jingxian Wang, Andrea Castelletti, Mariana Madruga De Brito, and Barbara Pernici. 2025. [Drought perceived impacts via text mining of social media](#). *Environmental Research: Water*, 1(4):045007.

Marthe Linda Kris Wens, Michael Hagenlocher, Anastasiya Shyrokyaya, Micha Werner, Andrea Toreti, Pedro Henrique Lima Alencar, Amobichukwu Chukwudi Amanambu, Jesse Eugene Bell, Won-Ho Nam, Sumiran Rastogi, Juan Rivera, Celia Ramos Sánchez, Matteo Zampieri, Anne-Sophie Sabino-Siemons, Kelly Smith, Faranak Tootoonchi, Patricia Trambauer, Michelle Theresia Helena van Vliet, Marit van Thiel, and 5 others. 2026. Ten key insights and gaps to inform drought risk research, policy and practice. *Nat. Water*, 4(6):711–721.

A Further details about the corpus development

Table 6 presents the keyword searches applied to select the articles in the Retriever media archive. To assess the relevance of the articles retrieved for each query, we examined the titles and headlines of the top 20 most relevant and the 20 least relevant articles. If a majority of articles were relevant in this subset, we included the full query in our search.

To filter out texts that did not mention municipality or urban area names, we used a list of official Swedish municipalities and the most recent list of urban areas from Statistics Sweden⁹. Urban areas are updated yearly, but we assume that year-to-year variations are negligible. Using regex patterns, the article’s main text was searched for municipality names in their nominative and genitive forms, including demonym suffixes. Urban areas were searched for only in their nominative form.

Figure 4 shows the distribution of cosine similarity values computed from TF-IDF representations of the articles after filtering out texts that did not contain municipality or urban area names. Most article pairs have low to moderate similarity (below 0.7), but there is a visible cluster of highly similar texts above 0.8–0.9. This distribution and the manual examination of text clusters confirmed that 0.8 can be used as a threshold value to identify duplicate and near-duplicate articles.

B Relevance classification prompt

Prompt 1: Relatedness classification

You will be given a news article in Swedish. Determine whether the article is relevant. An article is considered relevant if it contains information about:

- Water use restrictions or irrigation bans (e.g. bevattningsförbud) issued by a municipality or a water utility

⁹<https://www.scb.se/hitta-statistik/statistik-efter-amne/boende-bebyggelse-och-mark/bebyggelseomraden/tatorter-och-smaorter>

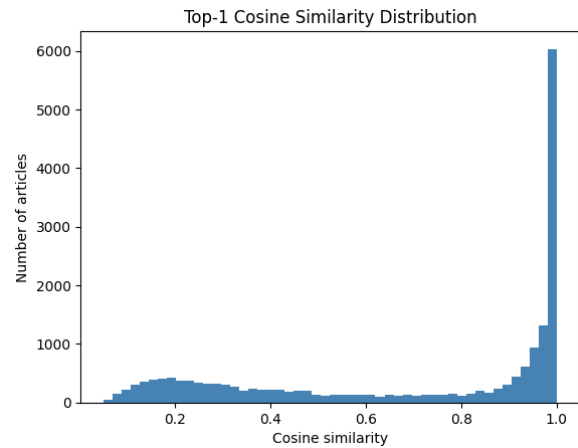


Figure 4: Cosine similarity distribution

- Official water saving recommendations
- Statements that no restrictions are needed

This information typically concerns municipal water management and often comes from a municipality or municipal water utility (VA-bolag), but the issuing authority may not always be explicitly stated in the text.

An article is NOT relevant if it only discusses:

- Water abstraction bans or permit restrictions issued by a regional authority (länsstyrelsen)
- Hypothetical bans that have not been formally decided
- Water saving technologies or tips for households
- General advice on saving water without official recommendations
- Economic or political discussions about water
- Water pricing or taxation
- Environmental issues without specific restrictions or recommendations

Even short articles can be relevant. If municipal restrictions or recommendations are mentioned, even briefly, the article is still considered relevant if it provides concrete information such as the location, timing, or implementation of the measure. Return True if the article is relevant. Return False if the article is not relevant.

REMEMBER: All texts provided are articles, even if they look like briefs, summaries, metadata, article lists. Ignore non-relevant content, only focus on relevant content.

	Target	Query	Articles
1	Water restrictions	bevattningsförbud* OR bevattningsstopp OR ((förbud* OR förbj* stopp* OR begräns) NEAR/2 (bevattn* OR vattna)) OR vattenrestriktion* OR bevattningsrestriktion* OR vattningsrestriktion* OR (restriktion* NEAR/2 vattenanvändning*) OR (totalförbud near/2 vatten)	12154
2	Save water	(spara OR bespara OR sparsam*) NEAR/3 (vatten OR vattnet OR dricksvatten OR dricksvattnet) OR vattenbesparing	6672
3	Save water (synonyms and expressions)	((minska OR ''dra ner'' OR ''skära ner'' OR begräns* OR anpassa OR ''håll ner'' OR ''använda mindre'' OR försiktig* OR sparsam*) NEAR/2 (vattenförbrukning* OR vattenanvänd*)) OR ((hushålla OR ''inte slösa'' OR ''slösa inte'' OR ransonera OR snåla OR ''tappa upp'') NEAR/2 (vatten OR vattnet OR dricksvatten OR dricksvattnet)) OR ((uppmana* OR manar OR varnar OR varna OR vädja*) NEAR/5 (bevattn* OR vattenförbrukning*))	3802
4	Water shortages	kommun* NEAR/5 (vattenbrist OR ''brist på vatten'' OR ''dricksvattenbrist'')	976

Table 6: Search queries used to construct the document corpus and total articles retrieved on 25/02/2026

C Data extraction prompt example

Prompt 2: Extract the municipality

Identify the Swedish municipality or municipalities (kommun/kommuner) where the extracted measure_type applies. An article may describe multiple measures and multiple municipalities. For each municipality, create a separate output entry associated with the relevant measure_type.

- If the same measure_type applies to multiple municipalities, return one entry per municipality (do NOT group municipalities together).
- If different municipalities are associated with different measure_types, assign each municipality to the correct measure_type and return separate entries.

Important:

- The municipalities must be official Swedish municipalities (kommuner).
- Only include municipalities that are explicitly mentioned in the text verbatim and directly linked to the measure.
- When returning municipality names, output only the municipality name itself (e.g. "Uppsala", even if the text says "Uppsala kommun", or "Region Uppsala").
- Do not normalize or translate municipality names.

- You may only infer the municipality if it can be identified with high certainty from the article source (e.g. "Uppsala kommun", "24 Kalmar") or context (i.e., municipality inhabitants).
- Do not include municipalities that are not affected by, subject to, or clearly linked to a water measure.
- If there is any ambiguity, do not return a municipality.

D Deterministic output for OpenAI models

Even though only low or no reasoning levels are recommended by OpenAI for deterministic workflows (https://developers.openai.com/cookbook/examples/gpt-5/gpt-5_new_params_and_tools#42-takeaways), OpenAI also uses prompt caching for the GPT-5 series, among others, making it difficult to determine if multiple runs are producing identical results or simply reliant on an internal caching mechanism to reduce latency (<https://developers.openai.com/api/docs/guides/prompt-caching>).

E Additional Results

This section contains additional results, including development set results for the preliminary experiment in Table 7, additional evaluation scores for the cause category classification on the development set in Table 8, comparison of performance on the test-set across different GPT-5 series models in Table 9, and the performance on the test set across ascending temperatures for the Llama3.3 model compared to GPT-5.5 performance across different reasoning levels in Table 10.

F Examples of errors by GPT-5.5 and Llama 3.3

This section presents article text excerpts translated into English, together with the corresponding annotations and selected examples of model errors for different extraction fields. The examples of errors are presented in Table 11 (GPT-5.5) and Table 12 (Llama 3.3).

Metric / Set	GPT-5.5		Llama3.3 (70B)		Deepseek R1 (70B)		GPT-OSS (120B)		Baseline	
	All	Match	All	Match	All	Match	All	Match	All	Match
article date	0.18	0.00	0.31	0.02	0.31	0.00	0.45	0.00	0.30	0.02
municipality	0.24	0.07	0.49	0.27	0.47	0.23	0.61	0.30	0.67	0.54
detailed location	0.26	0.10	0.87	0.82	0.87	0.81	0.91	0.84	0.72	0.60
location type	0.25	0.08	0.60	0.43	0.54	0.34	0.80	0.63	0.79	0.70
measure type	0.26	0.10	0.72	0.60	0.64	0.48	0.84	0.71	0.81	0.73
measure date	0.35	0.27	0.64	0.62	0.62	0.61	0.54	0.62	0.36	0.24
ban imposed date	0.06	0.04	0.48	0.54	0.40	0.54	0.48	0.51	0.25	0.30
ban lifted date	0.02	0.03	0.04	0.06	0.10	0.14	0.04	0.06	0.05	0.06
cause category	0.34	0.20	0.88	0.85	0.82	0.74	0.90	0.82	0.88	0.84
avg. score	0.22	0.10	0.56	0.47	0.53	0.43	0.62	0.50	0.54	0.45
coverage	96.42%		100%		100%		96.43%		100%	

Table 7: Comparison of extraction performance on the **development set**. Temperatures and reasoning levels are set to 0.0 and “none”, respectively. The results under the “Match” columns shows the result when extra or missed entries found by the LLMs are ignored. Coverage is the percentage of articles the models produced at least one entry for.

		GPT-5.5			Llama 3.3			Support
Label		Precision	Recall	F1	Precision	Recall	F1	
0	High water consumption	0.762	0.800	0.780	0.600	0.150	0.240	20
1	Lack of rain	0.806	0.833	0.820	0.000	0.000	0.000	30
2	Low groundwater levels	0.944	0.850	0.895	1.000	0.150	0.261	20
3	Low surface water levels	0.615	0.800	0.696	0.000	0.000	0.000	10
4	Low water levels unspecified	—	—	—	—	—	—	0
5	NA	0.833	0.714	0.769	0.000	0.000	0.000	21
6	Other	0.500	0.500	0.500	0.000	0.000	0.000	4
7	Power outage	—	—	—	—	—	—	0
8	Precaution	0.000	0.000	0.000	0.000	0.000	0.000	2
9	Warm temperature	1.000	1.000	1.000	0.000	0.000	0.000	25
10	Water repairs	1.000	1.000	1.000	0.500	0.167	0.250	6
11	Water shortage	0.519	0.824	0.636	0.222	0.706	0.338	17
12	Waterworks at capacity	0.800	0.889	0.842	0.000	0.000	0.000	9

Table 8: Precision, recall, and F1 scores for multi-label classification of the cause category for the **development set**. The support column represents the number of examples available in the gold annotations for each label. Labels with rows marked as dashes (‘—’) do not appear in the development set gold annotations.

Metric / Set	GPT-5.5		GPT-5.4		GPT-5.2		GPT-5.4-mini	
	All	Match	All	Match	All	Match	All	Match
article date	0.13	0.00	0.15	0.00	0.14	0.00	0.22	0.00
municipality	0.17	0.04	0.21	0.08	0.22	0.09	0.28	0.07
detailed location	0.19	0.07	0.22	0.09	0.23	0.11	0.32	0.13
location type	0.16	0.03	0.20	0.07	0.19	0.06	0.27	0.06
measure type	0.27	0.16	0.25	0.13	0.28	0.16	0.36	0.18
measure date	0.30	0.26	0.24	0.19	0.30	0.29	0.43	0.38
ban imposed date	0.11	0.10	0.13	0.12	0.14	0.13	0.16	0.16
ban lifted date	0.03	0.02	0.04	0.02	0.07	0.06	0.03	0.02
cause category	0.32	0.22	0.38	0.28	0.37	0.27	0.49	0.35
avg. score	0.19	0.10	0.20	0.11	0.21	0.13	0.28	0.15
coverage	91.56%		91.57%		92.77%		81.33%	

Table 9: Comparison of extraction performance on the **test set** for different GPT-5 models at reasoning level “none”. The results under the “Match” columns shows the result when extra or missed entries found by the LLMs are ignored. Coverage is the percentage of articles for which the models produced at least one entry. The lowest error rates are shown in **boldface**.

set	Llama 3.3 (70B) temperature						GPT-5.5 reasoning level			
	0	0.25	0.5	0.75	1.0	1.5	none	low	med.	high
article date	0.31	0.33±0.00	0.36±0.02	0.37±0.02	0.40±0.02	0.43±0.01	0.13	0.13	0.15	0.13
municipality	0.49	0.53±0.01	0.58±0.04	0.57±0.01	0.60±0.02	0.62±0.01	0.17	0.17	0.19	0.16
detailed location	0.87	0.87±0.00	0.87±0.03	0.84±0.03	0.83±0.03	0.85±0.02	0.19	0.19	0.20	0.17
location type	0.60	0.60±0.01	0.60±0.03	0.60±0.01	0.64±0.01	0.63±0.02	0.16	0.16	0.18	0.16
measure type	0.72	0.75±0.02	0.78±0.01	0.76±0.02	0.80±0.02	0.80±0.01	0.27	0.24	0.26	0.27
measure date	0.64	0.54±0.03	0.55±0.02	0.54±0.03	0.51±0.01	0.52±0.04	0.30	0.30	0.30	0.25
ban imposed date	0.48	0.43±0.01	0.41±0.02	0.37±0.02	0.35±0.02	0.34±0.03	0.11	0.14	0.14	0.14
ban lifted date	0.04	0.10±0.02	0.12±0.01	0.10±0.01	0.12±0.02	0.12±0.01	0.03	0.03	0.02	0.03
cause category	0.88	0.91±0.00	0.90±0.01	0.89±0.01	0.90±0.01	0.91±0.02	0.32	0.35	0.36	0.33
average score	0.56	0.56±0.00	0.57±0.01	0.56±0.01	0.57±0.00	0.58±0.01	0.19	0.19	0.20	0.18

Table 10: Performance on all entries for Llama3.3 at different temperatures and GPT-5.5 at different reasoning levels on the **test set**. Evaluation is performed on all articles. Standard deviation is shown for Llama3.3 for non-deterministic runs with temperature > 0, representing the variance between three separate runs. GPT-5 runs at different reasoning levels were performed only once.

Error	Text excerpt (translated)	True	Predicted
Measure type	"In several municipalities, a watering ban or a partial watering ban has already been imposed. For example Flen, [...]"	ban continues	ban imposed
Measure date	"Date: 2020-06-17", "Since earlier this year there has been a ban in Båstad and Osby municipalities"	Not mentioned	2020-06-17
Cause category	"[...] avoid using a hose for watering and please be mindful of your consumption.", "While water supplies in Bollnäs and Ovanåker remain plentiful for now, residents are still urged to save water." "That recommendation was introduced [...], when temperatures began to rise..."	Warm temperature, Precaution	Warm temperature, High water consumption
Cause category	"Water scarcity is spreading across the country, and Karlshamn is no exception. Alarming low water levels in Lake Mien [...].", "... the precipitation [this morning] could not have come at a better time."	Lack of rain, Low surface water levels, Water shortage	Low surface water levels, Water shortage
Number of entries	"At present, there are bans in 12 municipalities including Östhammar in Uppsala county.", "In Enköping we have low water levels and encourage everyone to save water."	2 separate entries: Uppsala, Enköping	Only 1 entry reported: Enköping

Table 11: Examples of errors by GPT-5.5 (at reasoning effort 'none')

Error	Text excerpt (translated)	True	Predicted
Municipality and number of entries	"There is a ban in Högsby and Hultsfred municipalities [...]"	"Högsby" (one entry), "Hultsfred" (second entry)	"Högsby and Hultsfred" (single entry)
Detailed location	"The water levels in our groundwater aquifers are still low..."	"Not mentioned"	"our groundwater aquifers" (returned in Swedish)
Location type	"Gotland", "Acute water shortage [...] in Visby"	Part of the municipality	Entire municipality
Measure date	"Ban in Flöghult municipality during the summer"	"Not mentioned"	"2021 summer"
Ban imposed date	"Date: 2019-06-13 [...] The ban is from July 4th to August 6th [...]"	2019-07-04	2019-06-13
Ban lifted date	"the ban remains until further notice"	"Not mentioned"	"until further notice" (returned in Swedish)
Cause category	"[...] the cause of the watering ban is the low groundwater levels".	Low groundwater levels	Water shortage

Table 12: Examples of errors by Llama 3.3 (using deterministic output)