Text2Story 2025

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

Angelina Parfenova
Technical University of Munich
Lucerne University of Applied Sciences and Arts
angelina.parfenova@tum.de

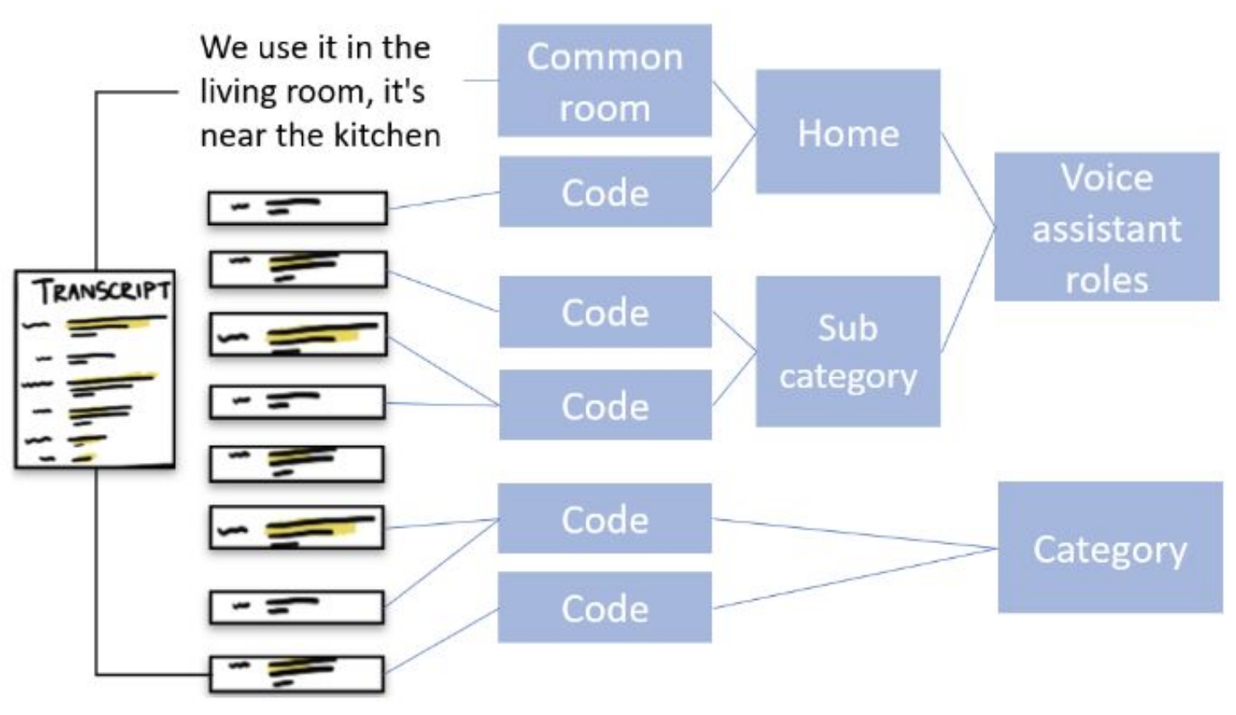
Motivation

Challenge: Holocaust children’s diaries contain deeply emotional and historically significant narratives, requiring nuanced interpretation.

Problem: Traditional qualitative coding is manual, time-consuming, inconsistent, and not scalable for large datasets.

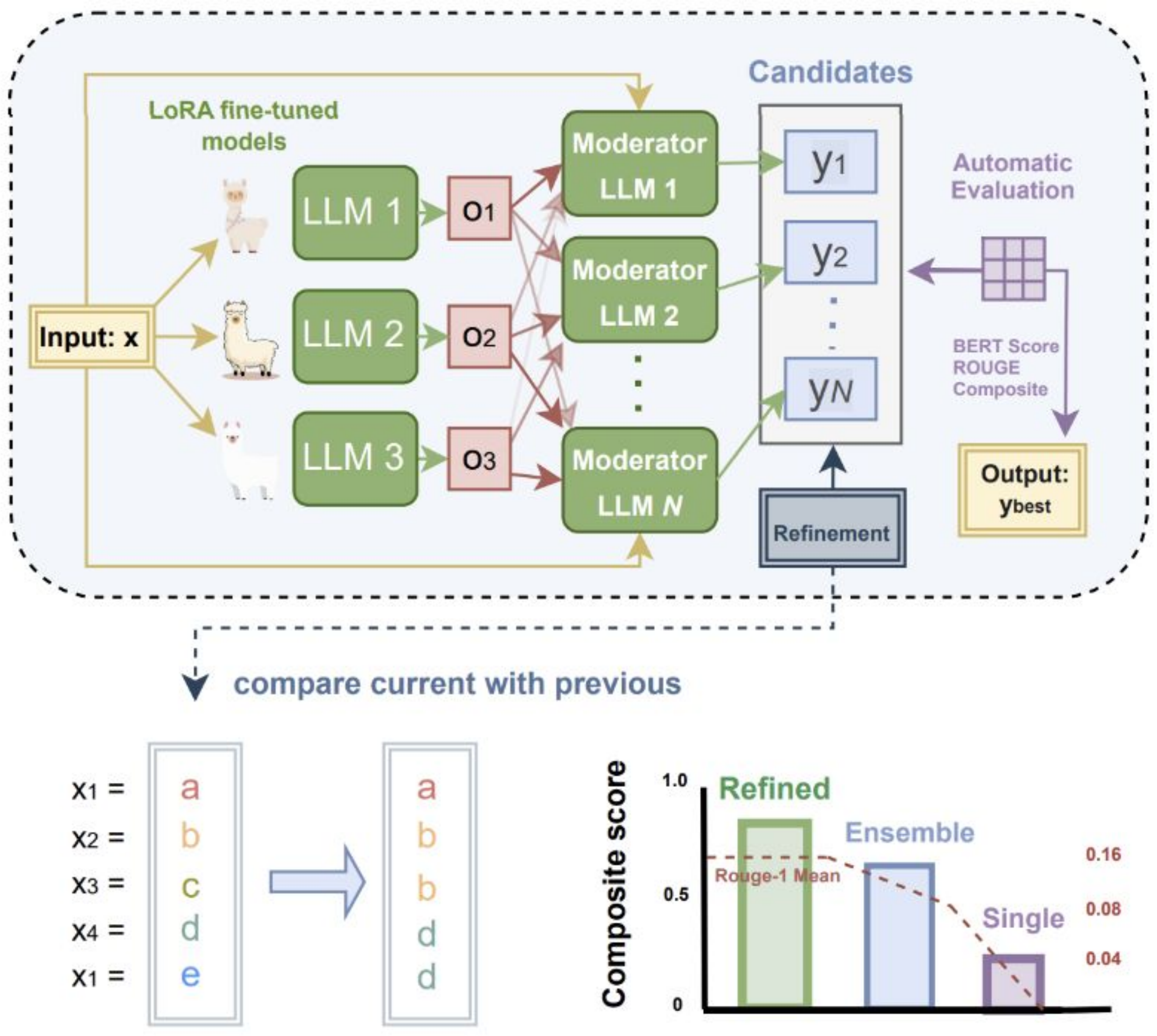
Opportunity: Advances in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) provide automation while preserving emotional depth and historical specificity.

Current inductive coding practice



In qualitative research, coding involves assigning a 'code' to each significant statement or dialogue segment within an interview to summarize its main idea. For instance, a citation such as "*I feel I can't stop using my fitness tracker, it is controlling my whole life*," would be associated with a label denoted as "*addiction*“.

Pipeline



First, an input x is processed independently by three smaller LLMs (7B and 8B parameter sizes). The outputs from these models are then evaluated by a set of N Moderators (Moderator_1, Moderator_2, … Moderator_N), which refine and consolidate the results. Finally, **code merging** is performed to ensure consistency across similar inputs, producing the optimal output.

Datasets

Code	Quote
Mobility	Well since it’s in a smartphone, it’s usually always with me.
Helper	For me the most important thing is to have an assistant like Google assistant, meaning it doesn’t talk to you itself, it only performs tasks.
Playing	For example I remember playing cities with her…

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

For the fine-tuning we propose utilizing a curated dataset comprising coded interview citations collected from social scientists, both academic researchers and students doing qualitative research in social science, augmented with SemEval data.

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.

Composite Score

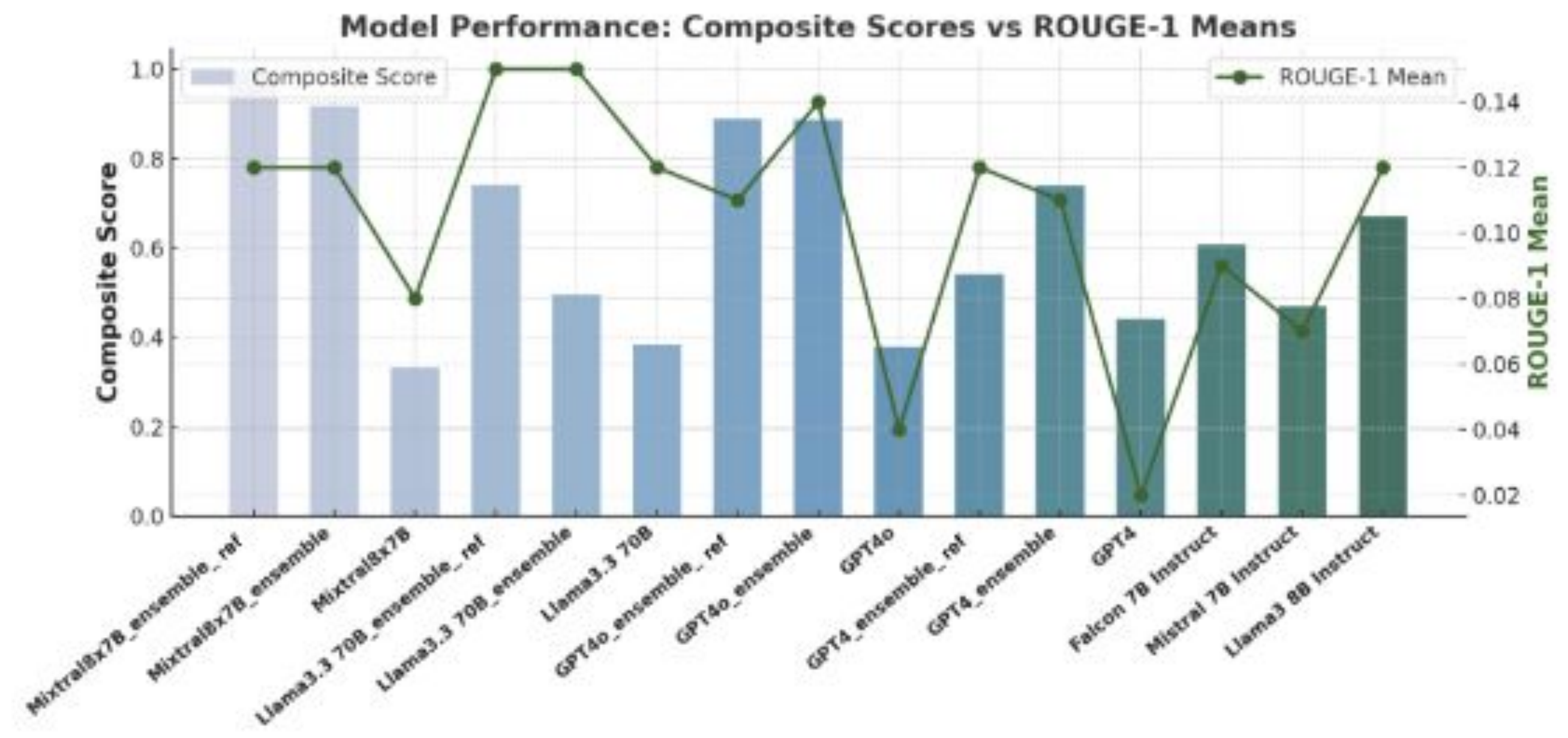
$$\mathscr{C} = \frac{1}{4}\left[\tilde{S}_c + \tilde{M} + (1-\tilde{L}) + (1-\tilde{J})\right]$$

To systematically assess the performance of ensemble models in inductive coding tasks, we introduce a composite score that integrates four evaluation scores: (1) cosine similarity, (2) METEOR score, (3) code length penalty, (4) Jensen-Shannon divergence.

Metric	Composite Score	BERTScore F1	ROUGE-1	ROUGE-L	Code Length	Human Rating
Composite Score	1.00	-0.06	0.51	0.44	-0.02	**0.73***
BERTScore F1	-0.06	1.00	-0.09	-0.11	**-0.65***	-0.14
ROUGE-1	0.51	-0.09	1.00	**0.99***	-0.35	0.49
ROUGE-L	0.44	-0.11	**0.99***	1.00	-0.32	0.44
Code Length	-0.02	**-0.65***	-0.35	-0.32	1.00	-0.41
Human Rating	**0.73***	-0.14	0.49	0.44	-0.41	1.00

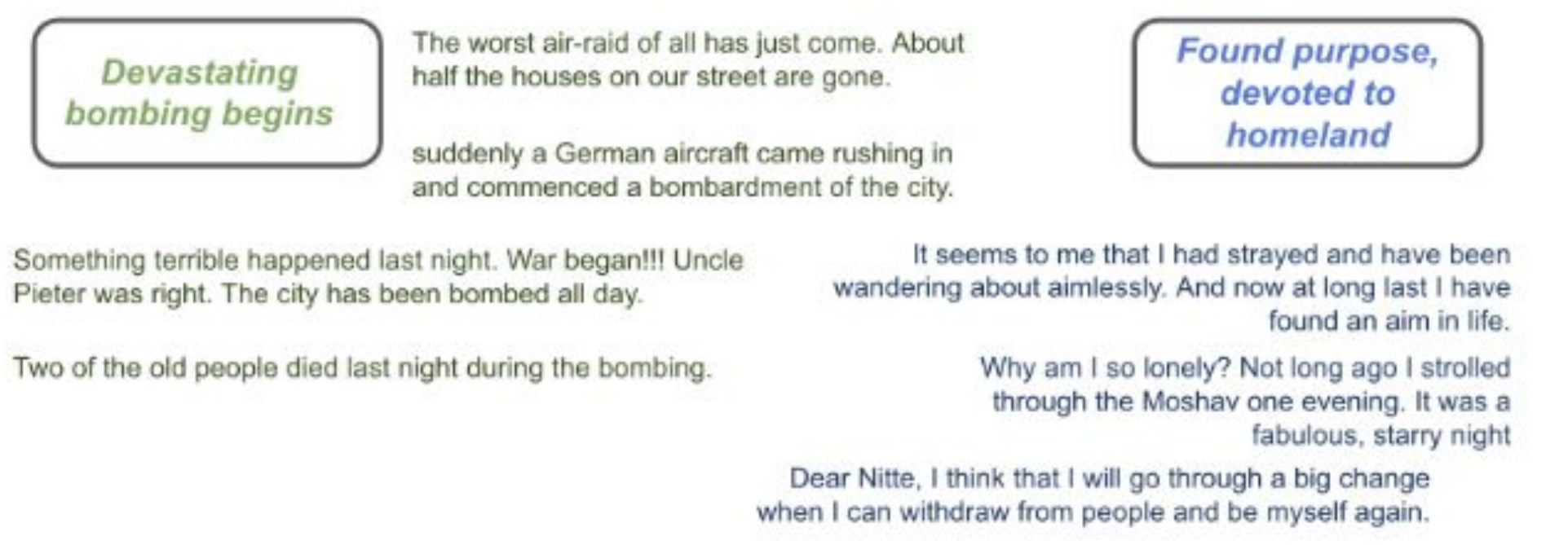
To assess the validity of the composite score, we conducted a human evaluation study where experts rated the quality of codes generated by anonymized models.

Experiments



Performance comparison of different models on inductive coding tasks, measured by Composite Scores (bars) and ROUGE-1 Mean (line). Ensemble models with refinement outperform individual models and non-refinement ensembles.

Results


Devastating bombing begins
Found purpose, devoted to homeland

The worst air-raid of all has just come. About half the houses on our street are gone.

suddenly a German aircraft came rushing in and commenced a bombardment of the city.

Something terrible happened last night. War began!!! Uncle Pieter was right. The city has been bombed all day.

Two of the old people died last night during the bombing.

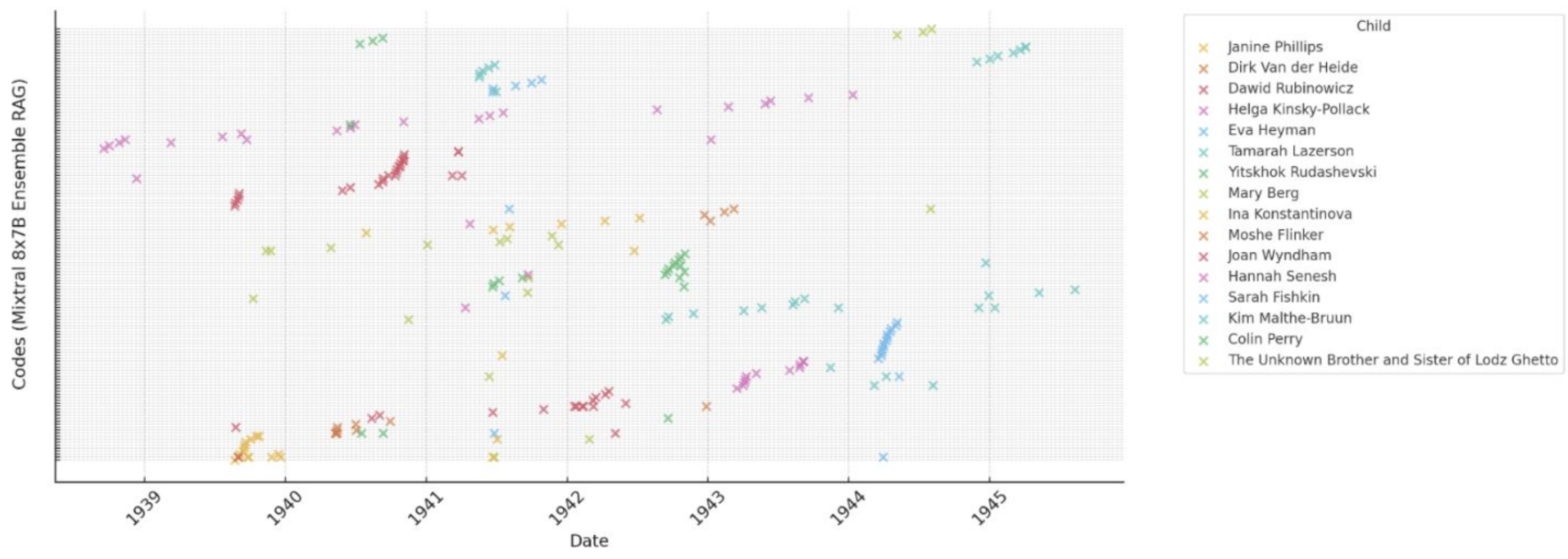
It seems to me that I had strayed and have been wandering about aimlessly. And now at long last I have found an aim in life.

Why am I so lonely? Not long ago I strolled through the Moshav one evening. It was a fabulous, starry night

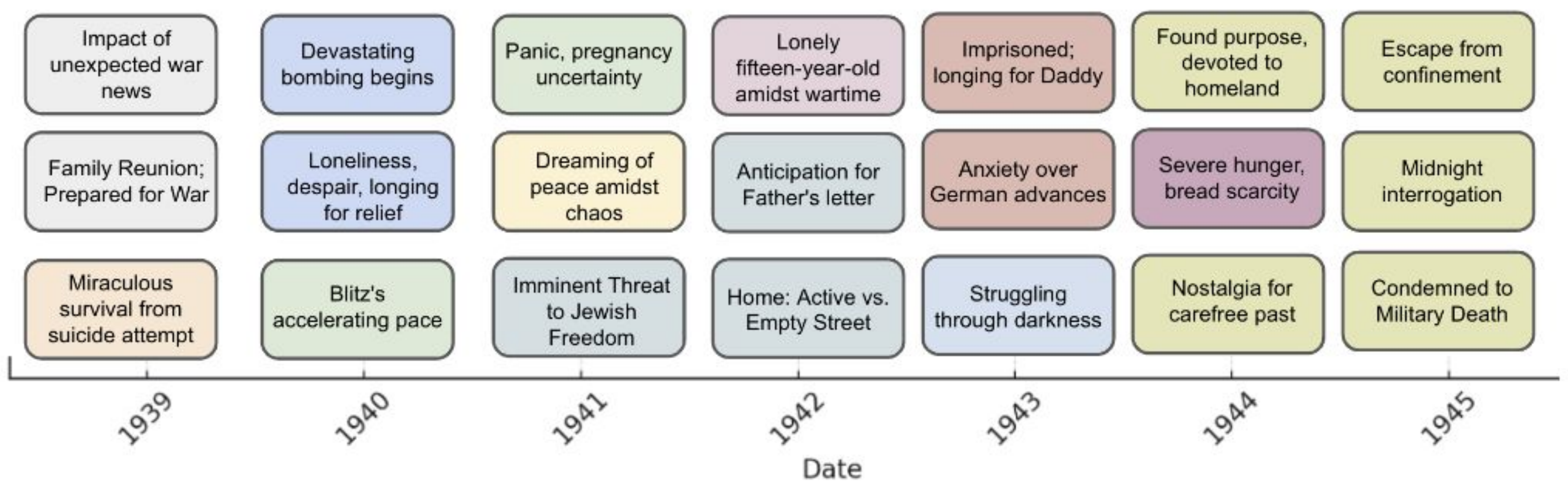
Dear Nitte, I think that I will go through a big change when I can withdraw from people and be myself again.

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

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.



The Figure 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.


Impact of unexpected war news
Family Reunion; Prepared for War
Miraculous survival from suicide attempt
Devastating bombing begins
Loneliness, despair, longing for relief
Blitz's accelerating pace
Panic, pregnancy uncertainty
Dreaming of peace amidst chaos
Imminent Threat to Jewish Freedom
Lonely fifteen-year-old amidst wartime
Anticipation for Father's letter
Home: Active vs. Empty Street
Imprisoned; longing for Daddy
Anxiety over German advances
Struggling through darkness
Found purpose, devoted to homeland
Severe hunger, bread scarcity
Nostalgia for carefree past
Escape from confinement
Midnight interrogation
Condemned to Military Death
1939
1940
1941
1942
1943
1944
1945
Date