Unveiling Hidden Stories: Automated Narrative Extraction from Holocaust Diaries with Ensemble LLMs

Angelina Parfenova[1,2,*,†]

[1]Technical University of Munich, Munich, Germany
[2]Lucerne University of Applied Sciences and Arts, Rotkreuz, Switzerland

Abstract

This paper presents a novel application of ensemble-based large language models (LLMs) with Retrieval-Augmented Generation (RAG) for automated inductive coding of Holocaust children's diaries. Our approach integrates multiple smaller LLMs, fine-tuned via Low-Rank Adaptation (LoRA), and employs a moderator-based mechanism to simulate collaborative human consensus. We evaluate our best model on a curated dataset of diaries, demonstrating significant improvements in coding consistency and specificity. Our results highlight the potential of ensemble-based LLMs with RAG for analyzing sensitive historical texts, offering a scalable and efficient alternative to manual coding while preserving the nuanced emotional and thematic content of the diaries.

Keywords

Inductive coding, Holocaust diaries, Ensemble models, Retrieval-Augmented Generation, Qualitative analysis

1. Introduction

The analysis of Holocaust children's diaries presents unique challenges due to the emotionally charged and historically significant nature of the content. These diaries, often written under extreme duress, capture narratives of loss, resilience, and survival, which require sensitive and nuanced interpretation [1]. Traditional qualitative coding methods, while effective for small-scale studies, are manual, time-consuming, labor-intensive, and prone to inconsistencies, particularly when applied to large datasets [2]. Recent advances in natural language processing (NLP) have enabled the automation of qualitative coding tasks, offering promising solutions for scaling textual analysis and extracting meaningful narratives from documents [3, 4, 5, 6]. However, existing approaches often struggle to capture the emotional depth, historical specificity, and narrative richness of diaries, limiting their applicability to such sensitive texts [7].

Inductive coding is a qualitative analysis approach in which codes emerge directly from the data rather than being predefined. A *code* represents a concise label that captures the core meaning of a text segment. This approach is a part of thematic analysis, a method for identifying and structuring patterns in qualitative data [8]. The process typically involves iteratively generating codes, clustering them into broader categories, and refining themes to represent the data's underlying structure. Inductive coding is particularly useful for exploratory studies, such as historical text analysis, where themes emerge organically. However, manual thematic analysis is time-consuming and subjective, posing scalability challenges for large textual datasets.

In this work, we propose a novel framework for automated inductive coding using ensemble-based large language models (LLMs) with Retrieval-Augmented Generation (RAG) [9]. Our approach uses the strengths of multiple smaller LLMs (7B and 8B parameters) in an ensemble framework, combining their outputs and feeding them into larger moderator LLM to generate high-quality codes that reflect the thematic and emotional complexity of the texts. To ensure consistency and reduce redundancy, we integrate RAG, which references previously assigned codes to maintain coherence across similar inputs.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

angelina.parfenova@tum.de (A. Parfenova)
0000-0002-5245-5571 (A. Parfenova)

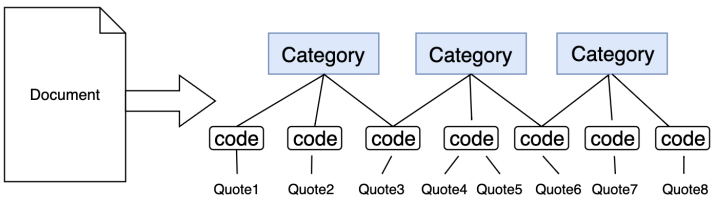

Figure 1: Illustration of the thematic analysis process. Segments from a transcript are first highlighted and assigned initial codes. These codes are then grouped into subcategories based on emerging patterns or shared meanings. Finally, subcategories are consolidated into broader categories, representing higher-level themes that capture the core insights from the data.

This combination of ensemble modeling and RAG addresses key limitations of existing methods [10, 5], offering a scalable and efficient alternative to manual coding while preserving the nuanced content.

We apply our framework to a curated dataset of Holocaust children's diaries, demonstrating its effectiveness in capturing recurring themes such as family separation, fear, and hope. Our results show significant improvements in coding consistency, specificity, and alignment with human-coded benchmarks, highlighting the potential of ensemble-based LLMs with RAG for analyzing sensitive historical texts.

2. Background

Qualitative data analysis (QDA) is one of the main methods in social science research, allowing researchers to identify, categorize, and interpret patterns within textual data [11, 12]. Central to this process is the concept of *coding*, where meaningful segments of text are assigned concise labels, or *codes*, that capture their essence (see Figure 1). According to Saldana [2], a code is "a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data." In thematic analysis, one of the most widely used methods in QDA, these codes are further grouped into broader categories to reveal hierarchical relationships and underlying themes within the data [13].

Recent advances in natural language processing (NLP) have introduced the use of large language models (LLMs) to automate qualitative coding tasks [14, 10, 15]. However, two critical challenges remain unaddressed in this domain. First, traditional evaluation metrics such as BERTScore and ROUGE, while effective for summarization tasks, are insufficient to assess the quality of qualitative codes [10, 5]. Recent work by Chen et al. [5] introduced unsupervised metrics tailored for code evaluation, but these approaches lack the ability to directly compare model outputs to human annotations. In this work, we address this gap by proposing a supervised evaluation framework that aligns model-generated codes with human-coded benchmarks.

Second, while individual LLMs demonstrate remarkable performance, their outputs often vary due to differences in training data, architectures, and model parameters [16, 17]. This variability mirrors the subjectivity inherent in human coding, where individual coders may interpret the same text differently. To address this challenge, ensemble methods, techniques that combine multiple models, were explored to combine the strengths of diverse models and improve overall performance [18, 19]. For example, Jiang et al. [20] demonstrated the effectiveness of ensembling in complex natural language generation tasks, while Cai et al. [21] highlighted the potential of mixture-of-experts (MoE) frameworks for specialized sub-tasks.

This study builds upon the concept of ensemble methods but diverges from existing approaches by adopting a *moderator-based framework*. Unlike fusion techniques that combine outputs probabilistically, our approach incorporates a final decision-making model tasked with selecting the best candidate or proposing a novel output. This design reflects the dynamics of human collaboration with a leader, where consensus is driven by a final arbiter, rather than by averaging or blending opinions. By employing this moderator model, we aim to mimic the decision-making process and demonstrate its effectiveness in

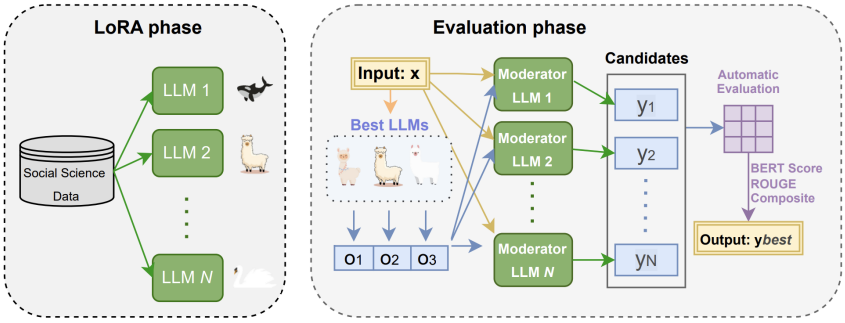

Figure 2: Workflow for integrating LLMs in qualitative data analysis. In the LoRA phase, multiple LLMs are fine-tuned using social science data. In the Evaluation phase, an input x is processed by selected LLM moderators to generate output candidates $(y_1, y_2, \ldots, y_N)$. These candidates are evaluated using metrics like BERTScore and ROUGE, with the best output (y_{best}) selected based on composite scores.

automating inductive coding tasks, particularly for sensitive historical texts such as personal diaries.

In the context of Holocaust studies, NLP has been increasingly applied to analyze historical texts, including survivor testimony, diaries, and archival documents. For instance, Schwartz et al. [22] used topic modeling to identify recurring themes in Holocaust survivor testimonies, while Eisenstein et al. [23] employed sentiment analysis to explore emotional patterns in wartime diaries. However, these studies often rely on traditional NLP techniques, which struggle to capture the nuanced emotional and thematic content of Holocaust texts.

Ensemble learning is a well-established strategy for improving model performance by combining the strengths of multiple models, often referred to as "weaker models" [18, 24]. Common approaches include weighting individual models based on their performance or aggregating diverse outputs to produce a unified result. For example, the Mix-of-Experts (MoE) framework [21] employs specialized sub-models to make predictions and merges their outputs for improved accuracy. Similarly, LLM-Blender [20] demonstrates the potential of ensembling by combining ranked outputs from multiple models to achieve superior performance in complex natural language generation tasks.

3. Methodology

Our pipeline consists of three key stages: (1) input processing by multiple smaller LLMs, (2) moderation and refinement of outputs by larger LLMs, and (3) retrieval-augmented generation to ensure consistency. The steps are described below.

3.1. Ensemble Model Framework

Our ensemble framework combines three smaller LLMs (7B and 8B parameters) to process each input diary entry independently. These models were fine-tuned using Low-Rank Adaptation (LoRA) [25] on a diverse corpus of social science data (see Table 1, enabling them to capture domain-specific patterns while maintaining computational efficiency. LoRA fine-tuning allows for efficient adaptation of pre-trained models to specialized tasks, such as inductive coding, without requiring extensive retraining or large-scale datasets.

The outputs from these models are evaluated by a moderator model, which refines and consolidates the results (see Appendix B). The moderator is tasked with assessing the quality and relevance of the generated codes, ensuring that the final output reflects a consensus among the ensemble. This approach reduces variability and improves the quality of the generated codes, addressing the inherent subjectivity of individual LLMs [16, 17].

3.2. Retrieval-Augmented Generation (RAG)

RAG is integrated into our pipeline to ensure consistency and reduce redundancy in the coding process. RAG operates by referencing a database of previously assigned codes, which are retrieved based on semantic similarity to the current input. For each input, RAG computes the cosine similarity between the input embedding $\phi(x)$ and the embeddings of previously assigned codes $\phi(p_i)$. If the similarity exceeds τ, the retrieved code is reused; otherwise, a new code is generated. The integration of RAG also addresses the challenge of code redundancy, a common issue in automated qualitative coding. By aligning new outputs with historical coding decisions, RAG ensures that similar inputs receive consistent labels.

3.3. Evaluation Metrics

We evaluate our approach using a combination of quantitative metrics (e.g., composite score, ROUGE [26], BERTScore [27]) and qualitative analysis. The composite score, which incorporates semantic, lexical, and structural alignment, serves as the primary metric for assessing coding quality.

Composite Score To provide a comprehensive evaluation of coding quality, we introduce a Composite Score ($\mathscr{C}$) that combines multiple normalized metrics:

$$\mathscr{C} = \frac{1}{4}\left[\tilde{S}_c + \tilde{M} + (1 - \tilde{L}) + (1 - \tilde{J})\right], \tag{1}$$

where: $\tilde{S}_c$: Normalized cosine similarity between code embeddings [28], measuring semantic alignment with human-coded references; $\tilde{M}$: Scaled METEOR score [29], which balances precision and recall while accounting for synonymy and stemming; $\tilde{L}$: Normalized code length percentile, where shorter codes are preferred to avoid verbosity; $\tilde{J}$: Normalized Jensen-Shannon divergence [30], which quantifies the distributional similarity between generated and reference codes. Each metric is normalized using min-max scaling:

$$\tilde{\cdot} = \frac{\cdot - \min}{\max - \min}, \tag{2}$$

ensuring that all components contribute equally to the Composite Score. The terms $(1 - \tilde{L})$ and $(1 - \tilde{J})$ invert the code length and divergence metrics, respectively, so that higher values indicate better performance across all dimensions.

4. Experiments and Results

Our experiments began with the training and evaluation of ensemble models using a dataset of 1,000 *code-quote* pairs compiled from social science research studies and the SemEval-2014 Task 4 dataset [31] (see Table 1). The dataset included 600 examples from social science studies and 400 examples from reviews, each annotated by 3–5 coders to establish mutually agreed *golden standard* codes. The dataset was split into training (900 examples) and test (100 examples) sets, with hyperparameters selected based on training performance.

Model Selection and Fine-Tuning We evaluated several open-source LLMs, including Llama3 [32], Falcon [33], Mistral [34], Vicuna [35], Gemma [36], and TinyLlama [37]. Each model was fine-tuned using Low-Rank Adaptation (LoRA) [25] on the training dataset, enabling efficient adaptation to the inductive coding task. The fine-tuned models generated outputs o_i for each input x, which were evaluated using BERTScore and ROUGE. The top three performing models—Llama3, Falcon, and Mistral—were selected for the ensemble framework (see Appendix A).

Table 1
Overview of dataset characteristics used for LoRA training. (A) Data sources and descriptions, including 600 quotes from social science studies and 400 quotes from SemEval 2014 Task 4. (B) Dataset statistics and splits, with 900 examples for training and 100 for testing. This dataset was annotated by multiple coders to create a *golden standard* and served as the foundation for fine-tuning the base 7B and 8B models.

N Quotes	Description
	Social Science Studies Data: 600 quotes
78	Study about interaction with self-tracking devices (interviews)
22	Study about life transitions and mobility (interviews)
82	Study about interaction with voice assistants (interviews)
28	Study about museums and cultural experiences (interviews)
25	Study on doctors' experiences with pregnant women (interviews)
110	Study on universal and national values (interviews)
24	Study on procrastination and budget planning (interviews)
56	Study on technology interactions and user feedback (reviews)
175	Study about social expectations (interviews)
	SemEval 2014; Task 4: 400 quotes
211	Restaurant reviews
189	Laptop reviews

Statistic	Overall	Train	Test
Total Quotes	1000	900	100
Social Science Data	600	550	50
SemEval Data	400	350	50
Num of Data Sources	11	11	11
Unique Codes	680	624	94
Avg. Quote Length	$254.75_{274.28}$	$280.89_{280.89}$	$234.80_{201.61}$
Avg. Code Length	$19.95_{10.43}$	$20.04_{10.70}$	$19.27_{10.53}$

Ensemble and RAG Integration The ensemble framework combines the outputs of the top three models, which are then refined by a set of moderators. RAG is integrated to ensure consistency across similar inputs. For each new input x, RAG computes the cosine similarity between its embedding $\phi(x)$ and previously assigned code embeddings $\phi(p_i)$. If the similarity exceeds a threshold τ, the existing code is reused; otherwise, a new code is generated.

Table 2
Performance comparison of individual models, standard ensembles, and RAG-enhanced ensembles across key metrics. RAG column indicates the similarity threshold between generated code and codes before it (can look only in the past). Models are evaluated using BERTScore (Precision, Recall, and F1), ROUGE (1, 2, and L), Composite Score, Average Code Length, and Number of Unique Codes.

Model	RAG	BERTScore			ROUGE			Composite Score	Code length	Unique Codes
		P	*R*	*F1*	*1*	*2*	*L*			
Mixtral8x7B	-	0.83	0.84	0.83	0.08	0.01	0.08	0.33	6.83	100
Mixtral8x7B Ensemble	-	0.83	0.85	0.84	0.12	0.01	0.08	0.91	4.02	100
Mixtral8x7B Ensemble + RAG	0.7	0.84	0.85	0.84	0.12	0.01	0.11	0.99	4	71
Llama3.3 70B	-	0.84	0.86	0.85	0.12	0.03	0.12	0.38	3.5	96
Llama3.3 70B Ensemble	-	0.85	0.86	0.85	0.15	0.02	0.15	0.50	3.57	100
Llama3.3 70B Ensemble + RAG	0.5	0.85	0.88	0.86	0.15	0.03	0.15	0.74	3.49	53
GPT-4	-	0.83	0.84	0.84	0.02	0.00	0.02	0.44	4.39	24
GPT-4 Ensemble	-	0.83	0.85	0.84	0.11	0.02	0.10	0.74	5.01	100
GPT-4 Ensemble + RAG	0.8	0.83	0.85	0.84	0.12	0.02	0.10	0.54	4.62	89
GPT-4o	-	0.85	0.86	0.86	0.04	0.00	0.04	0.37	3.8	43
GPT-4o Ensemble	-	0.85	0.87	0.86	0.14	0.02	0.14	0.74	4.26	100
GPT-4o Ensemble + RAG	0.7	0.85	0.87	0.86	0.11	0.00	0.11	0.54	4.42	71
Llama3 8B Instruct	-	0.83	0.86	0.84	0.12	0.02	0.11	0.61	8.45	98
Falcon 7B Instruct	-	0.83	0.85	0.84	0.09	0.01	0.09	0.47	12.76	100
Mistral 7B Instruct	-	0.83	0.85	0.84	0.07	0.01	0.07	0.67	11.77	100

The integration of RAG into our ensemble framework substantially reduces redundancy in generated codes by aligning new outputs with previously assigned codes. As demonstrated in Table 2, RAG-enhanced ensembles produce more concise outputs, achieving an average code length reduction from 6.83 to 4.00 tokens—a 41.5% improvement over non-RAG ensembles.

Further analysis highlights the impact of RAG on code diversity. While the human gold standard comprises 47 unique codes with an average length of 2.79 tokens, non-RAG models exhibit excessive code proliferation, often generating unique codes for each input. In contrast, RAG integration significantly reduces this redundancy, with Llama3.3 70B Ensemble+RAG and Mixtral 8x7B Ensemble+RAG producing 53 and 71 unique codes, respectively. This brings the models closer to human-like coding efficiency, as illustrated in Table 2.

5. Holocaust Dataset Analysis

To evaluate the generalizability of our framework, we applied the best-performing ensemble model (Mixtral 8x7B with RAG) to a curated dataset of 224 Holocaust children's diaries. The dataset was constructed from the book *Children in the Holocaust and World War II: Their Secret Diaries* by Laurel Holliday [38]. We selected diary entries that were explicitly labeled with both day and year, ensuring temporal consistency and facilitating the analysis of chronological patterns in the children's experiences.

5.1. Results

Temporal Distribution of Diary Entries The dataset spans from 1939 to 1945, capturing key moments in World War II from the perspective of children. Figure 4 shows the distribution of diary entries over time, revealing a notable increase in the density of entries around major historical events. For example, the invasion of Poland in 1939 and the intensification of bombings and deportations in later years are reflected in the children's writings. This temporal distribution demonstrates how the evolving wartime environment influenced the frequency and content of their diary entries.

Thematic Analysis of Codes Our framework generated a diverse array of codes that reflect the children's experiences and emotional states. Early entries, such as those from Janine Phillips in August and September 1939, focus on themes like *Impact of unexpected war news* and *Family Reunion; Prepared for War*. As the war progressed, the model identified more intense and emotionally charged themes, such as *Devastating bombing begins*, *War-time scarcity; community support*, and *Fear of war's soul-crushing impact*.

Recurring codes like *Loneliness, despair, longing for relief* and *Severe hunger, bread scarcity* illustrate the isolation and deprivation on the children. At the same time, the model captured moments of resilience, such as *Found purpose, devoted to homeland* and *Dreaming of peace amidst chaos*, highlighting the children's capacity for hope and adaptation even in dire circumstances. These findings demonstrate the model's ability to capture both the emotional depth and thematic complexity of the diaries.

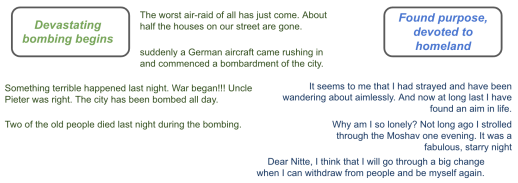

Figure 3: Illustrative diary excerpts with codes.

Table 3: Most frequently occurring codes.

Code	Frequency
Impact of unexpected war news	10
Devastating bombing begins	8
Found purpose, devoted to homeland	6
Ordered to shovel snow	6
Emotional turmoil	4
Imprisoned; longing for Daddy	4
Jews displaced, possessions limited	4
Experiencing Joy, Relieved	3
Struggling through darkness	3

Individual Variations The diaries reveal significant individual variations in how children responded to their experiences. For instance, Janine Phillips' entries focus on the immediate shock and logistical challenges of war, while others, such as those from anonymous authors, emphasize personal reflections on family, loss, and survival (see Figure 5). For example, one entry describing the emotional toll of being separated from family members was labeled as *Longing for family; emotional isolation*, while another reflecting on the resilience of children in the face of adversity was coded as *Hope amidst despair; finding strength*. These examples highlight the model's sensitivity to the nuanced emotional and thematic content of the diaries.

Table 3 presents the most frequently occurring codes generated by the Mixtral 8x7B Ensemble RAG model. These codes reflect the dominant themes and emotional states documented by children during the Holocaust. The frequency of each code provides insight into the shared experiences and collective trauma of the children, as well as their individual responses to the evolving wartime environment.

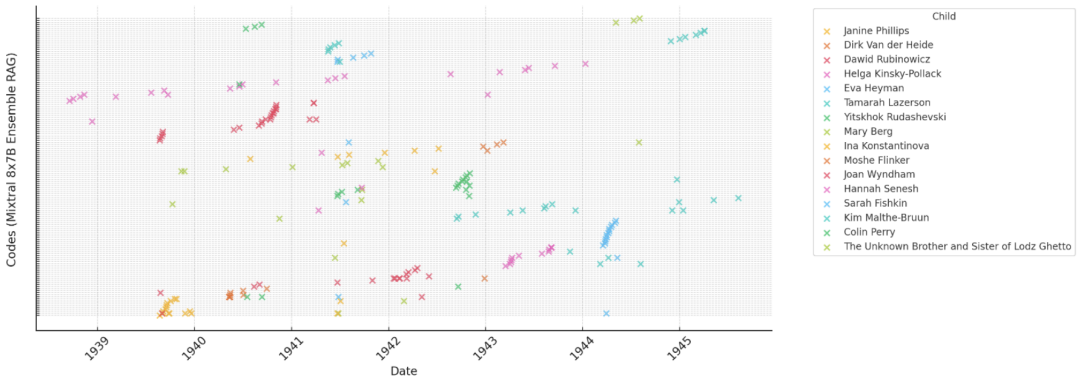

Figure 4: Timeline of diary entries from Holocaust children. This plot illustrates the number of entries per diary.

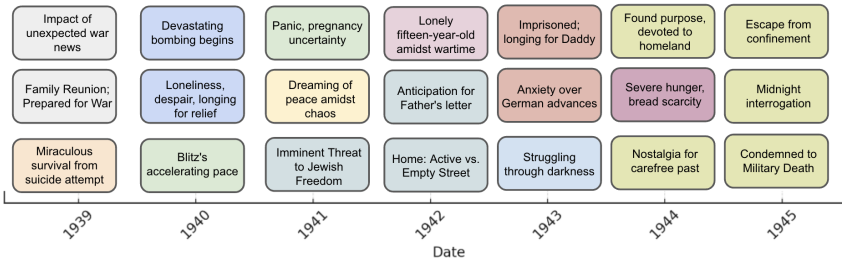

Figure 5: Final codes generated by the model, organized along a timeline from 1939 to 1945. Each colored box represents codes derived from different children, illustrating distinct thematic interpretations over the course of World War II.

6. Discussion

Our study demonstrates the effectiveness of ensemble-based LLMs with Retrieval-Augmented Generation (RAG) for automating inductive coding tasks. The results highlight the framework's ability to capture the emotional and thematic complexity of sensitive historical texts while maintaining consistency and reducing redundancy. Below, we discuss the key implications of our findings, address limitations, and outline directions for future research.

6.1. Ensembles Improve Coding Consistency

A major finding of our study is that ensemble models consistently outperform individual models in inductive coding tasks, as shown in Table 2. This suggests that aggregating multiple model outputs helps reduce inconsistencies, reflecting the consensus-building process employed by human coders in thematic analysis.

The increased consistency observed in ensemble-generated codes aligns with findings from prior research on LLM evaluation, which suggest that individual models often introduce unwanted variability in their outputs due to differences in training data and architectural biases [16, 19]. In contrast, ensemble methods mitigate this variability by integrating diverse inputs, thereby improving robustness. Our results indicate that this effect holds even for smaller models, making ensemble approaches a practical solution for qualitative coding tasks.

6.2. RAG Enhances Code Stability

The integration of RAG significantly improves code stability, as demonstrated by higher composite and ROUGE scores in RAG-enhanced ensembles (Table 2). By referencing previously assigned codes, RAG reduces redundancy and promotes consistency across similar inputs. This is particularly evident in the

reduction of unique code counts (e.g., 53 for Llama3.3 70B+RAG vs. 100 for non-RAG models) and code length (41.5% reduction), bringing model outputs closer to human-like efficiency.

In the context of Holocaust diaries, RAG's ability to align new outputs with historical coding decisions is crucial for capturing recurring themes such as fear, loss, and resilience. For example, entries describing the emotional toll of family separation are consistently labeled as *Longing for family; emotional isolation*, while reflections on the resilience of children are coded as *Hope amidst despair; finding strength*. This consistency enhances the interpretability and usability of the generated codes, making the framework a valuable tool for analyzing large collections of historical texts.

6.3. Balancing Abstraction and Specificity

This finding reflects a fundamental trade-off in LLM-based coding: while abstraction improves generalizability, excessive abstraction can obscure critical nuances. Prior work has noted that LLMs trained on diverse corpora tend to favor generalized patterns over domain-specific details [16, 14]. Our results suggest that ensemble approaches can mitigate this issue by combining diverse levels of abstraction, thereby producing more balanced and contextually grounded outputs. For example, the Mixtral 8x7B ensemble generates codes like *Devastating bombing begins* and *Found purpose, devoted to homeland*, which capture both the emotional depth and thematic specificity of the diaries.

6.4. Insights into Holocaust Diaries

The application of our framework to Holocaust children's diaries provides valuable insights into the experiences of children during World War II. The frequent codes generated by the model, such as *Impact of unexpected war news*, *Devastating bombing begins*, and *Found purpose, devoted to homeland*, reflect the diversity of responses to the war, from shock and despair to resilience and hope. These findings contribute to a deeper understanding of the emotional and psychological impact of the Holocaust on children, shedding light on their capacity for adaptation and survival in the face of unimaginable hardship.

Moreover, the framework's ability to capture individual variations in the diaries—such as Janine Phillips' focus on the immediate shock of war versus other children's reflections on family

Despite its successes, our framework has several limitations that need consideration. First, the reliance on pre-trained LLMs introduces potential biases inherent in the training data, which may affect the quality and fairness of the generated codes. While ensemble methods and RAG mitigate some of these biases, further work is needed to develop bias detection.

Second, the evaluation of automated coding frameworks remains challenging, as no single metric can fully capture the nuances of human judgment. While our composite score combines multiple dimensions of coding quality, it may not fully reflect the interpretative depth required for sensitive historical texts. Future work should explore more sophisticated evaluation frameworks, incorporating human preference modeling and interactive evaluation setups.

Finally, the generalizability of our framework to other languages and cultural contexts remains untested. The Holocaust diaries analyzed in this study are written in English, and the framework's performance on multilingual or non-Western texts may differ. Extending the framework to other languages and cultural settings could reveal additional challenges and opportunities for improvement.

7. Conclusion

Our study demonstrates the potential of ensemble-based LLMs with RAG for automating inductive coding tasks in sensitive and historically significant contexts. The framework's ability to capture the emotional and thematic complexity of Holocaust children's diaries, while maintaining consistency and scalability, highlights its value for qualitative research. By addressing the limitations and exploring future directions outlined above, we can further enhance the interpretability, fairness, and generalizability of automated coding, opening new possibilities for research in history and social science.

References

[1] P. Levi, The Drowned and the Saved, Summit Books, 1986.

[2] J. Saldana, The Coding Manual for Qualitative Researchers, SAGE Publications, 2016.

[3] D. Boyd, K. Crawford, Critical questions for big data, Information, Communication & Society 15 (2013) 662–679.

[4] D. Matter, M. Schirmer, N. Grinberg, J. Pfeffer, Close to human-level agreement: Tracing journeys of violent speech in incel posts with gpt-4-enhanced annotations, 2024. arXiv:2401.02001.

[5] J. Chen, A. Lotsos, L. Zhao, C. Wang, J. Hullman, B. Sherin, U. Wilensky, M. Horn, A computational method for measuring "open codes" in qualitative analysis, 2024. URL: https://arxiv.org/abs/2411.12142. arXiv:2411.12142.

[6] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science?, Computational Linguistics 50 (2024) 237–291.

[7] M. Hirsch, Family Frames: Photography, Narrative, and Postmemory, Harvard University Press, 1997.

[8] V. Braun, V. Clarke, Thematic analysis: A reflexive approach, International Journal of Qualitative Research 11 (2019) 301–310.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[10] D. Parfenova, et al., Automating qualitative analysis with llms, Proceedings of ACL 2024 (2024).

[11] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to wordnet: An on-line lexical database, International journal of lexicography 3 (1990) 235–244.

[12] J. W. Creswell, 30 Essential Skills for the Qualitative Researcher, SAGE Publications, 2016.

[13] V. Braun, V. Clarke, Thematic Analysis: A Practical Guide, SAGE Publications, 2021.

[14] P. Tornberg, Using large language models for automated qualitative coding in the social sciences, Nature Machine Intelligence 5 (2023) 576–586.

[15] T. Fischer, C. Biemann, Exploring large language models for qualitative data analysis, in: M. Hämäläinen, E. Öhman, S. Miyagawa, K. Alnajjar, Y. Bizzoni (Eds.), Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, Miami, USA, 2024, pp. 423–437. URL: https://aclanthology.org/2024.nlp4dh-1.41/. doi:10.18653/v1/2024.nlp4dh-1.41.

[16] S. Bubeck, V. Chandak, et al., Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712 (2023).

[17] H. Touvron, T. Lavril, et al., Llama: Open and efficient foundation language models, in: Proceedings of the 2023 Annual Conference on Machine Learning (ICML), 2023, pp. 123–134.

[18] O. Sagi, L. Rokach, Ensemble learning: A survey, Wiley interdisciplinary reviews: data mining and knowledge discovery 8 (2018) e1249.

[19] D. Jiang, et al., Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, Proceedings of ACL 2023 (2023).

[20] D. Jiang, X. Ren, B. Y. Lin, Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023. URL: https://arxiv.org/abs/2306.02561. arXiv:2306.02561.

[21] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, J. Huang, A survey on mixture of experts, 2024. URL: https://arxiv.org/abs/2407.06204. arXiv:2407.06204.

[22] D. Schwartz, et al., Topic modeling holocaust survivor testimonies, Journal of Digital Humanities 8 (2019) 45–60.

[23] J. Eisenstein, et al., Sentiment analysis of wartime diaries, Computational Linguistics 47 (2021) 601–630.

[24] A. Aniol, M. Pietron, J. Duda, Ensemble approach for natural language question answering problem, in: 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), IEEE, 2019, pp. 180–183.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adap-

tation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.
[26] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.
[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, CoRR abs/1904.09675 (2019). URL: http://arxiv.org/abs/1904.09675. arXiv:1904.09675.
[28] H. Steck, C. Ekanadham, N. Kallus, Is cosine-similarity of embeddings really about similarity?, in: Companion Proceedings of the ACM Web Conference 2024, WWW '24, ACM, 2024, p. 887–890. URL: http://dx.doi.org/10.1145/3589335.3651526. doi:10.1145/3589335.3651526.
[29] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.
[30] M. L. Menéndez, J. Pardo, L. Pardo, M. Pardo, The jensen-shannon divergence, Journal of the Franklin Institute 334 (1997) 307–318.
[31] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: P. Nakov, T. Zesch (Eds.), Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35. URL: https://aclanthology.org/S14-2004. doi:10.3115/v1/S14-2004.
[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[33] F. Pineda, R. Milliere, A. Vlachos, M. Ott, R. Yates, A. Glaese, et al., The falcon series of language models, arXiv preprint arXiv:2306.01116 (2023).
[34] M. A. Team, Mistral: Efficient pretraining of transformer language models, 2023. URL: https://mistral.ai.
[35] C. Li, F. Xie, et al., Vicuna: An open-source chatbot, FastChat: Open Assistant (2023). Available at https://vicuna.lmsys.org/.
[36] G. A. R. Team, Gemma: An instructable, open-source large language model, 2024. URL: https://gemma.ai.
[37] Z. Jiang, et al., Tinyllama: Distilling large language models for efficiency, arXiv preprint arXiv:2310.05637 (2023).
[38] L. Holliday, Children in the Holocaust and World War II: Their Secret Diaries, Washington Square Press, 1995.

A. Detailed fine-tuning results

These results (see Table 4) demonstrate the performance of various models when fine-tuned on the task of open coding using different prompts. BERTScore and ROUGE are reported.

B. Moderator prompt template

Table 4
Detailed Fine-tuning Results. The following table presents the detailed results from fine-tuning experiments, including precision (P), recall (R), F1 score, and ROUGE across different models and prompts.

Model	BERTScore			ROUGE		
	P_{std}	R_{std}	$F1_{std}$	1	2	L
Summarize the main idea of a sentence\n						
Llama3	$0.713_{0.060}$	$0.758_{0.040}$	$0.734_{0.062}$	0.141	0.033	0.153
Falcon	$0.746_{0.073}$	$0.782_{0.097}$	$0.764_{0.095}$	0.176	0.047	0.189
Mistral	$0.729_{0.076}$	$0.787_{0.093}$	$0.756_{0.078}$	0.178	0.047	0.195
Vicuna	$0.731_{0.063}$	$0.777_{0.095}$	$0.753_{0.079}$	0.163	0.028	0.182
Gemma	$0.712_{0.084}$	$0.738_{0.078}$	$0.745_{0.080}$	0.163	0.030	0.168
TinyLlama	$0.718_{0.072}$	$0.775_{0.090}$	$0.757_{0.087}$	0.164	0.052	0.158
Summarize the main idea of a sentence.						
Llama3	$0.718_{0.072}$	$0.788_{0.089}$	$0.750_{0.073}$	0.181	0.059	0.166
Falcon	$0.738_{0.099}$	$0.787_{0.103}$	$0.761_{0.096}$	0.213	0.077	0.210
Mistral	$0.719_{0.072}$	$0.768_{0.086}$	$0.742_{0.075}$	0.157	0.055	0.148
Vicuna	$0.733_{0.079}$	$0.787_{0.095}$	$0.758_{0.081}$	0.193	0.068	0.185
Gemma	$0.719_{0.071}$	$0.779_{0.089}$	$0.746_{0.072}$	0.172	0.049	0.166
TinyLlama	$0.736_{0.083}$	$0.788_{0.092}$	$0.760_{0.081}$	0.207	0.074	0.199
Can you tell me what the main idea of this sentence is in just a few words?						
Llama3	$0.688_{0.055}$	$0.778_{0.084}$	$0.729_{0.061}$	0.116	0.034	0.110
Falcon	$0.753_{0.105}$	$0.787_{0.108}$	$0.768_{0.102}$	0.236	0.104	0.239
Mistral	$0.742_{0.106}$	$0.795_{0.106}$	$0.766_{0.101}$	0.246	0.106	0.235
Vicuna	$0.691_{0.060}$	$0.783_{0.087}$	$0.732_{0.063}$	0.168	0.047	0.164
Gemma	$0.711_{0.075}$	$0.786_{0.093}$	$0.746_{0.079}$	0.171	0.057	0.168
TinyLlama	$0.725_{0.083}$	$0.789_{0.090}$	$0.754_{0.079}$	0.178	0.067	0.177
From the perspective of a social scientist, summarize the following sentence as you would in thematic coding\n						
Llama3	$0.698_{0.059}$	$0.784_{0.083}$	$0.738_{0.062}$	0.130	0.033	0.119
Falcon	$0.745_{0.109}$	$0.792_{0.105}$	$0.766_{0.102}$	0.210	0.089	0.211
Mistral	$0.688_{0.060}$	$0.785_{0.086}$	$0.732_{0.064}$	0.139	0.041	0.131
Vicuna	$0.713_{0.080}$	$0.778_{0.094}$	$0.743_{0.080}$	0.169	0.061	0.166
Gemma	$0.721_{0.085}$	$0.784_{0.093}$	$0.749_{0.082}$	0.180	0.070	0.177
Tinyllama	$0.718_{0.073}$	$0.776_{0.083}$	$0.745_{0.072}$	0.165	0.053	0.158
From the perspective of a social scientist, summarize the following sentence as you would in thematic coding.						
Llama3	$0.685_{0.082}$	$0.781_{0.064}$	$0.733_{0.081}$	0.136	0.025	0.154
Falcon	$0.754_{0.066}$	$0.778_{0.091}$	$0.759_{0.088}$	0.181	0.048	0.190
Mistral	$0.740_{0.067}$	$0.780_{0.088}$	$0.756_{0.071}$	0.172	0.045	0.187
Vicuna	$0.718_{0.071}$	$0.780_{0.094}$	$0.753_{0.073}$	0.165	0.039	0.185
Gemma	$0.700_{0.072}$	$0.780_{0.085}$	$0.746_{0.080}$	0.180	0.046	0.187
TinyLlama	$0.729_{0.076}$	$0.778_{0.089}$	$0.754_{0.080}$	0.169	0.046	0.183
If you were a social scientist doing thematic analysis, what code would you give to this citation?						
Llama3	$0.692_{0.060}$	$0.785_{0.083}$	$0.735_{0.064}$	0.064	0.043	0.126
Falcon	$0.736_{0.093}$	$0.785_{0.101}$	$0.759_{0.092}$	0.206	0.076	0.200
Mistral	$0.686_{0.057}$	$0.785_{0.082}$	$0.731_{0.061}$	0.132	0.044	0.123
Vicuna	$0.719_{0.070}$	$0.789_{0.091}$	$0.751_{0.073}$	0.183	0.063	0.169
Gemma	$0.724_{0.085}$	$0.784_{0.091}$	$0.751_{0.082}$	0.170	0.066	0.168
Tinyllama	$0.720_{0.071}$	$0.778_{0.083}$	$0.747_{0.072}$	0.186	0.053	0.182
What is the gist of this sentence?						
Llama3	$0.680_{0.064}$	$0.780_{0.086}$	$0.725_{0.066}$	0.129	0.042	0.121
Falcon	$0.731_{0.091}$	$0.780_{0.098}$	$0.754_{0.089}$	0.182	0.080	0.179
Mistral	$0.726_{0.079}$	$0.785_{0.095}$	$0.753_{0.079}$	0.165	0.057	0.160
Vicuna	$0.720_{0.070}$	$0.781_{0.089}$	$0.748_{0.072}$	0.172	0.055	0.162
Gemma	$0.707_{0.077}$	$0.773_{0.091}$	$0.737_{0.076}$	0.152	0.059	0.146
Tinyllama	$0.713_{0.057}$	$0.773_{0.079}$	$0.741_{0.061}$	0.143	0.032	0.139
Explain in a couple of words the primary thought expressed in the following text\n						
Llama3	$0.691_{0.062}$	$0.783_{0.085}$	$0.733_{0.066}$	0.120	0.038	0.110
Falcon	$0.734_{0.078}$	$0.778_{0.090}$	$0.754_{0.078}$	0.171	0.049	0.165
Mistral	$0.698_{0.067}$	$0.780_{0.088}$	$0.735_{0.070}$	0.141	0.038	0.131
Vicuna	$0.703_{0.072}$	$0.780_{0.088}$	$0.738_{0.072}$	0.155	0.048	0.148
Gemma	$0.706_{0.064}$	$0.786_{0.086}$	$0.742_{0.066}$	0.177	0.053	0.170
Tinyllama	$0.720_{0.077}$	$0.784_{0.091}$	$0.750_{0.078}$	0.168	0.071	0.163
Explain in a couple of words the primary thought expressed in the following text.						
Llama3	$0.700_{0.068}$	$0.784_{0.055}$	$0.747_{0.063}$	0.142	0.025	0.152
Falcon	$0.752_{0.088}$	$0.779_{0.061}$	$0.760_{0.086}$	0.183	0.042	0.193
Mistral	$0.738_{0.070}$	$0.790_{0.090}$	$0.759_{0.073}$	0.173	0.047	0.183
Vicuna	$0.717_{0.066}$	$0.780_{0.094}$	$0.752_{0.099}$	0.161	0.025	0.182
Gemma	$0.708_{0.068}$	$0.778_{0.079}$	$0.746_{0.098}$	0.172	0.039	0.186
TinyLlama	$0.728_{0.073}$	$0.778_{0.091}$	$0.755_{0.089}$	0.168	0.053	0.168

Listing 1: Moderator Prompt Template with Model Suggestions

```
You will be given a paragraph from the text, which is: {textdescription}.

Definition of the code: A word or short phrase that symbolically
assigns a summative, salient, essence-capturing, and/or evocative
attribute for a portion of language-based or visual data.

Here is the excerpt to code:
{row['Paragraph']}

Here are three coding suggestions from previous models:
1. {row['Llama3_Code']}
2. {row['Falcon_Code']}
3. {row['Mistral_Code']}

Please suggest a code taking into account all these answers.
Output should be the code with no longer than 5 words.
```