Inductive Categorization for Conceptual Analysis with LLMs: A Case Study from the Humanitarian Encyclopedia

Loryn Isaacs[1], Santiago Chambó[1], Pilar León-Araúz[1]

[1] University of Granada, Department of Translation and Interpreting,
Puentezuelas, 55, 18071, Granada, Spain
E-mail: lisaacs@ugr.es, santiagochambo@ugr.es, pleon@ugr.es

Abstract

Corpus-based conceptual analysis for the Humanitarian Encyclopedia (HE) grapples with vast amounts of lexical data, where semantic triples are grounded in lexical units ranging from single words to full sentences. To enhance conceptual representations, categorizing these units into semantic groups is crucial. While traditional inductive qualitative analysis is labor-intensive, researchers are now replicating these methods using LLM-assisted workflows. Following this trend, our paper presents an observational study with data on the concept of FORCED DISPLACEMENT extracted from humanitarian texts. In this initial assessment, we test LLM inductive categorization using local Magistral and DeepSeek R1 models against manual categorization. To assess baseline similarities, we provide models with minimal, zero-shot instruction, while also requiring structured outputs (categories containing members identified previously by annotators). We evaluate model fitness to complete the task using a similarity score and qualitative categorization behaviors. Models had low overall similarity scores when given little instruction and hundreds of spans to classify in one batch, consistently omitting spans despite being prompted not to do so. The results underscore the complexity of categorizing data for a single, domain-specific concept. We discuss several aspects of model behavior and steps for improving similarity to human annotation.

Keywords: inductive categorization; humanitarian domain; large language model; structured output; conceptual analysis

1. Introduction

The Humanitarian Encyclopedia (HE) is a project aimed at producing corpus-informed descriptions of 129 key specialized concepts by complementing traditional entry authoring with corpus-based conceptual analyses, which are provided by a team of linguists (Odlum & Chambó, 2022). Humanitarian concepts are anticipated to exhibit a considerable degree of conceptual variation (on conceptual variation, see León-Araúz, 2017 and Hampton, 2020) due to the recent professionalization of the sector (Eberwein & Saurugger, 2013), diverse organization epistemologies (Dauvin & Siméant-Germanos, 2002; Sezgin & Dijkzeul, 2015), and the relatively nascent development of Humanitarian Studies as an academic field (Gorin, 2024). Our approach to conceptual variation consists of quantifying the differing semantic triples in which humanitarian concepts are found (e.g. AID DEPENDENCE *caused_by* UNEMPLOYMENT; AID DEPENDENCE *caused_by*

DISPLACEMENT) and disaggregating them according to corpus metadata (based on organization type, country or year of publication, etc.) to draw comparisons.

The HE is building on Frame-based Terminology (FBT; Faber, 2015, 2022), a pre-existing method for conceptual analysis with the potential to fulfill the HE's requirements by integrating techniques from other disciplines. Following Kantner and Overbeck's recommendations (2020), part of our research efforts is currently centered on incorporating qualitative methods into a workflow that would enable conceptual analysts to subsume large amounts of diverse lexical data into manageable semantic triples, ensuring the traceability and transparency of modeling decisions.

In this study, we propose accelerating an inductive categorization process, intended for integration into a workflow for the HE, by leveraging large language models (LLMs) to create categories representing concepts and instances within the second argument position of semantic triples. To assess feasibility, we conducted an observational and comparative study with data on the concept of FORCED DISPLACEMENT and its causes. We compared the outputs of an LLM-driven inductive categorization process with locally run versions of Magistral and Deepseek R1 against manual categorization by two analysts. Manually generated categories and those of the LLMs were assessed with a score derived from Jaccard similarity coefficient. A quantitative analysis of the models' capacity to complete the task and a qualitative analysis of the categorization results were also conducted.

The remainder of this article is structured as follows. Section 2 discusses the challenges of adapting FBT for HE purposes and reviews previous works on LLM-driven inductive categorization. Section 3 details the study's materials and methods. Section 4 presents the results, and Section 5 concludes the paper, outlining future research directions.

2. Expanding FBT for the Humanitarian Domain through Inductive Categorization with LLMs

A key difference between conceptual analysis for the HE and previous FBT projects lies in the volume and nature of the lexical data utilized. FBT projects primarily focus on producing conceptual descriptions based on a specialized lexicon. Consequently, conceptual characteristics that are not expressed in truly specialized terms are likely to be excluded from representations (see example below). Reasonably, the goal is to build resources that are readily accessible for audiences with domain-specific lexical and knowledge needs, such as translators, interpreters and technical writers. While corpus evidence in the form of knowledge-rich contexts (see Condamines, 2022) is considered essential, expert validation also plays a crucial role. Previous FBT projects have employed corpus metadata to explore conceptual variation, but they focused on generating flexible definitions and conceptual networks for subdomains of a specialized field (León-Araúz et al. 2013; San Martín & León-Araúz, 2013; San Martín, 2022a,

2022b). Representing conceptual variation is limited to the most frequent triples found in the corpus based on subdomain prototypicality (León-Araúz, 2017) without triple quantification or substantiation.

In contrast, the HE aims to derive conceptual descriptions from lexical data irrespective of their terminological value. While analyses start from a specialized term in the humanitarian domain (i.e., 'forced displacement'), second arguments in triples are substantiated with lexical data regardless of its domain specialization. The HE's broader approach to conceptual variation necessitates differentiating prominent and marginal characteristics, which inherently relies on a quantitative dimension based on corpus metrics (Kantner & Overbeck, 2020, p. 186). The target audience also differs significantly; the end-users of the conceptual analyses are entry authors, who are interested in comparing their conceptual descriptions with findings from corpus analysis. The focus here is on providing a conceptual model that is representative of broader humanitarian discourse and that enables disaggregation for detecting conceptual variation.

Let us consider the concept of FORCED DISPLACEMENT. A traditional FBT analyst would typically identify terms such as 'genocide' and 'evacuation order' as causally linked to FORCED DISPLACEMENT, resulting in triples like FORCED DISPLACEMENT *caused_by* GENOCIDE and FORCED DISPLACEMENT *caused_by* EVACUATION ORDER. However, an HE analyst would additionally consider other items for conceptual modeling. Examples include 'exterminate people in large numbers' or 'the authorities ordered the evacuation of everyone in the area'. While these items happen to align with the triples modeled above, other valuable cases lacking concise and nominal counterparts run the risk of being overlooked in a traditional FBT analysis. Datasets in HE conceptual analyses comprise diverse lexical units ranging from single-word and multi-word expressions to clauses, and occasionally, full sentences. These datasets exhibit a long-tailed distribution: a few low-frequency n-grams account for modest absolute frequencies at the head, while the tail comprises a diverse majority of one-hit n-grams. Examples of such distributions can be found in Chambó & León-Araúz (2023).

For the purposes of the HE, these unique units must be considered and processed after corpus extraction. Low-frequency units have the potential to encode valuable information that can be abstracted through semantic grouping to substantiate semantic triples. Failing to treat these items has two drawbacks. On the one hand, conceptual characteristics conveyed through diverse natural language strategies would be overlooked. On the other, reports would be overly granular, difficult to understand and would fail to communicate conceptual variation effectively. Figure 1 situates inductive categorization within the workflow currently being developed for the HE.

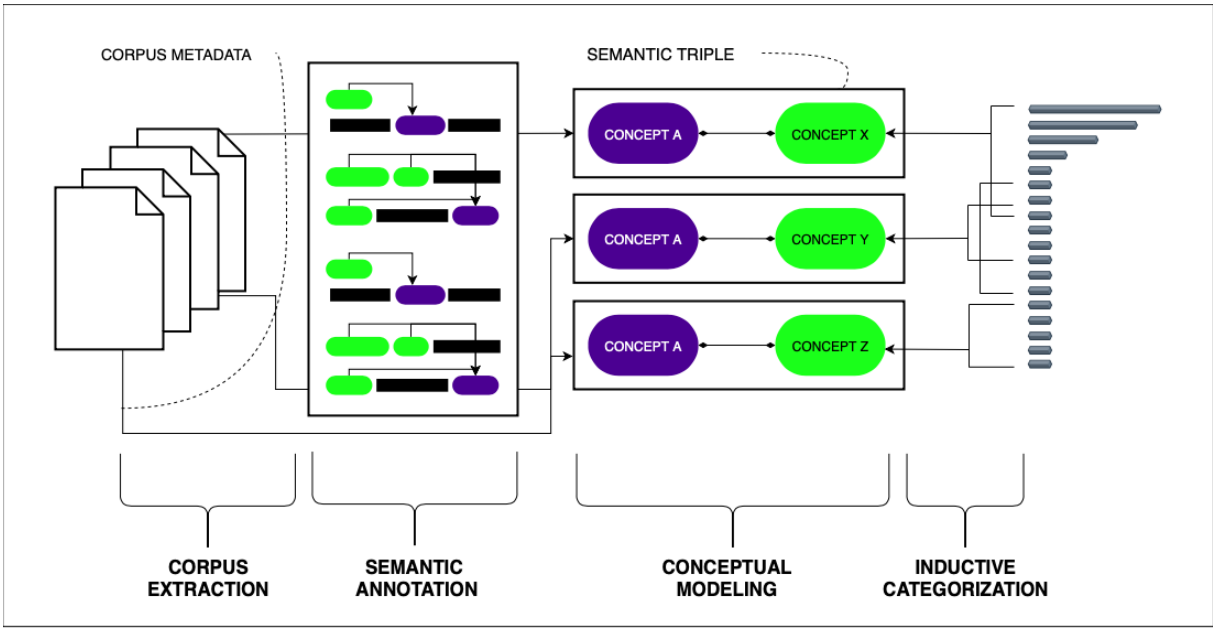

Figure 1: Conceptual analysis workflow for the HE

Classifying lexical items into manageable categories is the backbone of several qualitative research methods such as grounded theory, thematic analysis and content analysis (Babchuk & Boswell, 2023; Bingham, 2023; Naeem et al., 2023). Typically, teams of coders develop initial categories independently from source documents, refining them through collaboration. Given the labor-intensive nature of this activity, several LLM-assisted inductive coding frameworks have been proposed for applications in various disciplines.

In empirical legal studies, Drapal et al. (2023) developed a framework to support initial coding and category building using OpenAI's GPT-4. They assessed the model's performance in autonomously generating codes against the same task with expert feedback. Further categorization of initial codes was evaluated by comparing a zero-shot strategy with a set of manually derived categories. Performance was assessed through domain expert input, with 66% acceptability reported for initial coding and 82% for categorization.

De Paoli (2024) experimented with temperature values for GPT-3.5 and compared results with two thematic analysis studies, using temperature values of 0.5 and 1, while also constraining the number of categories to be developed by the LLM. Considerable overlap was found between expert- and LLM-generated categories. While most categories persisted at higher temperatures, the model either overlooked categories from one study or generated richer ones from another. De Paoli highlights the importance of establishing best practices for temperature settings, suggesting that while a temperature of 0 ensures result reproducibility, higher values may also contribute to validity.

Arlinghaus et al. (2024) introduced their LLM-assisted Inductive Categorization (LAIC) method, providing user-friendly open-access materials for researchers. A key contribution is the emphasis on incorporating an 'audit trail' into the method's design, a crucial element often overlooked in some existing literature. LAIC generates categories by repeatedly prompting a GPT model of choice via OpenAI's API with the same prompt with 10 runs at a temperature of 0 to ensure result dependability. Categories are collected and ordered by frequency counts in a spreadsheet. Manual intervention is required to select categories by identifying similar outputs and prioritizing those with higher frequencies. In the final step, the model is prompted to assign passages to one of the user-selected, LLM-generated codes.

In education research, Barany et al. (2024) compared categories developed using GPT 4 via chat interface from four distinct LLM integration workflows: no LLM use, LLM use only in category development, LLM use only in category refinement, and full LLM use. LLM-assisted configurations generated valuable categories overlooked by both manual and LLM-only workflows.

Katz et al. (2024) developed the Extract, Embed, Cluster, and Summarize (EECS) workflow, a fully automated pipeline mirroring traditional qualitative research. It performs initial coding with a locally run dolphin fine-tuned version of Mistral-7b. Initial codes are then embedded into a vector space and algorithmically clustered. A cluster representative is automatically selected by comparing each initial code with a concatenated string including all codes, and an initial set of categories is created, having the LLM generate a summary. Subsequent refinement is conducted with Retrieval-Augmented Generation, comparing a small selection of categories via cosine similarity rather than feeding the entire set into the LLM.

In a study with healthcare qualitative interviews, Mathis et al. (2024) employed LLaMA-2-70B-Instruct on a local server. To assess the similarity between manual and LLM-generated categories, they used the Sentence-T5-xxl model, which converts categories into vector spaces for comparison via cosine similarity. Initial categories were generated by prompting the LLM with the source text. A key distinction of this study lies in its application of chain-of-thought and reflection prompting strategies during the category merging and refining phases, alongside an explicit evaluation phase where the LLM is prompted to identify flaws and faulty logic.

Finally, Bakharia et al. (2025) compared multiple LLMs against traditional topic modeling techniques, with a strong emphasis on traceability and verification against hallucinations. A crucial verification phase was introduced between LLM initial coding and category merging. Intermediate categories were produced by clustering with non-negative matrix factorization. LLMs discarded approximately one-third of the initial codes provided. Subsequently, the LLM was employed to validate and subsume clusters, with instructions to support categorization decisions with explicit reasoning. Smaller open-source models struggled to substantiate codes with quotes, while larger

proprietary LLMs performed better, generating between 7.7% and 8.3% invalid substantiating quotes for categories.

All the workflows reviewed here, except for Katz et al. (2024), incorporate some form of manual intervention to guide and validate outputs during intermediate phases of LLM-assisted inductive categorization. Findings from Barany et al. (2024) suggest that human supervision in such tasks has the potential to reveal valid categories overlooked in purely manual or LLM-generated categorization. Aside from faster completion, additional reported benefits include leveraging LLM output to aid category refinement in discussions or to serve as another coder (De Paoli, 2024; Gao et al., 2024).

A significant obstacle, however, is the reported poor performance of LLMs when provided with a full list of initial codes. This is critical because the initial set of codes largely determines the categories formed through a "constant comparative method" (Bingham, 2023, p. 6). Consequently, some studies advocate for reducing the number of initial codes via semantic clustering before prompting the LLM. For instance, Katz et al. (2024) employ an ingenious approach where initial codes are iteratively added to a final set of categories by comparing them with a selection of the most semantically similar, previously added codes. A new category is established if semantic similarity is low.

A further challenge is guaranteeing the link between corpus evidence and a given category. De Paoli (2024) and Bakharia et al. (2025) reported on hallucinations that render traceability difficult. In conceptual analysis for the HE, analysts initially identify textual fragments semantically linked to a concept under study before categorization. Maintaining the connection between a given textual fragment and its assigned category is paramount. While Arlinghaus et al. (2024) address this by first developing categories and then deductively coding the source text, this approach overlooks multi-category coding, a common practice in qualitative research.

In this study, we explored an LLM-assisted inductive categorization technique with a view to incorporating it into our conceptual analysis workflow for the HE. Key objectives included evaluating the performance of LLMs when fully exposed to an entire list of items for categorization, testing smaller locally run models, and determining the degree of overlap with manual categorization. Section 3 details the materials and methods employed.

3. Methods and Materials

3.1 Data Collection and Annotation

A random sample of 1,000 concordances containing 'forced displacement' were selected from humanitarian reports on Reliefweb[1], a communication hub for humanitarian organizations managed by the United Nations Office for the Coordination of Humanitarian Affairs. Sample selection was conducted with a preexisting corpus of English Reliefweb (Isaacs et al., 2024) with 2.2 billion tokens from 933,928 documents and dates from 2000 to 2024. The sample, obtained with a local instance of the corpus management software NoSketch Engine (Kilgarriff et al., 2014; Rychlỳ, 2007), represented 6.39% of the occurrences of 'forced displacement' (out of 15,631 total).

The data were annotated with Doccano (Nakayama et al., 2018), where causal semantic relations were detected with 'forced displacement', including both its causes and consequences. Two human annotators with experience in conceptual analysis labeled triples containing *caused_by* relations, showing the direction of causality with 'forced displacement' and any number of spans within a sentence (see example in Figure 2). Annotating other related spans and triples (i.e. *located_at*), as well as providing comments, was optional.

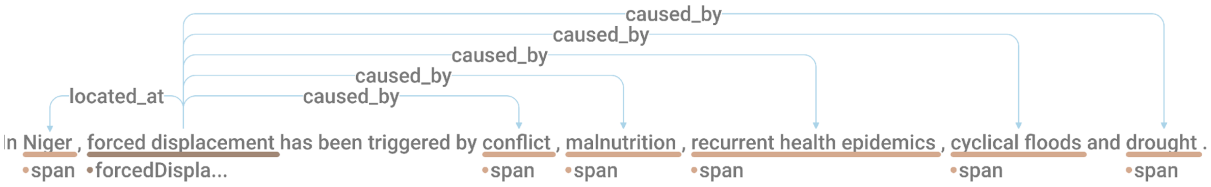

Figure 2: Human annotation of causality triples

The dataset consisted of 274 spans labeled as causes of 'forced displacement'. Two annotators categorized these spans inductively and independently, repeating the categorization task twice in order to record both their initial impressions and deliberated conclusions. In each pass they applied up to four categories to each span, and they later cooperatively developed a final harmonized classification with 34 categories (this experiment's gold standard). Annotators left 18 spans uncategorized for being too generic or redundant, such as 'complex situations', 'crisis', 'drivers', 'immediate and root causes', 'numerous interlinked causes.' Table 1 details the number of spans per category and a span example.

[1] https://reliefweb.int

Category label	N	Span example	Category label	N	Span example
Non-state armed actor	48	FARC	Inequality	4	growing economic disparities
Conflict	38	armed dispute	Political instability	4	chronic political instability
Violence	32	some form of direct or indirect violence or factors known as push factors	Weaponry and explosive hazards	5	UXO
State	29	Government of the Sudan	Environmental disruption	3	Environmental degradation
State armed actor	20	Israeli military	Food insecurity	3	physical and economic barriers to accessing food
Military operation	14	ground operations	Forced recruitment	3	forced recruitment of children and youth
Rights violations	14	human rights abuses	House demolitions	3	home demolitions
Displacement order	13	evacuation orders issued by the Israeli forces	Gender-based violence	2	gender-based violence
Resource extraction, development projects and land exploitation	12	oil and gas projects	Health issue	2	malnutrition
Coercive and oppressive practices	10	expulsion or other coercive acts	Impediments to accessing humanitarian services	2	decreasing humanitarian access
Crime	9	illicit crops	Lack of international action	2	international community's near silence and lack of action
Natural hazard	9	impacts of floods	Population movements and	2	ethnic cleansing

			demographic change		
Climate change	7	aggravating effects of climate change	Poverty	2	extreme poverty
Economic disruption	6	loss of livelihood	Vulnerability	2	related vulnerability
Expansionism	6	Policies on settlement expansion	Drought	1	drought
Freedom of movement restrictions	6	Restriction of the freedom of movement by the Wall	Intergenerational impact	1	… intergenerational impact on the Palestinian society…
Insecurity	6	precarious security situation	Limited access to basic services	1	lack of access to basic services (such as education and health)

Table 1: Categories developed manually

3.2 Language Model Framework and Configuration

To orchestrate the use of language models for linguistic annotation tasks, we developed a command line interface (CLI) written in Python and available on GitHub[2], relying on the language model framework LangChain (Chase, 2022), along with Ollama[3] and HuggingFace[4] integrations. To allow a high degree of freedom in model output and produce baseline results, we opted for a zero-shot modality, where models were only given a descriptive yet concise summary of the task, without specifying the target domain or research context. Two recent language models/families with permissive licenses available on Ollama were selected. Magistral Small 1.0, with 24 billion parameters (Mistral-AI et al., 2025), was selected as the primary model for evaluation, with the DeepSeek R1 family (DeepSeek-AI, 2025) utilized for comparison purposes, including its versions with 8, 32 and 70 billion parameters.

Recent language models, including those selected here, can be configured to produce a structured output to provide desired fields and their values rather than a single text block requiring further parsing. We developed a structured output amounting to a dictionary of categories with lists of members, where members can only include those

[2] https://github.com/engisalor/lmf
[3] https://ollama.com
[4] https://huggingface.co

previously identified manually. Requiring models to produce the same members as human annotators prevented hallucinations and orthographic modifications during inference, which simplified evaluation. However, fully evaluating the impact of the output structure's definition remained a task for the future.

The following system prompt (1) was designed to provide a minimal set of instructions. A human prompt was then appended to each run, specifying the input format of the members to be categorized. Data were supplied in four text formats (lines, CSV rows with index, JSON dictionary and Python list) to ensure the evaluation was not impacted by variation in how models parse formats. The default configuration preserved most of LangChain's defaults for the ChatOllama class as of version 0.3, except for the use of recommended default parameters for Magistral as specified on HuggingFace: a temperature of 0.7 (default 0.8) and a top p of 0.95 (default 0.9).

(1) You are an expert text annotator working on a content categorization project. Group the following phrases into several categories, using as many specific yet distinct categories as needed. Give each category a unique label that describes its members' shared meaning. Make sure to include every phrase in at least one category, never omitting input phrases. If appropriate, place a phrase in multiple categories, up to four.

3.3 Quantitative Evaluation Metrics

Magistral (24B) and three sizes of DeepSeek R1 models (8B, 32B, 70B) were sent the prompt templates with 274 spans (causes of 'forced displacement') in the four input formats. Ten runs were executed for each model, amounting to 40 categorization results per model. The degree of category similarity was computed as follows. Every combination of categories for the gold standard and a model output was compared to identify their degree of overlap. The number of unique members and shared members per category pairing were used to calculate a shared similarity percentage. The lists of members for each annotator were also aligned to compute Jaccard similarity. Table 2 shows results from several category pairs to exemplify the procedure.

Category labels		**N members**				**Similarity**	
X (gold std.)	**Y (model)**	**X**	**Y**	**Total**	**Shared**	**Shared %**	**Jaccard**
Lack of international action	international relations and complicity	2	2	2	2	100	1.0
Health issue	health and epidemics	2	3	3	2	66.67	0.5
Expansionism	settler violence and policies	6	5	7	4	57.14	0.4
Food insecurity	health and epidemics	3	3	4	2	50	0.33
Inequality	social and economic inequality	4	2	4	2	50	0.33

Table 2: Example of category pair similarity scores

As the similarity measures above only consider one pair of categories for annotators X and Y, we calculated annotator-level similarity by taking the sum of Jaccard similarity scores for every category combination for X and Y and dividing it by the larger number of categories identified among the annotators (20 if X had 20 categories and Y 10). This produced a relative similarity score, which was then normalized to values between 0 and 1, where 1 represents the gold standard's relative Jaccard similarity with itself and 0 complete dissimilarity. The model results with the highest scores were taken to be the most similar to the gold standard in terms of category membership (ignoring category label), barring further inspection of LLM errors.

The following section provides a quantitative description of the categorization results. Section 4.1 describes how models replicated the task at a high level (their ability to complete the task), Section 4.2 offers a quantitative measure of model categorization against the gold standard (the unified categories agreed upon by human annotators), and Section 4.3 compares the semantic content of LLM-generated and manually generated categories.

4. Results and Discussion

4.1 Language Model Fitness for Zero-shot Categorization

Table 3 shows the degree to which models categorized the 274 spans and how often spans appeared in multiple categories. The 8 billion parameter DeepSeek R1 model on average categorized about 10% of spans, omitting 90% from its structured output, while the 70B model had the highest average categorization rate (57%). Magistral Small's average results were positioned between the 32B and 70B DeepSeek models, with a propensity for multiple categorization similar to 70B. DeepSeek R1 70B was the only model to categorize 100% of spans at least once, whereas every model except Magistral produced results that failed the categorization task (0%). The average number of multi-category spans for the larger models was similar to the gold standard (32), although with quite different minimums and maximums (0 for all models, up to 355 for Magistral).

	Mean		
deepseek70	156.32	57.05	45.88
magistral24	119.48	43.6	45.98
deepseek32	106.22	38.77	32.28
deepseek8	10.18	3.71	3.65
	Maximum		
deepseek70	274	100	210
deepseek32	249	90.88	257
magistral24	223	81.39	355
deepseek8	78	28.47	62
	Minimum		
magistral24	16	5.84	0
deepseek32	0	0	0
deepseek70	0	0	0
deepseek8	0	0	0

Table 3: Degree of span categorization

While models were occasionally capable of categorizing all or most of the spans, the number of categories often had an inverse relationship with the percentage of spans classified. Table 4 shows the average of the top ten results in terms of categorization percentage per model, along with the number of categories and largest category size. While the percentage of spans successfully categorized was high, the number of categories was often low (few categories contain most spans). For example, the DeepSeek R1 70B model's top ten results by categorization percentage omitted only 10% of the spans, but also generated just 5 categories, wherein almost all spans (244.5 on average) were put in the same category. This is also apparent in Table 5, which shows the top ten results by categorization percentage; the largest category in every case has at least 221 members. In comparison, the gold standard had 34 categories and an average of 9.42 members per category, up to 48 ('non-state armed actor').

Model	**Categorized %**	**N Categories**	**Largest Category**
deepseek32	64.12	6.8	130.2
deepseek70	90.07	5	244.5
deepseek8	10.15	19.2	5.7
magistral24	64.70	7.9	130.5

Table 4: Averages for highest 10 categorization % results

Model	Categorized %	N Categories	Largest Category
deepseek70	100	2	274
deepseek70	100	9	274
deepseek70	100	1	274
deepseek70	99.64	1	273
deepseek70	98.91	2	271
deepseek70	95.99	16	263
deepseek32	90.88	1	249
deepseek70	83.21	1	228
magistral24	81.39	9	223
magistral24	80.66	1	221

Table 5: Top 10 categorization % results and category sizes

Overall, we observed a dynamic where models between 24 and 70 billion parameters eventually succeeded in classifying most of the spans, but when most spans were included the quality of categorization was worse, with almost all spans put in a single category. All models also had runs with poor, unusable results, omitting all or almost all spans. With poor results at both extremes of categorization percentage, quantitatively identifying the highest quality categories required further inspection.

4.2 Similarity against Manual Categorization

The relative, normalized Jaccard similarity score described in Section 3.3 was utilized to identify the most similar categorization results between models, human annotators (users in Figure 3), and the gold standard. This measurement served as a global, quantitative indicator of which results best matched the human annotators' harmonized interpretation. However, testing a model's capacity to reproduce the judgments of two annotators does not speak to the correctness or usefulness of results. For example, Table 2 shows how human annotators created two categories—'Health issues' and 'Food insecurity'—for the single 'Health and epidemics' category created by the model. As such, these quantitative results are meant to determine whether an LLM could serve as an additional, like-minded annotator. Figure 3 shows a box plot of the averaged results for each model, with an average similarity of 0.06 and maximum of just under 0.20 (Magistral). This is compared against the two annotators and their two passes in completing the task (0.16 and 0.33 for user A, 0.21 and 0.27 for user B).

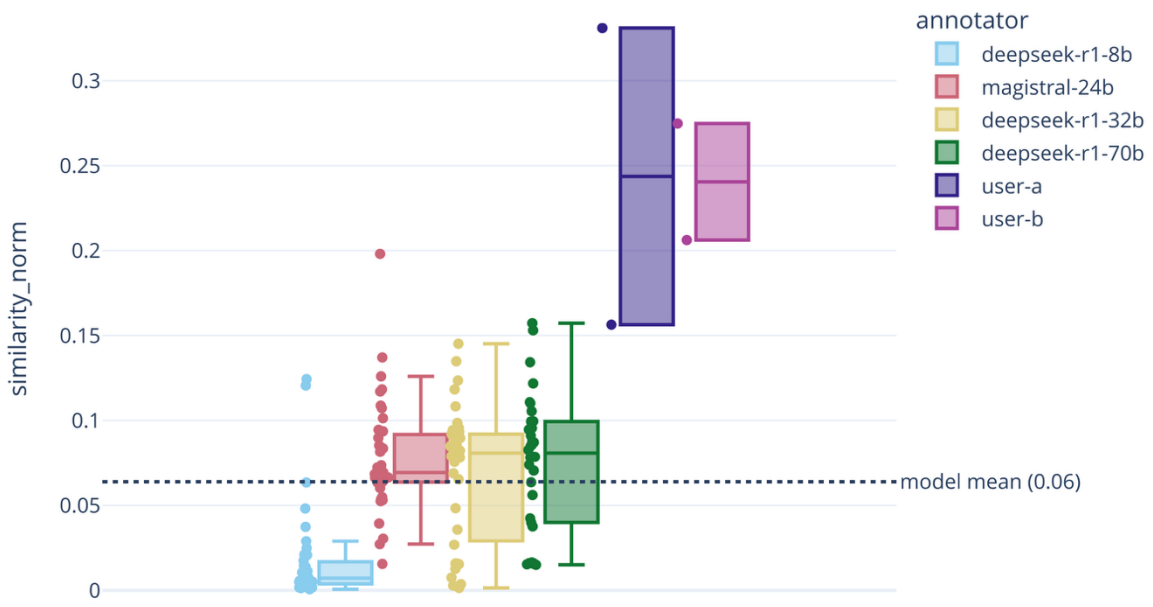

Figure 3: Similarity of generated categories with manual categories

The performance of each model according to the derived Jaccard similarity score is as follows. As normalization tests indicated that similarity scores for both Magistral and Deepseek R1 8B did not have normal distributions (p<.001), the Kruskal-Wallis test was used with the four models, with a result of p<.001. Follow-up Mann-Whitney U rank tests showed each of the three larger models having significant p values when compared with DeepSeek R1 8B (all p<.001). The larger models did not have p values under .05 when compared with each other. Additionally, when comparing the four input formats (JSON, CSV, Python list, and text lines), no values were significant, with the lowest being p=.15, for JSON and Python list formats.

To test the impact of temperature on similarity, ten runs were also executed with Magistral Small and a temperature of 0.0. This achieved the highest average (0.09), minimum (0.05) and maximum (0.23) similarity values among the models, but when compared with Magistral's earlier results (temperature=0.7) the Mann-Whitney result was low yet above the threshold, at p=.081. Descriptive statistics for all model runs are shown in Table 6, which also reflects how often models produced no categorizations (Count < 40), DeepSeek R1 70B having the most such cases.

			Similarity			
Model	**Temperature**	**Count**	**Mean**	**Std**	**Min**	**Max**
deepseek-r1-32b	0.7	39	0.07	0.04	0	0.15
deepseek-r1-70b	0.7	34	0.07	0.04	0.02	0.16
deepseek-r1-8b	0.7	39	0.02	0.03	0	0.12
magistral-24b	0.7	40	0.08	0.03	0.02	0.2
magistral-24b	0.0	40	0.09	0.03	0.05	0.23

Table 6: Model category similarity with gold standard

From a quantitative perspective, the results indicated low similarity with human annotator choices for most runs, although outliers fall within the range of human annotator similarity scores. While the smallest model studied could be discarded, the other three occasionally offered results somewhat more aligned with the gold standard. That said, computational costs associated with larger models may be too burdensome if results do not significantly exceed those of Magistral.

If relying on outliers was the best means in this study to obtain results similar to a team of annotators with specialized knowledge, further refinements would require introducing more techniques and variables. In particular, the zero-shot modality, combined with minimal prompt instructions, a structured output definition that eliminated hallucinations in span production, and (anecdotally) the processing of a large set of spans without chunking gave the task an increased degree of difficulty.

4.3 Semantic Overlap with Manual Categorization

Of the ten best runs, four were generated by Magistral 24B and six by the DeepSeek-R1 family. Top run 1 (Magistral 24B) generated 25 categories with 7 sharing between 0.3 and 1.0 Jaccard similarity with equivalent manual categories. Top run 2 (Magistral 24B) generated 22 categories, including 7 categories with values similar to top run 1. It was the only categorization to include a dedicated category for 'Israeli-Palestinian conflict specific terms'. Interestingly, manual annotators had previously debated a similar categorization but rejected it in favor of maintaining generic categories. However, the LLM also generated a unique category, 'Complex situations and cumulative effects', which was deemed both useful and reflective of the data's semantics. Conversely, it produced categories considered duplicates, specifically 'Crimes against humanity' and 'Human rights violations.'

Top run 3 (DeepSeek R1 70B) generated 15 categories; only two of these surpassed the 0.25 mark with manual categorizations. This run also produced four sets of duplicate categories: 'Infrastructure and development' with 'Infrastructure development'; 'Economic factors' with 'Economic development and priorities'; 'Social and humanitarian issues' with 'Humanitarian crises'; and 'Legal and policy issues' with 'Legal and governance issues'. Top run 4 (DeepSeek R1 70B) offered 14 categories, with three cases exhibiting a similarity score from 0.27 to 0.42 with manual categories. It included two closely duplicated categories: 'Environmental disasters' and 'Environmental factors'. Top run 5 (DeepSeek R1 32B) generated 16 categories, four of which range between 0.43 and 0.66 in similarity with manual categorizations.

Top run 6 (Magistral 24B) produced 27 categories, with five sharing 0.25–0.5 similarity with manual annotations. This run included a duplicate category with in vivo labels: 'The [West Bank] wall's impact' and 'The wall and its associated regime'. Additionally, two false categories, 'Government housing and construction policies' and 'Agricultural and economic practices', were identified as not reflecting the content of the spans,

alongside a useless category simply titled 'Title'. This run also generated two near duplicates, 'Infrastructure and development projects' and 'Oil and gas projects', as well as another duplicate pair: 'Security crises and threats' and 'Insecurity and threats.'

Top run 7 (DeepSeek R1 32B) generated 16 categories, with two exhibiting 0.27–0.30 similarity with comparable manual categories. It identified a useful category, 'Ethnic and cultural issues', alongside a rather unhelpful but interesting one, 'Forced displacement', which, while reflecting the semantic nature of the spans, is unsuitable for conceptual analysis because it links the concept of analysis to itself as a cause. This offers no utility, despite the valid argument that specific instances of forced displacement may lead to other cases. Top run 8 (DeepSeek R1 70B) produced 16 categories, with four showing between 0.33–0.5 similarity. Another unhelpful category was considered too broad, 'Legal and political frameworks', but a unique and useful category, 'Terrorism and insurgency', was also revealed. This run additionally produced 'Internally displaced persons' (IDPs), a category based on references to IDP issues reflected present in the spans. However, no span suggests IDPs are a direct cause of forced displacement.

Top run 9 (Magistral 24B) generated 18 categories, with five demonstrating 0.25–0.6 similarity with manual annotation. It also produced an unhelpful category, 'Violent and extractive systems of colonialism both past and present', which featured an in vivo label. Finally, top run 10 (DeepSeek R1 8B) generated the highest number of categories at 44, predominantly labeled in vivo, and was subsequently excluded from analysis. Table 7 details the semantic overlap between top runs 1–9 and the manual categories, as well as the unique and useful categories identified through LLM-generated categorization.

	Top LLM runs by similarity with manual categories								
Category label	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Non-state armed actor	X	X	X	X	X	X	X	X	X
Conflict	X	X	X	X	X	X	X	X	X
Violence	X	X	X	X		X	X	X	X
State	X	X	X	X	X	X			X
State armed actor	X	X	X	X	X	X		X	X
Military operation	X	X		X	X		X	X	X
Rights violations	X	X	X	X	X	X	X	X	X
Displacement order		X	X	X	X	X	X		
Resource extraction, development projects and land exploitation	X	X	X	X	X	X	X	X	X
Coercive and oppressive practices				X					
Crime	X			X			X	X	
Natural hazard	X	X		X	X	X	X	X	X

Climate change	X	X	X	X	X	X	X	X	X
Economic disruption	X	X	X	X		X	X	X	
Expansionism	X	X							
Freedom of movement restrictions						X			
Insecurity	X	X			X	X	X		X
Inequality	X				X	X			X
Political instability	X	X					X	X	X
Weaponry and explosive hazards									
Environmental disruption	X	X	X	X	X	X	X	X	X
Food insecurity		X			X	X			X
Forced recruitment				X				X	
House demolitions					X				
Gender-based violence	X								
Health issue	X				X		X		
Impediments to accessing humanitarian services		X	X			X	X	X	X
Lack of international action	X	X			X				
Population movements and demographic change	X	X		X	X				X
Poverty		X						X	X
Vulnerability									X
Drought									
Intergenerational impact		X							
Limited access to basic services									
Complex situations and cumulative effects		X							
Ethnic and cultural issues							X		
Terrorism and insurgency								X	

Table 7: Semantic overlap between manual and LLM-generated categories

(unique categories in italics)

5. Conclusion

Corpus-based conceptual analysis requires inductive categorization of large volumes of lexical data. This study explored LLM-generated inductive categorization with Magistral and DeepSeek R1 local models and a dataset of lexical spans denoting causes of the concept of FORCED DISPLACEMENT. Our study confirms findings from existing literature and provides insights into the application of LLM-assisted inductive categorization, a process useful for data-driven lexicographic projects that derive descriptions from large datasets of lexical data.

First, LLM-generated categorization shows potential by identifying valuable overlooked categories, by helping validate manual categories and by providing alternative category labels. Label similarity of the best ten model outputs against the gold standard showed significant semantic overlap. Nonetheless, output requires manual revision due to undesired behaviors, specifically incorrect labeling and the omission of spans from category assignment.

Second, providing minimal task instructions was unsuccessful for the 8- to 72-billion parameter models. This zero-shot experiment found that models only achieved similarity scores comparable with manual categorization in outlier occasions (out of 40 runs per model). Additionally, providing spans in different formats (e.g., JSON or CSV) showed no statistical difference. Further investigation is warranted into variations in prompting and structured output definition as well as including concept-specific information in prompts that may also help reduce incorrect labeling.

Third, output usefulness appears contingent on balancing the number of spans left unprocessed with the number of category labels generated. The Magistral and DeepSeek R1 models with 24 to 72 billion parameters exhibited a tendency to successfully categorize more spans only to reduce the number and validity of categories. Consequently, only intermediate results were usable. For tasks where the entire set of spans is provided for categorization, we recommend excluding underperforming LLM outputs automatically by comparing the number of categories with spans ignored. In future work, we will also investigate a per-code task strategy, as suggested by Dunivin (2025, p. 13).

Fourth, while quantitative methods can measure similarities in category assignment between model and manual results, a more comprehensive battery of evaluation tasks encompassing diverse modalities is required. In this study, optimal LLM categorization results were quantitatively identified against a gold standard, defined through consensus, by converting multiple Jaccard similarity scores into a single normalized measure. This assessment also focused on the number of categories rather than on category labels. Determining that a model can reliably assist annotators necessitates further investigation using more open tasks, without restricting model output to a pre-existing gold standard.

In conclusion, LLM outputs exhibited low overall similarity with manually generated categories, highlighting limitations in category granularity and redundancy. However, outlier runs achieved similarity scores comparable to annotators, while revealing useful insights that were not captured in manual efforts. This suggests their potential as complementary analytical tools. Future work will focus on exploring multi-category approaches, hybrid approaches incorporating human-in-the-loop validation, refined prompting strategies, and additional pre- and post-processing of data.

6. Acknowledgements

This research was funded by the Regional Government of Andalusia, Spain, with project PROYEXCEL_00369 (VariTermiHum), the 2023 Ramón Areces Foundation grant scheme for PhD theses in the Humanities and a 2024 Arqus Talent Fund grant.

7. References

Arlinghaus, C. S. (2024). LLM-Assisted Inductive Categorization: A Step-by-step guide. https://zenodo.org/records/13379684

Arlinghaus, C. S., Wulff, C., & Maier, G. W. (2024). Inductive Coding with ChatGPT - An Evaluation of Different GPT Models Clustering Qualitative Data into Categories. OSF. https://doi.org/10.31219/osf.io/gpnye

Babchuk, W. A., & Boswell, E. (2023). Grounded theory. In R. J. Tierney, F. Rizvi, & K. Ercikan (eds.) *International Encyclopedia of Education (Fourth Edition)*. Amsterdam: Elsevier, pp. 107-122. https://doi.org/10.1016/B978-0-12-818630-5.11013-9

Bakharia, A., Shibani, A., Lim, L.-A., McCluskey, T., & Buckingham Shum, S. (2025). From Transcripts to Themes: A Trustworthy Workflow for Qualitative Analysis Using Large Language Models. In M. Hlosta, I. Moser, A. Winer et al. (eds.) *Joint Proceedings of LAK 2025 Workshops co-located with 15th International Conference on Learning Analytics and Knowledge (LAK 2025)*. Dublin, Ireland, pp. 179-189. https://ceur-ws.org/Vol-3995/LLMQUAL_paper1.pdf

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., & Baker, R. S. (2024). ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In A. Olney, I. Chounta, Z. Liu, O. Santos, & I. Bittencourt (eds.) *Artificial Intelligence in Education. AIED 2024*. Cham, Springer, pp. 134-149. https://doi.org/10.1007/978-3-031-64299-9_10

Bingham, A. J. (2023). From Data Management to Actionable Findings: A Five-Phase Process of Qualitative Data Analysis. International Journal of Qualitative Methods, 22, 16094069231183620. https://doi.org/10.1177/16094069231183620

Chambó, S., & León-Araúz, P. (2023). Corpus-driven conceptual analysis of epidemic and coronavirus for the Humanitarian Encyclopedia: a case study. *Terminology*,

29(2), pp. 180–224. https://doi.org/10.1075/term.00069.cha

Chase, H. (2022). LangChain [Computer software]. Available at https://github.com/langchain-ai/langchain

Condamines, A. (2022). How the Notion of "Knowledge Rich Context" Can Be Characterized Today. *Frontiers in Communication*, 7. https://doi.org/10.3389/fcomm.2022.824711

Dunivin, Z. O. (2025). Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science*, 14. https://doi.org/10.1140/epjds/s13688-025-00548-8

Dauvin, P., & Siméant-Germanos, J. (2002). *Le travail humanitaire*. Paris: Presses de Sciences Po.

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4), pp. 997-1019. https://doi.org/10.1177/08944393231220483

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://arxiv.org/abs/2501.12948

Drapal, J., Westermann, H., & Savelka, J. (2023). Using large language models to support thematic analysis in empirical legal studies. In J. Spanakis, G. VanDijck, & G. Sileno (eds.) *Legal knowledge and information systems (Vol. 379), JURIX 2023: The Thirty-sixth Annual Conference*. Amsterdam: IOS Press, pp. 197-206. https://doi.org/10.3233/FAIA230965

Eberwein, W.-D., & Saurugger, S. (2013). The professionalization of international non-governmental organizations. In B. Reinalda (ed.) *Routledge Handbook of International Organization*. Abingdon: Routledge, pp. 257-269.

Faber, P. (2015). Frames as a framework for terminology. In H. J. Kockaert & F. Steurs (eds.) *Handbook of Terminology: Volume 1* Amsterdam: John Benjamins, pp. 14-33. https://benjamins.com/catalog/hot.1.fra1

Faber, P. (2022). Chapter 16. Frame-based Terminology. In P. Faber & M.-C. L'Homme (eds.) *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge (Vol. 23).* Amsterdam: John Benjamins, pp. 353-376. https://doi.org/10.1075/tlrp.23.16fab

Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., & Perrault, S. T. (2024). CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In F. F. Mueller, P. Kyburz, J. R. Williamson et al. (eds) *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3613904.3642002

Gorin, V. (2024). Humanitarian studies: A field still in the making. *Alternatives Humanitaires/Humanitarian Alternatives*, 25. https://www.alternatives-humanitaires.org/en/2024/03/20/humanitarian-studies-a-field-still-in-the-making/

Hampton, J. A. (2020). Investigating differences in people's concept representations. In T. Marques & A. Wikforss (eds.) *Shifting Concepts: The Philosophy and*

Psychology of Conceptual Variability. Oxford: Oxford University Press, pp. 67-82. https://doi.org/10.1093/oso/9780198803331.003.0005

Isaacs, L., Chambó, S., & León-Araúz, P. (2024). Humanitarian Corpora for English, French and Spanish. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (eds.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* pp. 8418–8426. https://aclanthology.org/2024.lrec-main.738

Kantner, C., & Overbeck, M. (2020). Exploring Soft Concepts with Hard Corpus-Analytic Methods. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische Textanalyse*. Berlin: De Gruyter, pp. 169-190. https://doi.org/10.1515/9783110693973-008

Katz, A., Gerhardt, M., & Soledad, M. (2024). Using generative text models to create qualitative codebooks for student evaluations of teaching. *International Journal of Qualitative Methods*, 23. https://doi.org/10.1177/16094069241293283

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1). https://doi.org/10.1007/s40607-014-0009-9

León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In P. Drouin, A. Francoeur, J. Humbley, & A. Picton (eds.) *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins. pp 213-258.

León Araúz, P., Reimerink, A. & García Aragón, A. (2013). Dynamism and context in specialized knowledge. *Terminology*, 19(1), pp. 31-61. https://doi.org/10.1075/term.19.1.02leo

Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., & De Paoli, S. (2024). Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs In Biomedicine*, 255. https://doi.org/10.1016/j.cmpb.2024.108356

Mistral-AI: Rastogi, A., Jiang, A. Q., Lo, A. et al. (2025). Magistral. https://doi.org/10.48550/arXiv.2506.10910

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods*, 22, https://doi.org/10.1177/16094069231205789

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). Doccano: Text Annotation Tool for Human [Computer software]. https://github.com/doccano/doccano

Odlum, A., & Chambó, S. (2022). *Horizontally integrating diverse definitions and debates on key concepts in an online Humanitarian Encyclopedia*. Accessed at: https://humanitarianencyclopedia.org/expertise-note/horizontally-integrating-diverse-definitions-and-debates-on-key-concepts-in-an-online-humanitarian-encyclopedia (15 July 2025).

Rychlỳ, P. (2007). Manatee/Bonito—A modular corpus manager. In P. Sojka & A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language*

Processing, RASLAN 2007, pp. 65–70.

San Martín, A., & León-Araúz, P. (2013). Flexible terminological definitions and conceptual frames. In C. Tao, Y. He, L. Toldo et al. (eds) *Proceedings of the ICBO2013 Workshops: International Workshop on Vaccine and Drug Ontology Studies (VDOS 2013), International Workshop on Definitions in Ontologies (DO 2013)*. https://ceur-ws.org/Vol-1061/Paper3_DO2013.pdf

San Martín, A. (2022a). A Flexible Approach to Terminological Definitions: Representing Thematic Variation. *International Journal of Lexicography*, 35(1), pp. 53-74. https://doi.org/10.1093/ijl/ecab013

San Martín, A. (2022b). Contextual Constraints in Terminological Definitions. *Frontiers in Communication*, 7. https://doi.org/10.3389/fcomm.2022.885283

Sezgin, Z., & Dijkzeul, D. (eds.). (2015). *The New Humanitarians in International Practice: Emerging Actors and Contested Principles*. Abingdon: Routledge. https://doi.org/10.4324/9781315737621