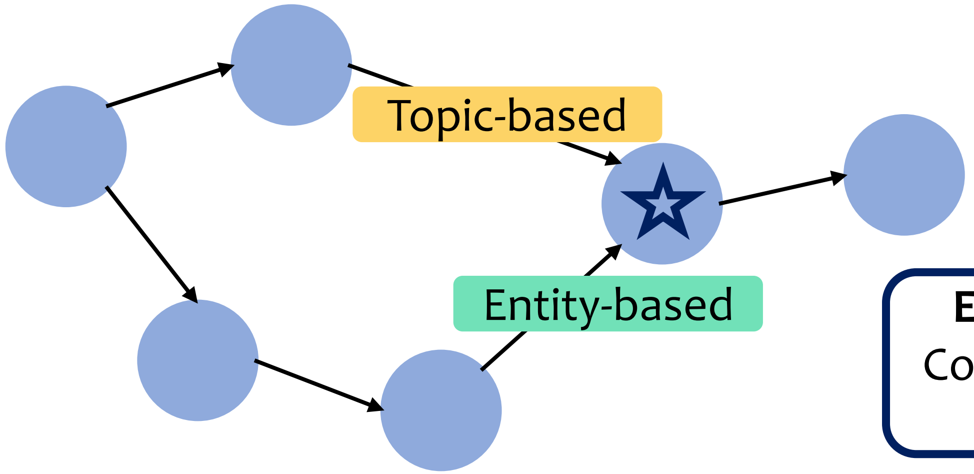
Text2Story'25 Workshop
Lucca, Italy – April 10, 2025

Explainable AI Components for Narrative Map Extraction

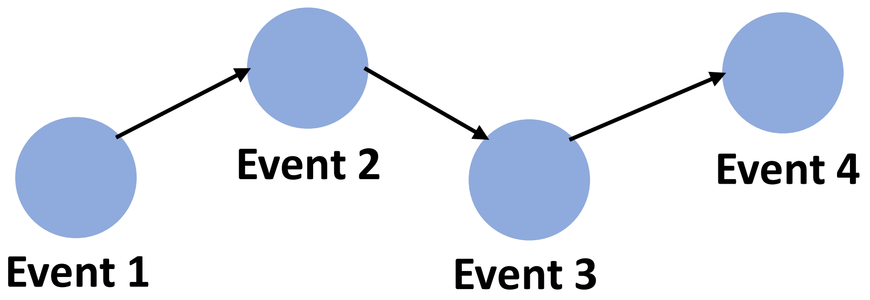
Brian Keith, Fausto German, Eric Krokos, Sarah Joseph, and Chris North



Explanations:
Connection Type
Key Events

Problem Context

What are Narrative Maps?



Graph-based representations that capture connections between events in a narrative

The "Black Box" Problem

Users don't understand why events are connected or how the narrative structure was determined

Our Challenge

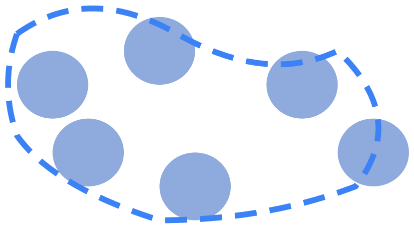
Provide meaningful explanations across multiple levels of abstraction to enhance human-AI collaboration.

Proposed XAI System Architecture

Multi-Level Explanation Approach

Low-Level

Document Relationships



Topical Clusters

- HDBSCAN + TF-IDF
- Keyword-based explanations
- UMAP visualization

Connections

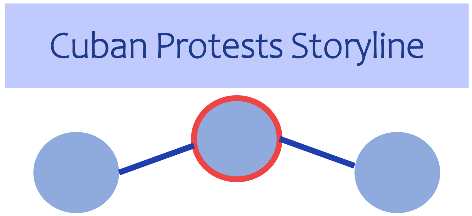
Event Relationships

Connection Labels

- Type of connection (low or high level).
- Detailed explanations with SHAP values

High-Level

Narrative Structure



Storyline Names and Important Events

- Naming algorithm
- Content-based event importance
- Structural-based event importance

Evaluation & Results

- User study with 10 participants.
- Analyzed data set of the 2021 Cuban protests.
- Insight-based evaluation approach.

Responses (5-point Likert Scale)

1 (“Strongly Disagree”) to 5 (“Strongly Agree”)

(a) Overall Explanations

	1	2	3	4	5	*Mean*	*SD*
Trust	0	0	1	3	6	**4.50**	*0.71*
Usefulness	0	0	1	5	4	**4.30**	*0.67*

(b) Important Events

	1	2	3	4	5	*Mean*	*SD*
Relevance	0	0	1	5	4	**4.30**	*0.67*
Usefulness	0	0	2	4	4	**4.20**	*0.79*

(c) Storyline Names

	1	2	3	4	5	*Mean*	*SD*
Correctness	1	4	1	2	2	**3.00**	*1.41*
Relevance	1	0	4	1	4	**3.70**	*1.34*
Usefulness	0	1	2	5	2	**3.80**	*0.92*

(d) Connections

	1	2	3	4	5	*Mean*	*SD*
Label Correctness	0	0	2	6	2	**4.00**	*0.67*
Label Usefulness	0	1	3	5	1	**3.60**	*0.84*
Explanation Usefulness	0	0	1	7	2	**4.10**	*0.57*
Comparison Usefulness	0	0	4	4	2	**3.80**	*0.79*

Key Insights

- Explanations significantly **increased user trust.**
- **Connection explanations** and **important events** were the most effective at building user confidence.

Conclusions

Key Contribution

Multi-level explanation framework that bridges low-level text processing and high-level narrative structures.

Practical Applications

- Journalism
- Intelligence Analysis
- Digital Humanities
- Crisis Management

Future Directions

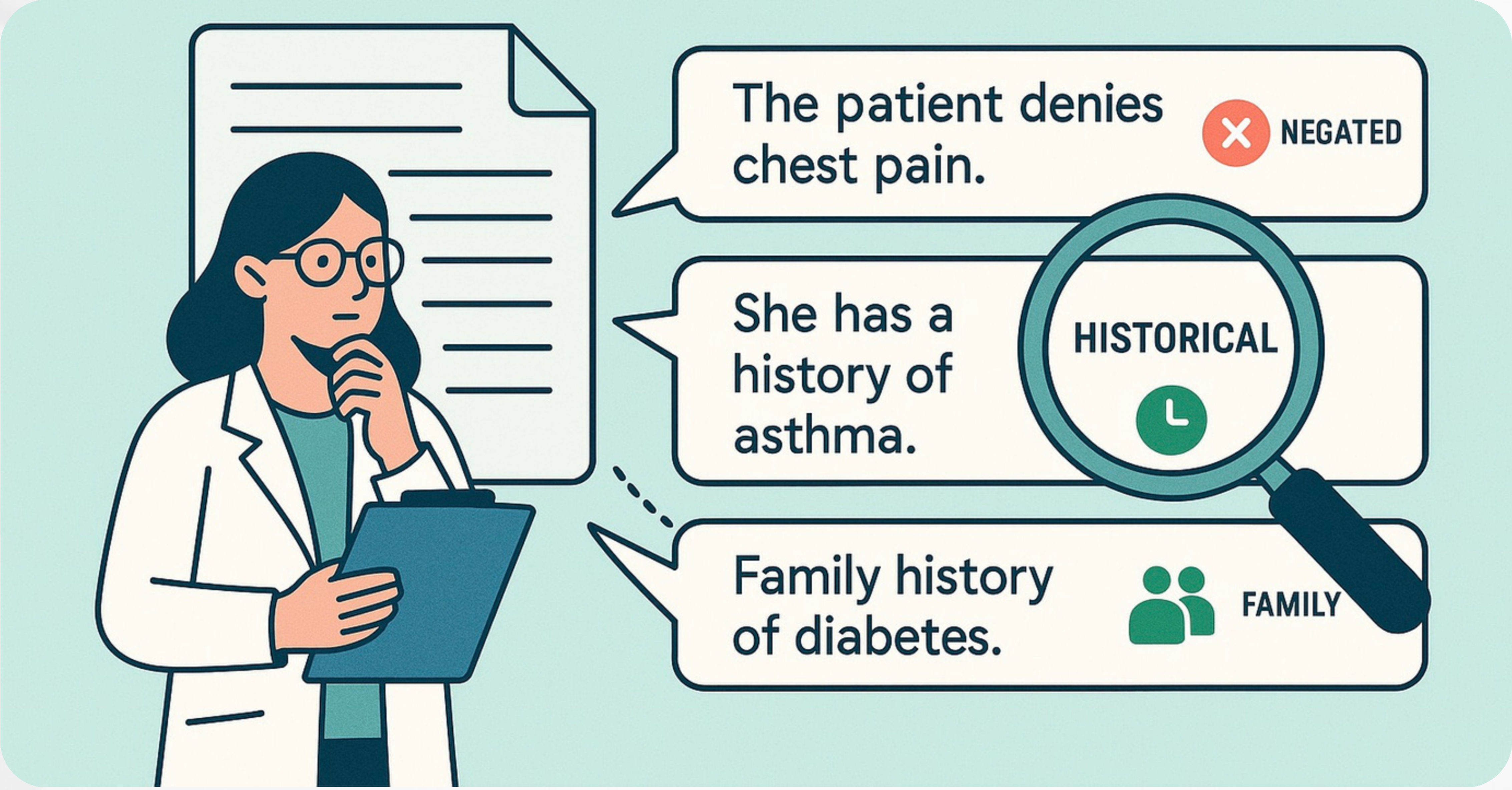
- **Causal explanations.**
- **Adaptive explanations** based on user needs.
- **Scalability improvements** for larger narrative collections.
- Using **LLMs** to generate explanations.
- Integration with **interactive narrative sensemaking tools**.

Thank you!

Beyond Negation Detection : Comprehensive Assertion Detection Models for Clinical NLP

Veysel Kocaman
Yigit Gul
Hasham Ui Haq
Cabir Celik
Mehmet Butgul
M. Aytug Kaya
David Talby

John Snow Labs Inc., Delaware, USA



Dataset Description

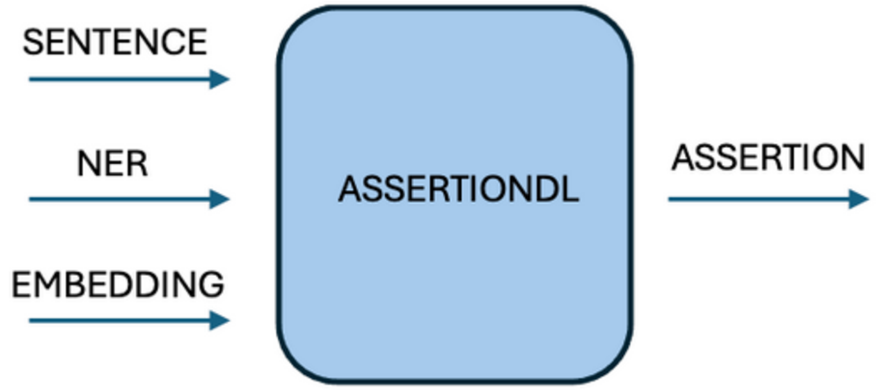
Text	Label	Description	Size
Overnight, the patient became **hypoxic** , dropping to the 80 's.	*present*	Confirms the presence of a medical condition.	8622
He gets **short of breath** with one flight of stairs.	*conditional*	Represents conditions that might occur under specific circumstances or conditions.	148
Small stroke, nearly recovered, likely **embolic from carotid artery** .	*possible*	Suggests uncertainty or potential presence of a condition.	652
There was no evidence of **diarrhea** during medical Lawrence Memorial Hospital stay.	*absent*	Indicates the negation or nonexistence of a medical condition.	2594
Mother suffer **MI** in her 50 's, died at age 59.	*associated with someone else (awse)*	Refers to medical conditions related to individuals other than the patient, such as family members.	131
Hydrocodone 5 mg with Tylenol , one to two tablets every four hours p.r.n. **pain** .	*hypothetical*	Denotes speculative or conjectural conditions that are not currently present.	445

The evaluation and benchmarking in this study are conducted exclusively on the official i2b2 dataset, which represents a comprehensive resource for assessing assertion detection frameworks in real-world clinical scenarios.

Methodology

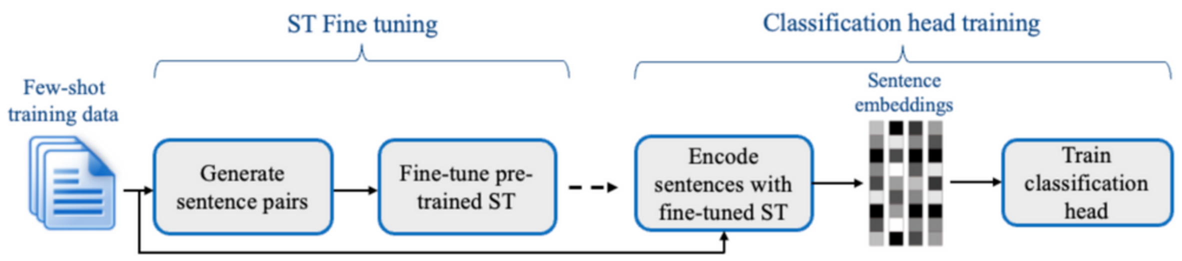
Assertion DL

AssertionDL is a Bi-LSTM-based classification model designed for assertion detection, built on a modified version of a previous architecture.



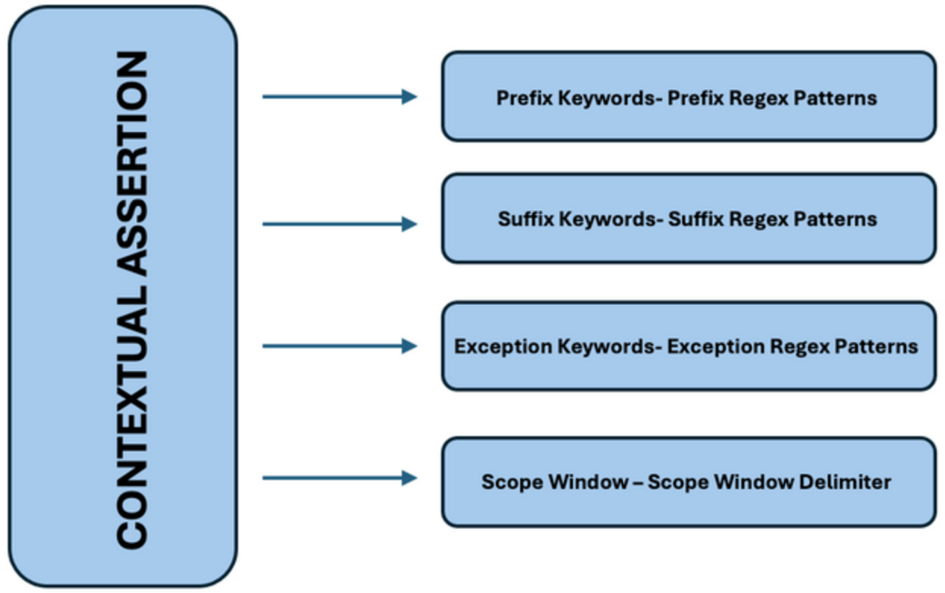
FewShot Assertion

FewShotAssertion is a transformer-based model built on a modified SetFit framework, leveraging sentence-transformer embeddings and contrastive learning for few-shot assertion detection.



Contextual Assertion

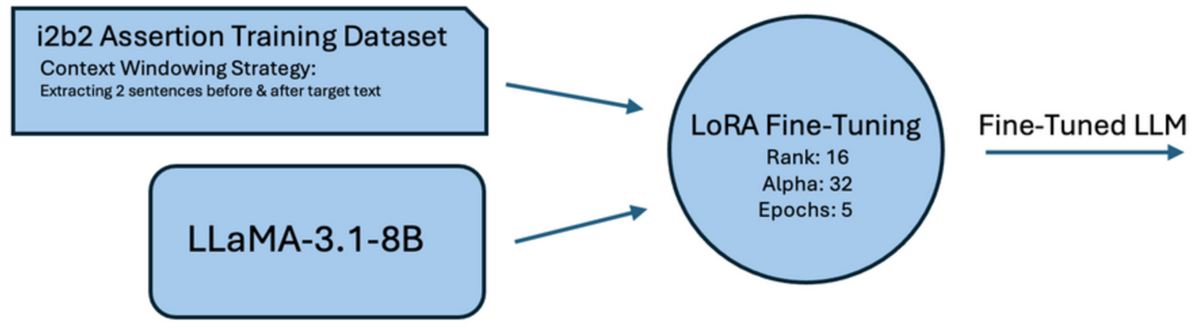
The Contextual Assertion module extends rule-based assertion detection by leveraging user-defined rules and contextual patterns



Methodology

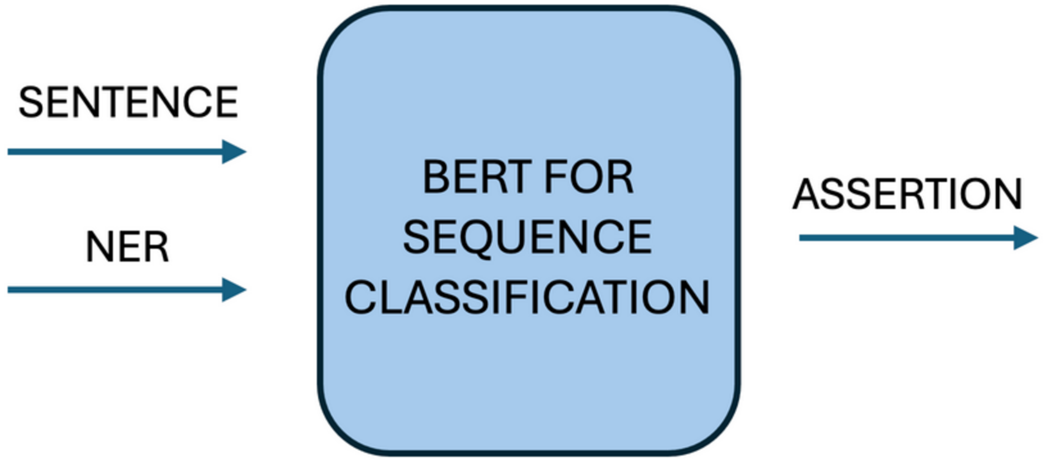
Fined-Tuned LLM

We fine-tuned LLama-3.1-8B on the i2b2 assertion training dataset using LoRA without quantization, ensuring parameter efficiency while preserving pre-trained knowledge.



Bert For Sequence Classifier

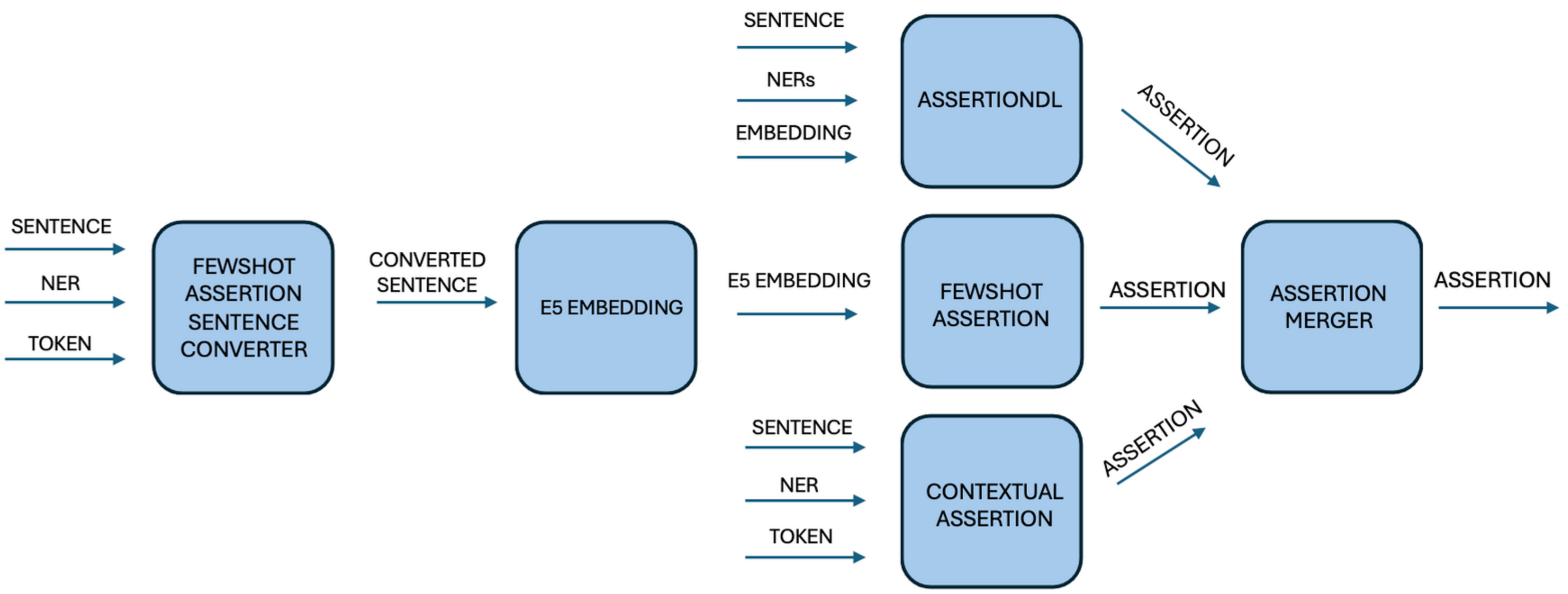
We implemented a transformer-based approach using BioBERT, a biomedical fine-tuned BERT model.



Methodology

Combined Pipeline

- The pipeline includes AssertionDL, FewShotAssertion, and Contextual Assertion models.
- A majority voting mechanism is applied to resolve conflicts in predictions across models.



Results

Model	present	absent	possible	hypothetical	conditional	awse*	weighted avg
*Combined Pipeline***	0.963	0.951	0.755	0.875	0.511	0.922	0.941
AssertionDL	0.941	0.898	0.672	0.761	**0.599**	0.886	0.907
FewShotAssertion	0.955	0.942	0.748	0.872	0.293	0.809	0.929
ContextualAssertion	-	0.929	0.708	-	-	0.835	0.883
Fine Tuned LLM	**0.976**	**0.975**	0.759	0.911	-	**0.943**	**0.962**
BFSC (BioBert)	0.975	0.972	**0.787**	0.918	0.590	0.913	0.957
GPT-4o	0.937	0.891	0.692	0.677	-	0.805	0.901
Azure Ai Text Analytics	-	0.761	0.583	0.763	0.569	0.800	0.727
AWS Med Comprehend	0.882	0.788	0.659	0.617	-	0.737	0.839
NegEx	-	0.897	-	-	-	-	0.897
BFSC latest best [11]	0.979	0.972	0.786	-	-	-	0.952
Prompt-based Bert [29]	0.971	0.968	0.763	**0.921**	0.485	0.875	0.951

Comparison of assertion models across various categories. Best performing model for each category is represented with bold characters. The models in the first section of this table are developed by JSL. In LLM and GPT-4o experiments, hypothetical and conditional labels are merged/treated as a single label.

Conclusion

- The study evaluates JSL's state-of-the-art assertion detection models, from lightweight DL models to fine-tuned LLMs.
- The fine-tuned LLM achieves the highest accuracy (96.2%), outperforming GPT-4o (90.1%) and commercial APIs, especially in Present, Absent, and Hypothetical assertions.
- However, the LLM is extremely costly, running 100× slower on a CPU and being thousands of times more expensive for just 1-2% better accuracy.
- Assertion DL, FewShot Assertion, and Bert For Sequence Classifier models offer efficient, competitive alternatives, with the Combined Pipeline (94.1%) outperforming all commercial solutions.
- Integrated with Spark NLP, these smaller, domain-specific models surpass GPT-4o, Azure AI, and AWS Medical Comprehend, providing scalable, cost-effective clinical NLP solutions.

Thank You

TEXT2STORY@ECIR'25

John Snow LABS

Can Zero-Shot Commercial API's Deliver Regulatory-Grade Clinical Text De-Identification ?

Veysel Kocaman, Muhammet Santas, Yigit Gul, Mehmet Butgul and David Talby

John Snow Labs Inc., Delaware, USA

Can Zero-Shot Commercial API's Deliver Regulatory-Grade Clinical Text De-Identification ?

Objective:

Evaluate Azure Health Data Services, AWS Comprehend Medical, Claude 3.7 Sonnet and Open AI GPT-4o against Healthcare NLP for PHI de-identification.

Dataset:

48 clinical documents annotated by medical experts.

Evaluation Metrics:

Assessed at both entity-level and token-level performance.

Results:

Healthcare NLP achieved the highest F1-score (96%) vs. Azure (91%), AWS (83%), and GPT-4o (79%).

Cost Efficiency:

Healthcare NLP reduces processing costs by over 80% compared to Azure and GPT-4o, thanks to its fixed-cost, local deployment model.

Key Findings:

Commercial zero-shot APIs fall short in accuracy, adaptability, and cost-efficiency. Healthcare NLP offers superior performance, customization, and scalability

Methodology

Original Text	Masked	Obfuscated
Harbor Hospital --------------------------------- 36 Park Avenue, 95108, San Diego, CA, USA Email: medunites@harborhospital.com, Phone: (818) 342-7353. TSICU MRN# 1482928 on 24/06/2019 by ambulance VIN: 1HGBH41JXMN109186. John Davies is a 62 y.o. patient admitted to ICU after an MVA on 22 Hoyt Street, at 23:00 hours. He works as a driver, and long hours of work reported. He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. Davies was seen at 23:12 minutes by attending physician Dr. Meyer Lorand and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. Frank M and was HD stable with normal vital signs and therefore and transferred (ID num 184378) for further radiological investigations.	<HOSPITAL> --------------------------------- <STREET>, <ZIP>, <CITY>, <STATE>, <COUNTRY> Email: <EMAIL>, Phone: <PHONE>. TSICU MRN# <MEDICALRECORD> on <DATE> by ambulance VIN: <VIN>. <PATIENT> is a <AGE> y.o. patient admitted to ICU after an MVA on <STREET>, at 23:00 hours. He works as a <PROFESSION>, and long hours of work reported. He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. <PATIENT> was seen at 23:12 minutes by attending physician Dr. <DOCTOR> and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. <DOCTOR> and was HD stable with normal vital signs and therefore and transferred (ID num <IDNUM>) for further radiological investigations.	MERCY HOSPITAL ARDMORE --------------------------------- 474 North Yellow Springs Street, 14235, Salt Lake City, Utah, US Email: dalton@mercyhospital.com, Phone: (765) 896 92 86. TSICU MRN# US:3025146 on 15/08/2019 by ambulance VIN: 1AAAA00AAAA111000. Meldon Lemon is a 58 y.o. patient admitted to ICU after an MVA on 390 40th street, at 23:00 hours. He works as a special educational needs teacher, and long hours of work reported. He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. Lemon was seen at 23:12 minutes by attending physician Dr. Evangeline Kelly and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. Lara Courier and was HD stable with normal vital signs and therefore and transferred (ID num 453267) for further radiological investigations.

De-Identification process identifies potential pieces of content with personal information about patients and removes them by replacing them with semantic tags or fake entities.

Methodology

Comprehensive Healthcare NLP Solution:

Healthcare NLP (Spark NLP) offers 2,500+ pre-trained models for medical text processing, including NER, information extraction, and clinical text analysis. The library provides advanced PHI de-identification using NER models, ensuring compliance with privacy regulations while maintaining data utility for research.

Azure Health Data Services:

Uses NLP to identify, label, redact, and surrogate PHI in clinical texts, ensuring compliance with HIPAA, GDPR, and CCPA.

Amazon Comprehend Medical:

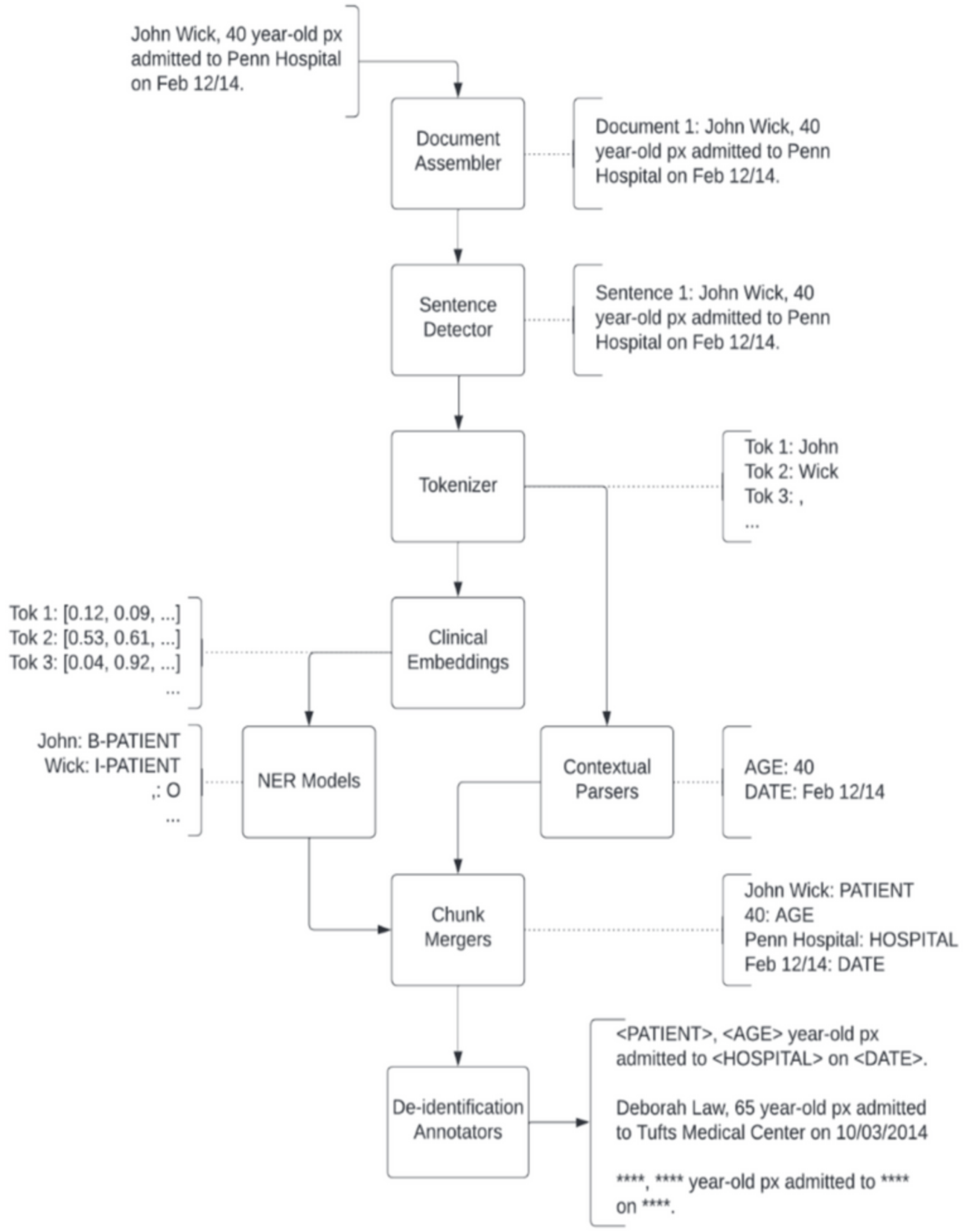
HIPAA-compliant NLP service that extracts medical entities and PHI from clinical texts using machine learning, aiding data automation and workflow optimization.

Open AI GPT-4o:

Multimodal model with improved classification accuracy and response times compared to GPT-4, potentially enhancing PHI identification and redaction via prompting. While previous GPT models have been studied for medical text de-identification, GPT-4o's effectiveness in this area remains unverified due to the lack of empirical evaluations.

Anthropic Claude 3.7 Sonnet:

Mid-tier model from Anthropic, balances speed and accuracy, making it a strong candidate for healthcare AI. It demonstrates high contextual awareness in PHI extraction, effectively identifying and categorizing patient data.

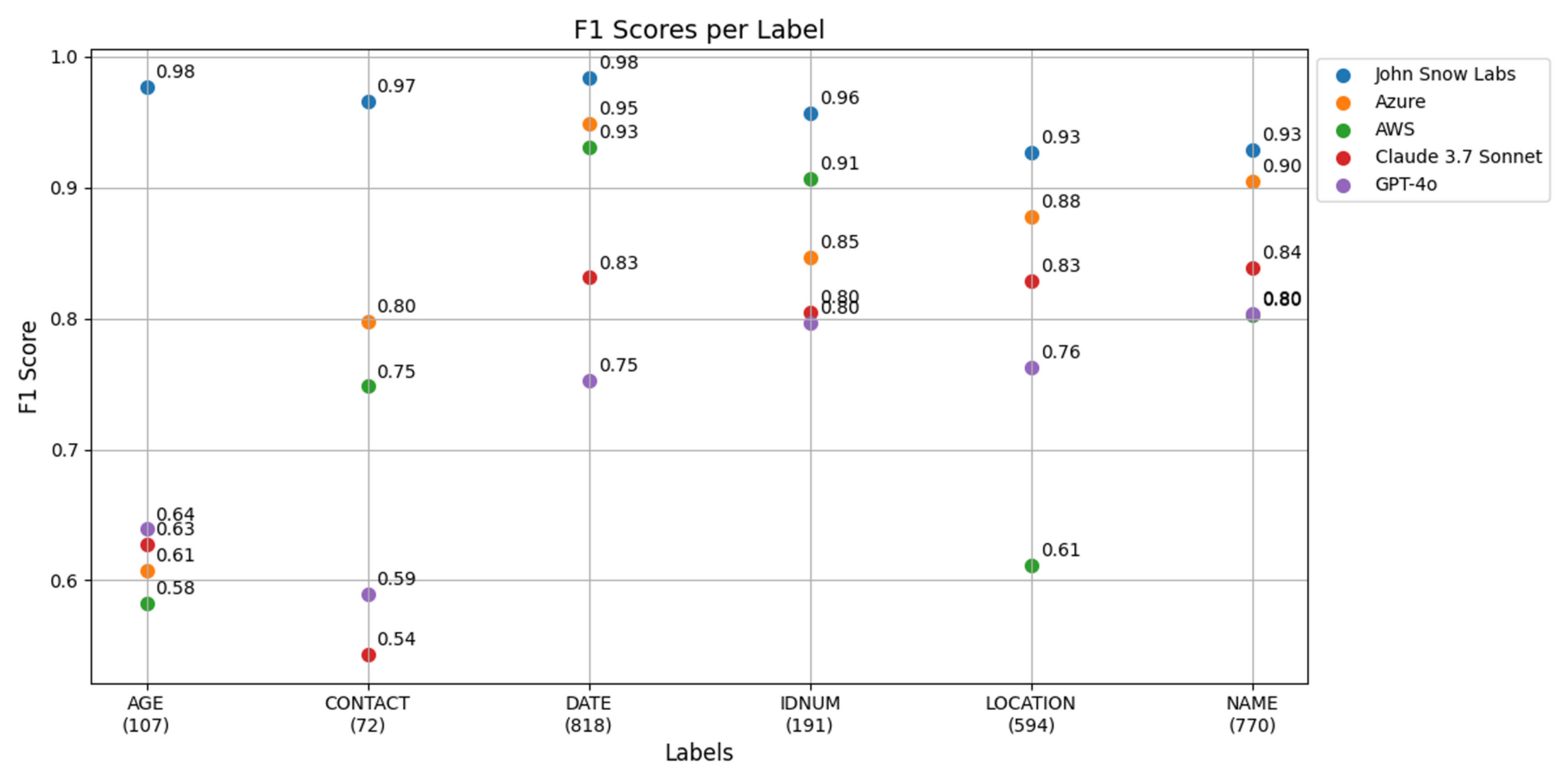


Results

Metric / Entity	Healthcare NLP			Azure			Amazon			GPT-4o			Claude 3.7 Sonnet		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
AGE	0.96	1.00	**0.98**	0.94	0.45	0.61	1.00	0.41	0.58	0.87	0.50	0.64	0.73	0.55	0.63
CONTACT	0.96	0.97	**0.97**	0.73	0.88	0.80	0.78	0.72	0.75	0.67	0.53	0.59	0.74	0.43	0.54
DATE	0.97	0.99	**0.98**	0.91	0.99	0.95	0.90	0.97	0.93	0.79	0.72	0.75	0.84	0.83	0.83
IDNUM	0.98	0.94	**0.96**	0.78	0.93	0.85	0.95	0.86	0.91	0.70	0.92	0.80	0.70	0.95	0.80
LOCATION	0.93	0.92	**0.93**	0.89	0.87	0.88	0.52	0.74	0.61	0.82	0.72	0.76	0.79	0.87	0.83
NAME	0.92	0.94	**0.93**	0.92	0.89	0.90	0.85	0.76	0.80	0.79	0.82	0.80	0.82	0.86	0.84
O	1.00	1.00	**1.00**	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Macro Avg	0.96	0.97	**0.96**	0.88	0.86	0.85	0.86	0.78	0.80	0.80	0.74	0.76	0.80	0.78	0.78
Non-PHI	1.00	1.00	**1.00**	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
PHI	0.96	0.97	**0.96**	0.91	0.92	0.91	0.81	0.85	0.83	0.81	0.77	0.79	0.81	0.84	0.83
Macro Avg	0.98	0.98	**0.98**	0.95	0.96	0.95	0.90	0.92	0.91	0.90	0.88	0.89	0.90	0.92	0.91
cost per 1M doc	**$2,418**			$13,125			$14,525			$21,400			$23,330		

Healthcare NLP, Azure, Amazon GPT-4o and Claude 3.7 Sonnet PHI Recognition and Benchmark Comparison (Sample size: 45172 PHI entities).

Clinical Text De-Identification of PHI Data



Conclusion

Performance Comparison:

Healthcare NLP achieved the highest accuracy, followed by Azure Health Data Services, Amazon Comprehend Medical, Claude 3.7 Sonnet and GPT-4o.

Evaluation Metrics:

Healthcare NLP ranked highest in precision, recall, and F1-score across both entity-level and token-level evaluations.

Adaptability & Customization:

Healthcare NLP allows pipeline modifications, while other solutions are black-box APIs with no customization.

Cost-Effectiveness:

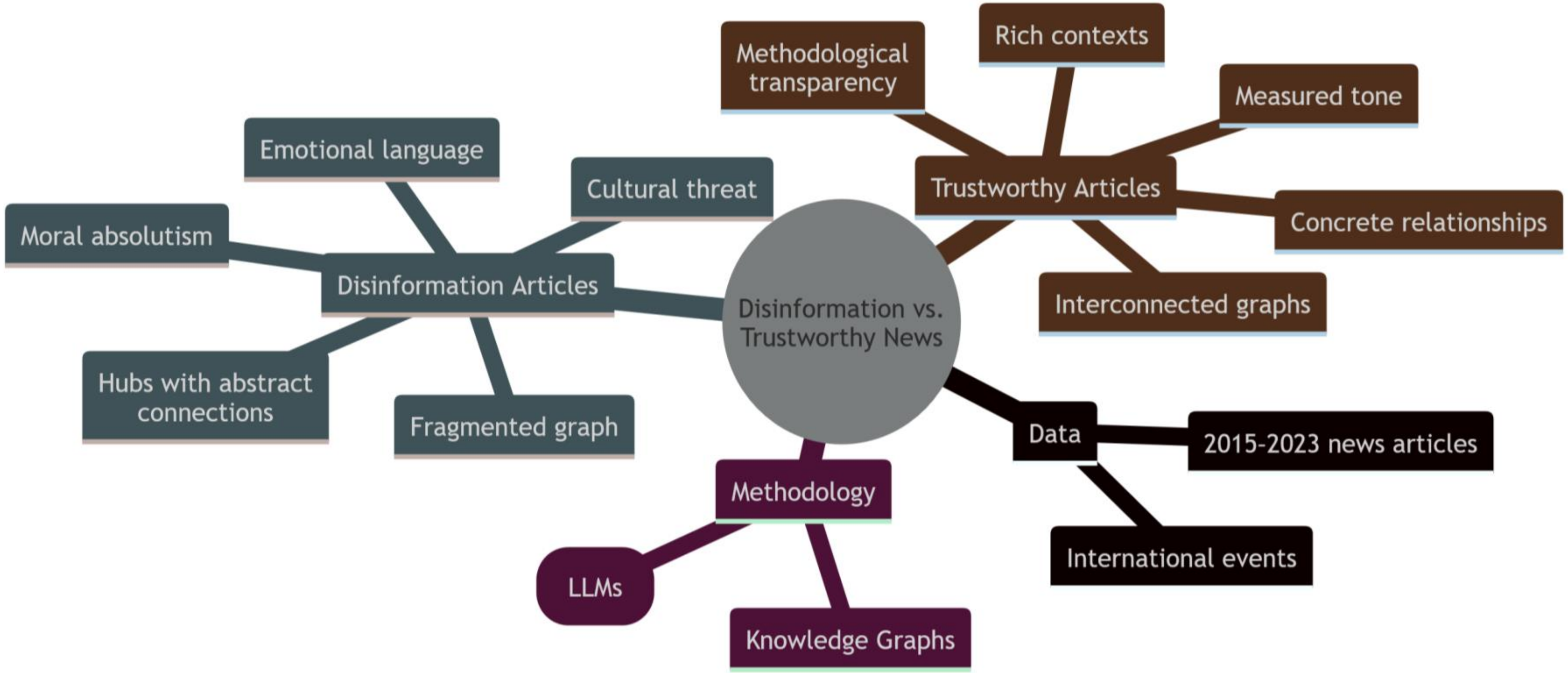
Healthcare NLP enables fixed-cost, local deployment, whereas cloud-based solutions have per-request pricing that scales with data volume.

Final Verdict:

Healthcare NLP outperformed alternatives by 5-10%, offering superior accuracy, flexibility, and cost-efficiency.

Thank You

Disinformation vs. Trustworthy News: A Knowledge Graph-Based Analysis of Narrative and Framing Patterns



Justina Mandravickaitė
Vytautas Magnus University
justina.mandravickaite@vdu.lt

Weaving Knowledge: Building Narratives with Ontologies for Heritage Crafts

Nicolò Pratelli and Valentina Bartalesi

Institute of Information Science and Technologies "Alessandro Faedo" (ISTI) of the National Research Council of Italy (CNR), Italy

CRAEFT

care, judgment, dexterity

Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

How can a 200-year-old porcelain technique live forever?

By turning **narratives** into **knowledge graphs**.

With the **Craeft Ontology**, we preserve the who, what, where, and why of heritage crafts: **semantically structured, visually engaging, and future-proof**.

🔍 Explore craft traditions as never before: visualize, query, and reshape their narratives.

Let's talk at the poster!

Newsletter-Factory: A Thematic Newsletter Generation Tool for Business Insights

- Newsletter:
  - Periodic publication of news
  - Based on a specific theme
  - Personalized stories
- Benefits for enterprises:
  - Track evolving business & tech.
  - Contribute to success & growth
  - Reduce information overload
- Newsletter-Factory Tool:
  - Automation tool
  - Near-duplicate detection
  - Metadata classification
  - Fine-grained classification using sub-categories
  - Human-in-the-loop for quality check and model improvement

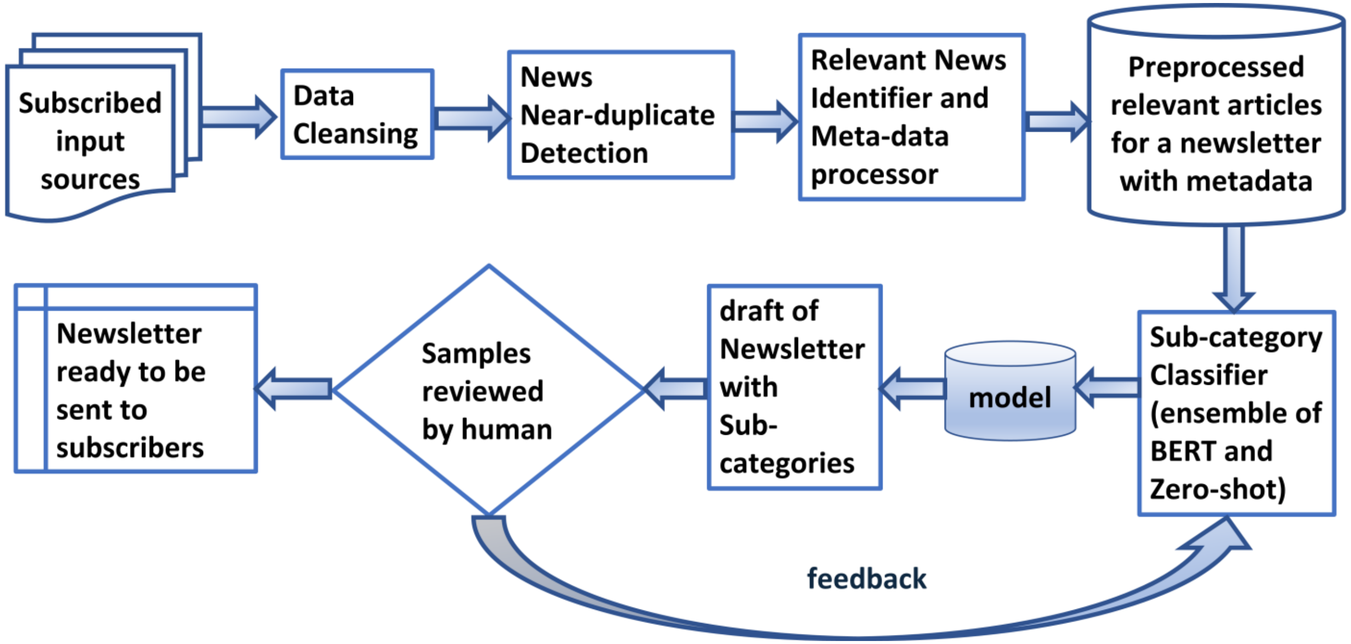


Figure: High-level block diagram of key components in Newsletter-Factory

tcs Research

TATA

Key Components of Newsletter-Factory Tool

- Near-duplicate detection and handling
  - Locality Sensitive Hashing
  - MinHash and Shingling
- Metadata classification
  - Pattern-based categorizers
  - ML classifiers
  - Identify newsletter theme
- Sub-category classification
  - Keyword-based classifier
  - Adapter fine-tuned BERT classifier
  - BART zero-shot classifier
- Human-in-the-loop
  - Inspection of draft
  - Ability to provide feedback
  - Model improvement using feedback

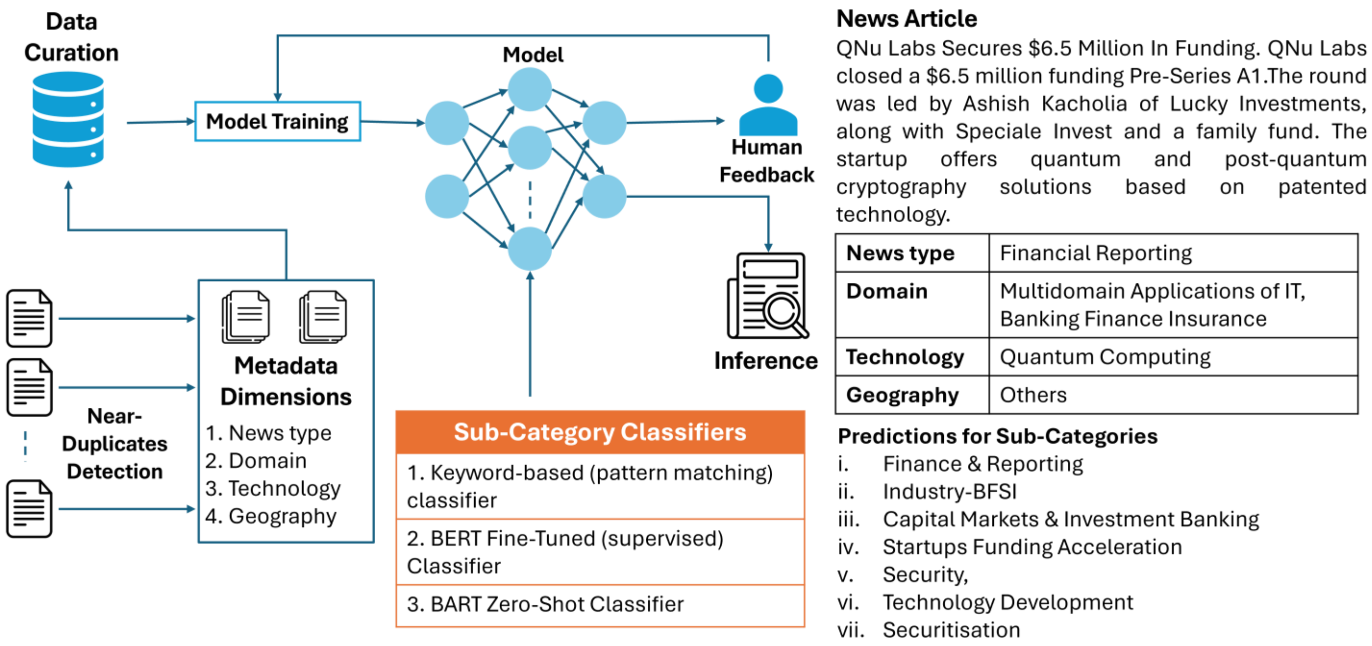


Figure: Illustrative example of sub-category classification step for enrichment of raw news data

tcs Research

TATA

tcs Research

TATA

Newsletter-Factory: A Thematic Newsletter Generation Tool for Curating Business Insights

Authors: Siddarth Tumre, Alok Kumar, Ajay Phade, Nihar Riswadkar, Sangameshwar Patil

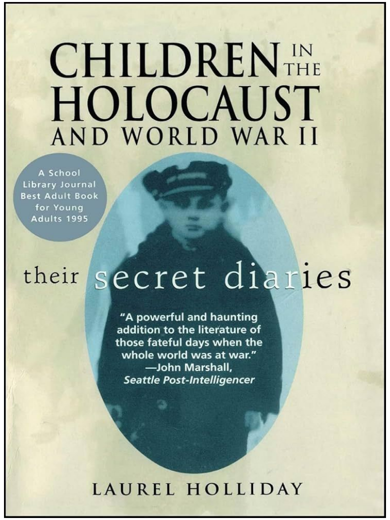
Presenter: Ajay Phade

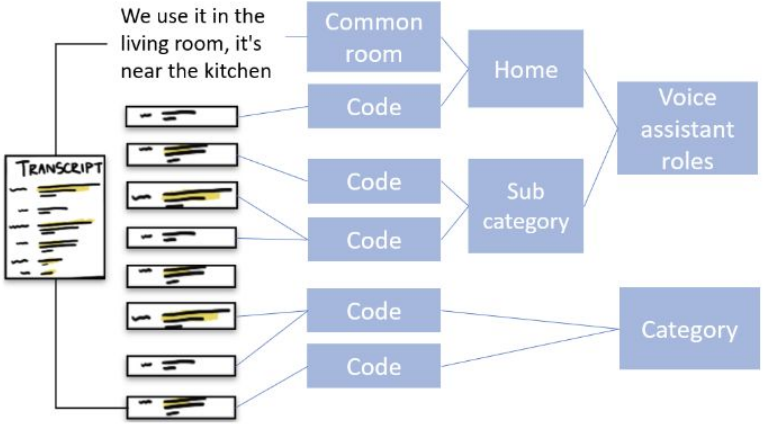
April 10th, 2025

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

Text2Story 2025

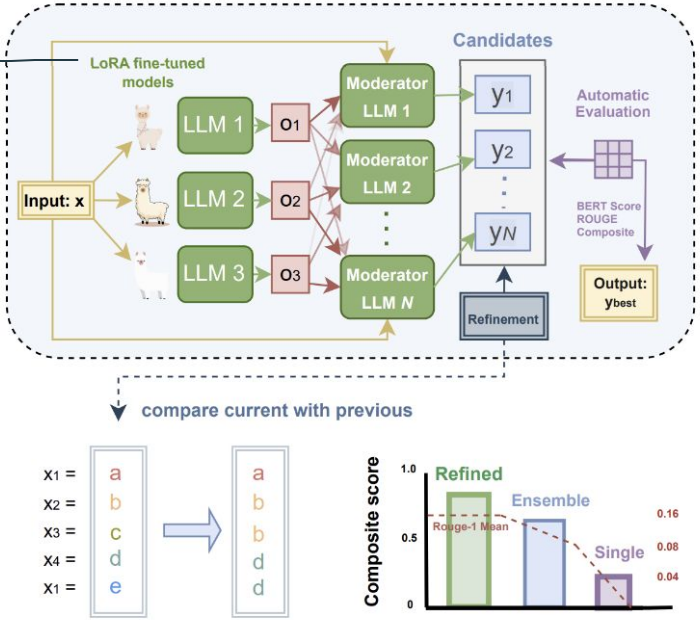
Qualitative coding



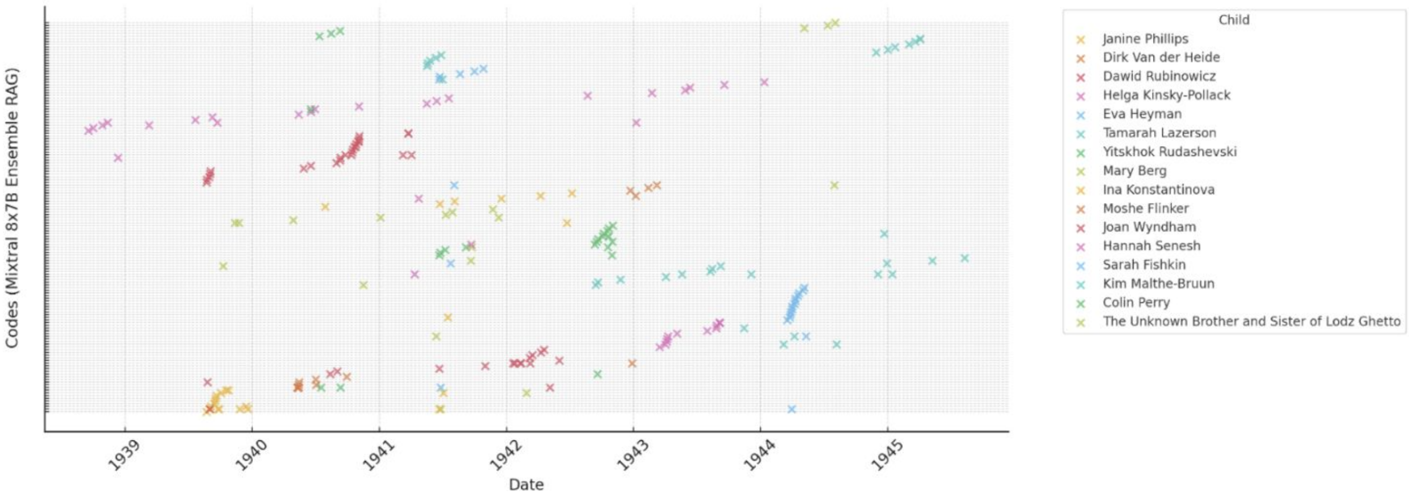


Moderator based framework

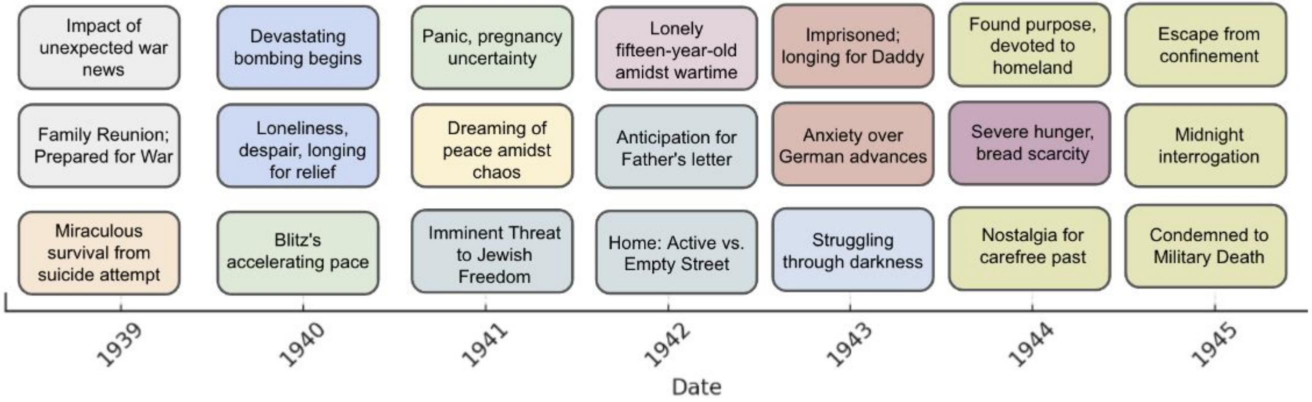
N Quotes	Description
Social Science Studies Data: 600 quotes	
78	Study about interaction with self-tracking devices (interviews)
22	Study about life transitions and mobility (interviews)
82	Study about interaction with voice assistants (interviews)
28	Study about museums and cultural experiences (interviews)
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



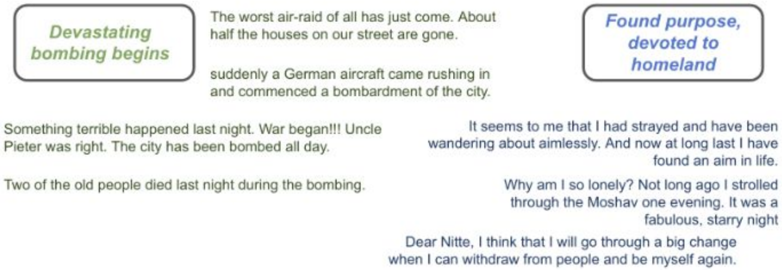
Results (2)


Codes (Mixtral 8x7B Ensemble RAG)
Date
1939
1940
1941
1942
1943
1944
1945
Child
Janine Phillips
Dirk Van der Heide
Dawid Rubinowicz
Helga Kinsky-Pollack
Eva Heyman
Tamarah Lazerson
Yitskhok Rudashevski
Mary Berg
Ina Konstantinova
Moshe Flinker
Joan Wyndham
Hannah Senesh
Sarah Fishkin
Kim Malthe-Bruun
Colin Perry
The Unknown Brother and Sister of Lodz Ghetto

Codes timeline



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

Monitoring narratives about the energy transition in Germany

Social (in)justice in the energy transition — from the digital debate to the living world (dissemination paper)

Jonas Rieger, **Lars Grönberg**, Carmen Loschke, Sibylle Braungardt

Text2Story'25
held in conjunction with ECIR 2025
Lucca, Italy

10.04.2025

Supported by:

Federal Ministry
for Economic Affairs
and Climate Action

on the basis of a decision
by the German Bundestag

tu dortmund university

fakultät statistik

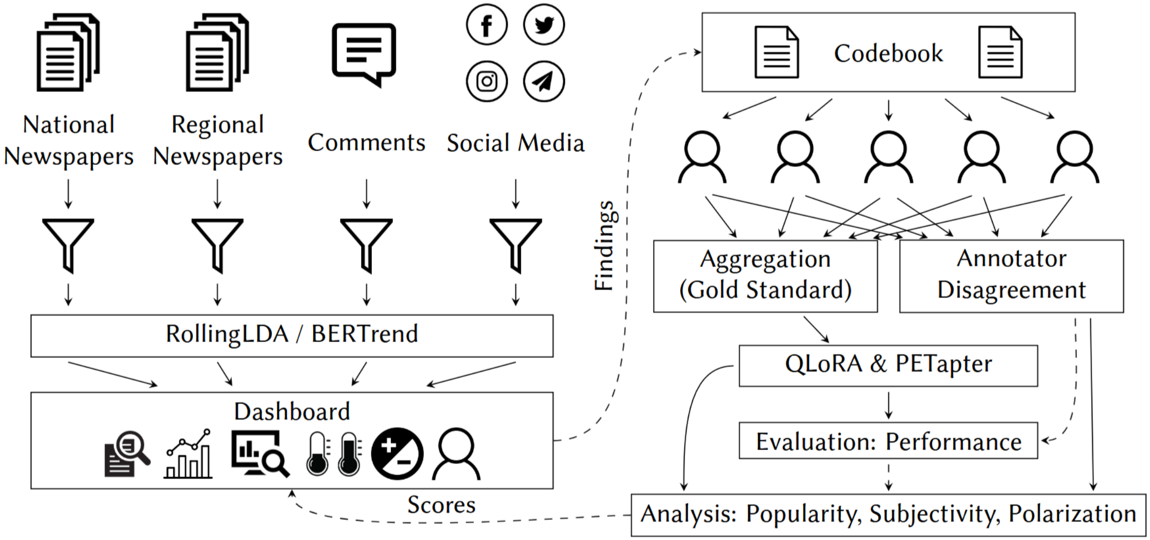
Öko-Institut e.V.
Institut für angewandte Ökologie
Institute for Applied Ecology

Our Research Goal and hypotheses

We aim to find evidence for the following hypotheses:

- The (media-perceptible) acceptance for the German energy transition dropped in the last two years (2023, 2024).
- The negative aspects regarding the energy transition are disseminated disproportionately often by certain media.
- The argument of social (in)justice is more often used as a counter-argument for the German energy transition as a whole than as an argument for a more socially just implementation of it.

Our complementary framework of topic modeling and classification



Discuss with us at our poster — and beyond

There are still **several challenges** to be solved (in particular, around the process of narrative extraction), for which we would be interested to exchange opinions and possible solutions.

Hence, **stop by our poster** and **discuss with us** — even after the poster session, e.g., via the GitHub issue tracker of our project!

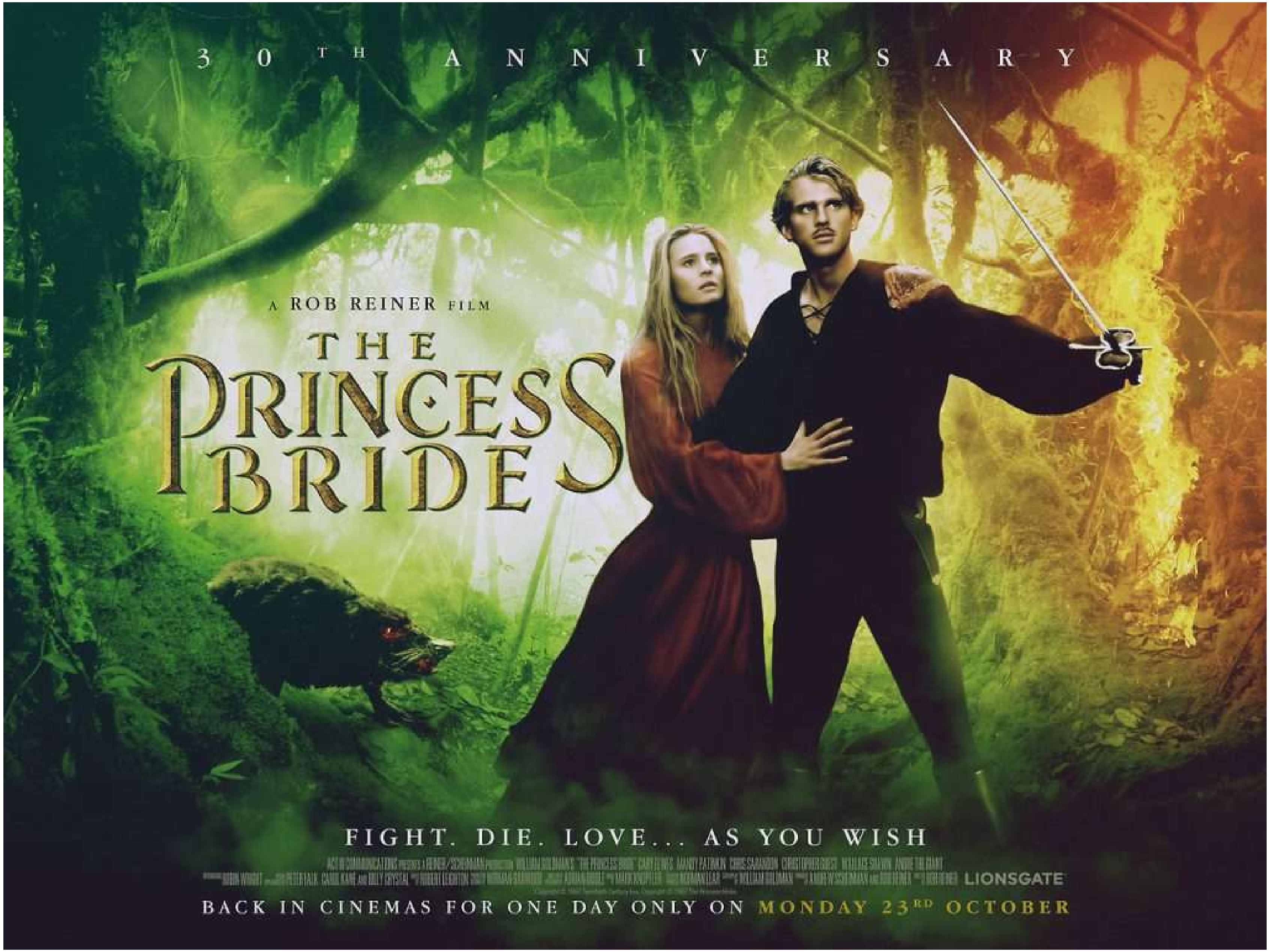
`https://github.com/LarsG321/Dissemination-Monitoring-narratives-about-the-energy-transition-in-Germany`

Accounting for the Importance of Changes in Event Actuality in the Representation of Narrative

Pablo Gervás, José Luis López Calle

Universidad Complutense de Madrid

Text2Story 2025
Eighth International Workshop on Narrative Extraction from Texts,
47th European Conference on Information Retrieval,
April 10th, 2025 - Lucca, Italy


30TH ANNIVERSARY
A ROB REINER FILM
THE PRINCESS BRIDE
FIGHT. DIE. LOVE... AS YOU WISH
LIONSGATE
BACK IN CINEMAS FOR ONE DAY ONLY ON MONDAY 23RD OCTOBER

Action Unit ID	Character	On Screen Action Dialogue Line Action	Truth Validation Status	Verifying Lapse	Conflicting AUID
1082	Westley	What was that?			
1083	Buttercup	If we surrender,	1099	16	
1084	Buttercup	and I return with you,	1106	22	
1085	Buttercup	will you promise	1087	2	
1086	Buttercup	not to hurt this man?	-1485	399	
1087	Humperdinck	May I live a thousand years and	-2125	1038	
1088	Humperdinck	never hunt again.	-2125	1037	
1089	Buttercup	He is a sailor on the pirate ship "Revenge."	923	-166	
1090	Buttercup	Promise	1092, 1093	2	
1091	Buttercup	to return him to his ship.	1092, 1093	2	
1092	Humperdinck	I swear	-1119	27	
1093	Humperdinck	it will be done.	-1119	26	
1094	Humperdinck	Once we're out of sight,			1092, 1093
1095	Humperdinck	take him back to Florin and	1119	24	1092, 1094
1096	Humperdinck	throw him in the Pit of Despair.	1119	23	1092, 1095
1097	Rugen	I swear	1118	21	
1098	Rugen	it will be done.	1118	20	

Action Unit ID	Character	On Screen Action Dialogue Line Action	Truth Validation Status	Verifying Lapse	Conflicting AUID	
1082	Westley	What was that?				
1083	Buttercup	If we surrender,	1099	16		
1084	Buttercup	and I return with you,	1106	22		
1085	Buttercup	will you promise	1087	2		REQUEST
1086	Buttercup	not to hurt this man?	-1485	399		
1087	Humperdinck	May I live a thousand years and	-2125	1038		PROMISE
1088	Humperdinck	never hunt again.	-2125	1037		
1089	Buttercup	He is a sailor on the pirate ship "Revenge."	923	-166		
1090	Buttercup	Promise	1092, 1093	2		COMMAND
1091	Buttercup	to return him to his ship.	1092, 1093	2		
1092	Humperdinck	I swear	-1119	27		PROMISE
1093	Humperdinck	it will be done.	-1119	26		
1094	Humperdinck	Once we're out of sight,			1092, 1093	
1095	Humperdinck	take him back to Florin and	1119	24	1092, 1094	
1096	Humperdinck	throw him in the Pit of Despair.	1119	23	1092, 1095	
1097	Rugen	I swear	1118	21		
1098	Rugen	it will be done.	1118	20		

Action Unit ID	Character	On Screen Action Dialogue Line Action	Truth Validation Status	Verifying Lapse	Conflicting AUID
1082	Westley	What was that?			
1083	Buttercup	If we surrender,	1099	16	
1084	Buttercup	and I return with you,	1106	22	
1085	Buttercup	will you promise	1087	2	
1086	Buttercup	not to hurt this man?	-1485	399	
1087	Humperdinck	May I live a thousand years and	-2125	1038	
1088	Humperdinck	never hunt again.	-2125	1037	
1089	Buttercup	He is a sailor on the pirate ship "Revenge."	923	-166	
1090	Buttercup	Promise	1092, 1093	2	
1091	Buttercup	to return him to his ship.	1092, 1093	2	
1092	Humperdinck	I swear	-1119	27	
1093	Humperdinck	it will be done.	-1119	26	
1094	Humperdinck	Once we're out of sight,			1092, 1093
1095	Humperdinck	take him back to Florin and	1119	24	1092, 1094
1096	Humperdinck	throw him in the Pit of Despair.	1119	23	1092, 1095
1097	Rugen	I swear	1118	21	
1098	Rugen	it will be done.	1118	20	

CONDITIONAL

REQUEST

PROMISE

COMMAND

PROMISE

COMMAND

Action Unit ID	Character	On Screen Action Dialogue Line Action	Truth Validation Status	Verifying Lapse	Conflicting AUID
1082	Westley	What was that?			
1083	Buttercup	If we surrender,	1099	16	
1084	Buttercup	and I return with you,	1106	22	
1085	Buttercup	will you promise	1087	2	
1086	Buttercup	not to hurt this man?	-1485	399	
1087	Humperdinck	May I live a thousand years and	-2125	1038	
1088	Humperdinck	never hunt again.	-2125	1037	
1089	Buttercup	He is a sailor on the pirate ship "Revenge."	923	-166	
1090	Buttercup	Promise	1092, 1093	2	
1091	Buttercup	to return him to his ship.	1092, 1093	2	
1092	Humperdinck	I swear	-1119	27	
1093	Humperdinck	it will be done.	-1119	26	
1094	Humperdinck	Once we're out of sight,			1092, 1093
1095	Humperdinck	take him back to Florin and	1119	24	1092, 1094
1096	Humperdinck	throw him in the Pit of Despair.	1119	23	1092, 1095

1119	Pit of Despair	Albino	Albino enter chamber
1120	Pit of Despair	Westley	Westley lie on table
1121	Pit of Despair	Albino	Albino approach Westley

CONDITIONAL

REQUEST

PROMISE

COMMAND

PROMISE

LIE

COMMAND

Action Unit ID	Character	On Screen Action Dialogue Line Action	Truth Validation Status	Verifying Lapse	Conflicting AUID
1082	Westley	What was that?			
1083	Buttercup	If we surrender,	1099	16	
1084	Buttercup	and I return with you,	1106	22	
1085	Buttercup	will you promise	1087	2	
1086	Buttercup	not to hurt this man?	-1485	399	
1087	Humperdinck	May I live a thousand years and	-2125	1038	
1088	Humperdinck	never hunt again.	-2125	1037	
1089	Buttercup	He is a sailor on the pirate ship "Revenge."	923	-166	
1090	Buttercup	Promise	1092, 1093	2	
1091	Buttercup	to return him to his ship.	1092, 1093	2	
1092	Humperdinck	I swear	-1119	27	
1093	Humperdinck	it will be done.	-1119	26	
1094	Humperdinck	Once we're out of sight,			1092, 1093
1095	Humperdinck	take him back to Florin and	1119	24	1092, 1094
1096	Humperdinck	throw him in the Pit of Despair.	1119	23	1092, 1095

1119	Pit of Despair	Albino	Albino enter chamber
1120	Pit of Despair	Westley	Westley lie on table
1121	Pit of Despair	Albino	Albino approach Westley

CONDITIONAL

REQUEST

PROMISE

COMMAND

PROMISE

LIE

COMMAND

Action Unit ID	Character	On Screen Action Dialogue Line Action	Assertive	Subjunctive	Imperative	Modal	Conditional	Result	Purpose	Interrogative
1083	Buttercup	If we surrender,					yes	yes		
1084	Buttercup	and I return with you,	Yes							
1085	Buttercup	will you promise				will				yes
1086	Buttercup	not to hurt this man?								yes
1087	Humperdinck	May I live a thousand years and				may				
1088	Humperdinck	never hunt again.	Yes							

Action Unit ID	Character	On Screen Action Dialogue Line Action	Assertive	Subjunctive	Imperative	Modal	Conditional	Result	Purpose	Interrogative
1083	Buttercup	If we surrender,					yes	yes		
1084	Buttercup	and I return with you,	Yes							
1085	Buttercup	will you promise				will				yes
1086	Buttercup	not to hurt this man?								yes
1087	Humperdinck	May I live a thousand years and				may				
1088	Humperdinck	never hunt again.	Yes							

Action Unit ID	Character	On Screen Action Dialogue Line Action	Assertive	Subjunctive	Imperative	Modal	Conditional	Result	Purpose	Interrogative
1083	Buttercup	If we surrender,					yes	yes		
1084	Buttercup	and I return with you,	Yes							
1085	Buttercup	will you promise				will				yes
1086	Buttercup	not to hurt this man?								yes
1087	Humperdinck	May I live a thousand years and				may				
1088	Humperdinck	never hunt again.	Yes							

Action Unit ID	Character	On Screen Action Dialogue Line Action	Assertive	Subjunctive	Imperative	Modal	Conditional	Result	Purpose	Interrogative
1083	Buttercup	If we surrender,					yes	yes		
1084	Buttercup	and I return with you,	Yes							
1085	Buttercup	will you promise				will				yes
1086	Buttercup	not to hurt this man?								yes
1087	Humperdinck	May I live a thousand years and				may				
1088	Humperdinck	never hunt again.	Yes							

Take Aways

Representation of narrative that includes means for identifying non-actual events

AND keeping track of when their non-actual status changes

Captures features critically relevant for understanding plot

Currently an initial baseline, a lot of further work required

Thank you!

`http://nil.fdi.ucm.es/`

Automated Identification of Competing Narratives in Political Discourse on Social Media

Sergej Wildemann • Erick Elejalde
L3S Research Center, Leibniz University Hannover, Germany

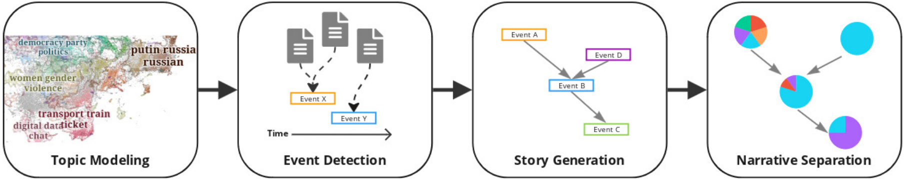
April 10, 2025 • Text2Story@ECIR'25

Objectives

- Unsupervised framework to identify and analyze competing narratives in political discourse on social media
- Focus on German politicians' tweets as the data source

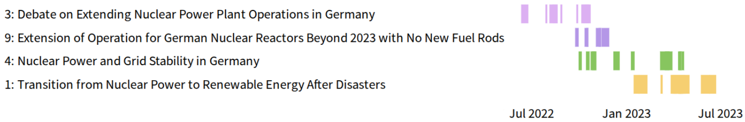
Approach:

- Define an unsupervised multi-stage pipeline
- Aggregate data into coherent stories and uncover distinct perspectives of user communities



Example: Nuclear Power Phase Out in Germany

Storylines

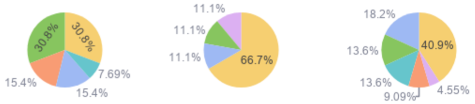


Inside a story

Community X

Our position: #nuclear power ☢ is nothing more than a mirage, because it is expensive and unsafe.	Scheuer is not only calling for the reactivation of old nuclear power plants, but also the construction of three new ones. That is madness.	The #nuclear phase-out remains! Now there is clarity from the federal government

Communities within Events



Community Y

Exiting the nuclear phase-out - our demand and the first step towards dismantling the green energy cuckoo land built by red-left-yellow-black.	It's madness to believe that extending #nuclear power is madness.	We have called for a shift away from the so-called energy transition for years. An end to climate madness. Build a secure and independent energy infrastructure.

Key Findings

- Well-separated user viewpoints validated by politicians' party memberships
  - User communities align with political parties (e.g., Greens vs. AfD).
- The framework can provide temporally/semantically coherent distinct perspectives on diverse key social issues as shown in representative cases
  - *Energy Crisis*: Communities agree on the need for relief but diverge on solutions (government intervention vs. questioning government approach).
  - *Migration Policy*: Communities frame the issue differently (economic needs vs. security).

Contact

Sergej Wildemann

wildemann@l3s.de

www.l3s.de

On the Challenges in Evaluating Visually Grounded Stories

Aditya K Surikuchi Raquel Fernández Sandro Pezzelle

UNIVERSITY OF AMSTERDAM
Institute for Logic, Language and Computation

Visually Grounded Story Generation

Visually Grounded Story Generation

S1. *A man ducks for cover. There are bullet holes in the wall near him, and he covers his face in fear. He tries to get away, but another person grabs him from behind. He holds him in place as the man struggles to get away. Suddenly, the bullets start flying. Both men duck or dive for safety. Food and bits of plaster go flying as the bullets fill the room. Two men are hit, and they both go flying back.*

Visually Grounded Story Generation

S1. *A man ducks for cover. There are bullet holes in the wall near him, and he covers his face in fear. He tries to get away, but another person grabs him from behind. He holds him in place as the man struggles to get away. Suddenly, the bullets start flying. Both men duck or dive for safety. Food and bits of plaster go flying as the bullets fill the room. Two men are hit, and they both go flying back.*

S2. *In the opulent room, a man in a tuxedo was playing the accordion, oblivious to the chaos unfolding around him. Suddenly, another man burst into the room, his face contorted with anger as he lunged at the first man. The second man's attack was met with a swift counter from the first, who managed to push him back. As they struggled, a cloud of dust filled the air, obscuring their movements. In the midst of the commotion, a third man appeared, his expression one of shock and fear as he watched the two men grapple.*

Visually Grounded Story Generation

🗣️ *A man ducks for cover. There are bullet holes in the wall near him, and he covers his face in fear. He tries to get away, but another person grabs him from behind. He holds him in place as the man struggles to get away. Suddenly, the bullets start flying. Both men duck or dive for safety. Food and bits of plaster go flying as the bullets fill the room. Two men are hit, and they both go flying back.*

🤖 *In the opulent room, a man in a tuxedo was playing the accordion, oblivious to the chaos unfolding around him. Suddenly, another man burst into the room, his face contorted with anger as he lunged at the first man. The second man's attack was met with a swift counter from the first, who managed to push him back. As they struggled, a cloud of dust filled the air, obscuring their movements. In the midst of the commotion, a third man appeared, his expression one of shock and fear as he watched the two men grapple.*

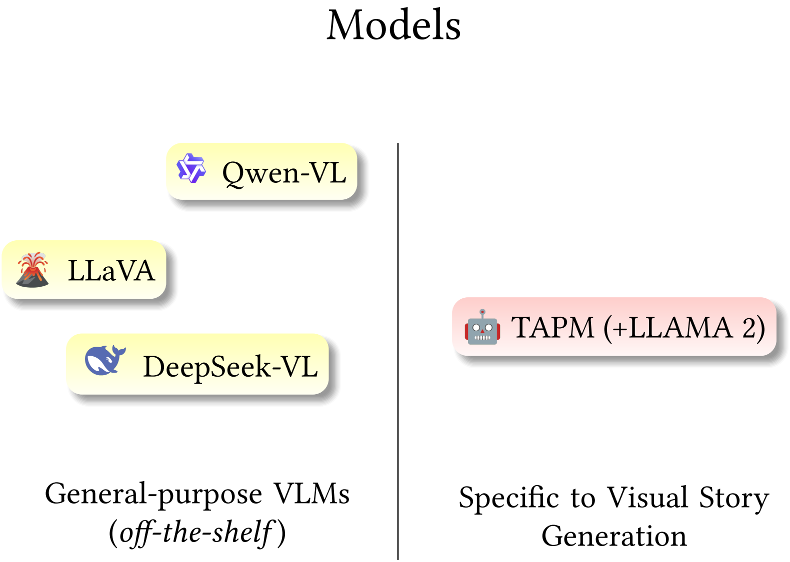
Datasets

VIST

- Sequences constructed using images from Flickr albums.
- Lacks consistency of entities (e.g., *characters*, *objects*)
- Corresponding stories are generally descriptive in nature

VWP

- Sequences constructed using scenes from movies
- Semantically well-connected with recurring characters
- Stories contain diverse entities, are longer, and coherent


Models
Qwen-VL
LLaVA
DeepSeek-VL
General-purpose VLMs
(off-the-shelf)
TAPM (+LLAMA 2)
Specific to Visual Story
Generation

Evaluation

Human evaluation is challenging in terms of:

- 💶 Scalability and costs
- ⚠️ Selecting qualified annotators and reliability

Evaluation

Human evaluation is challenging in terms of:

- 💶 Scalability and costs
- ⚠️ Selecting qualified annotators and reliability

Automatic Metrics:

- BLEU, METEOR, CIDEr, SPICE, ROUGE

Evaluation

Human evaluation is challenging in terms of:

💶 Scalability and costs

⚠️ Selecting qualified annotators and reliability

Automatic Metrics:

❌ ~~BLEU, METEOR, CIDEr, SPICE, ROUGE~~

✔️ Reference-free metrics that assess stories along different aspects—*Coherence*, *Visual grounding*, *Repetition*

Evaluation

Human evaluation is challenging in terms of:

💶 Scalability and costs

⚠️ Selecting qualified annotators and reliability

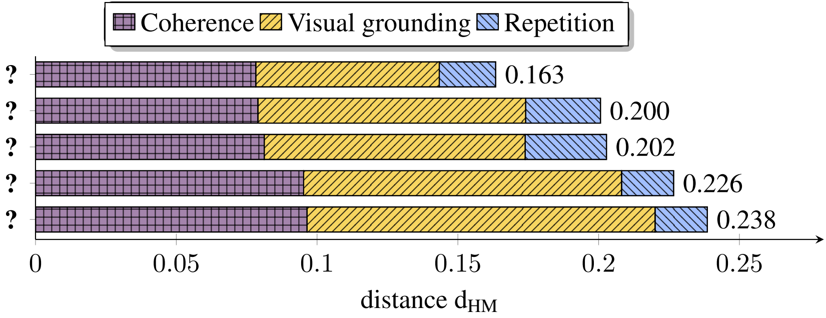
Automatic Metrics:

❌ ~~BLEU, METEOR, CIDEr, SPICE, ROUGE~~

✔️ Reference-free metrics that assess stories along different aspects—*Coherence*, *Visual grounding*, *Repetition*

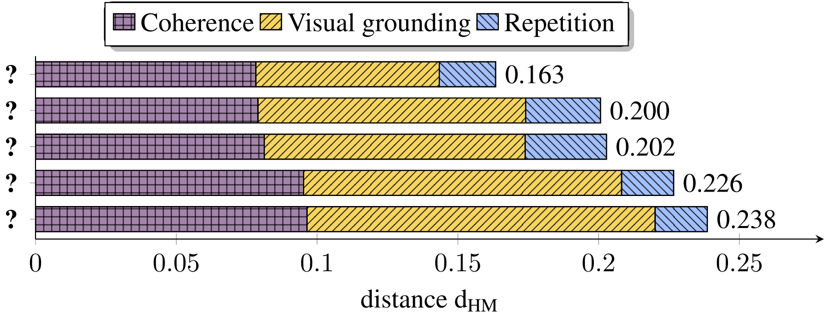
✔️ d_{HM} measure that computes distances between individual metric scores of model- and corresponding human stories

Results



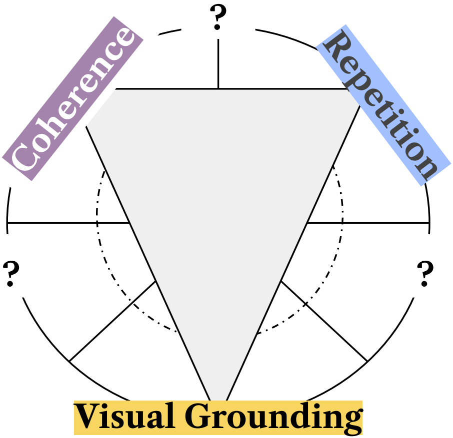
🤨 What is overall best performing model?

Results



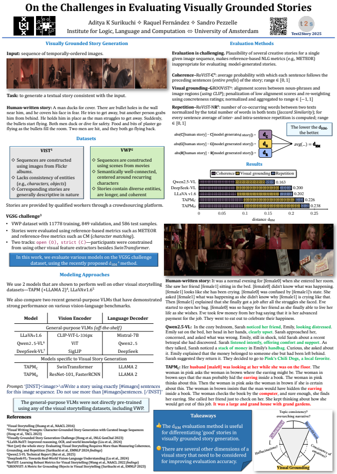
🤨 What is overall best performing model?

🤨 How did the models fare along each of the 3 dimensions?



🤔 Are there other dimensions relevant for evaluating visually-grounded stories?

To learn more, please stop by our poster:



On the Challenges in Evaluating Visually Grounded Stories

Aditya K Surikuchi ✧ Raquel Fernández ✧ Sandro Pezzelle
Institute for Logic, Language and Computation ⇔ University of Amsterdam

Text2Story 2025

Visually Grounded Story Generation

Input: sequence of temporally-ordered images.

Task: to generate a textual story consistent with the input.

Human-written story: A man ducks for cover. There are bullet holes in the wall near him, and he covers his face in fear. He tries to get away, but another person grabs him from behind. He holds him in place as the man struggles to get away. Suddenly, the bullets start flying. Both men duck or dive for safety. Food and bits of plaster go flying as the bullets fill the room. Two men are hit, and they both go flying back.

Datasets

VIST[1]
- ◇ Sequences are constructed using images from Flickr albums.
- ◇ Lacks consistency of entities (e.g., *characters, objects*)
- ◇ Corresponding stories are generally descriptive in nature

VWP[2]
- ◇ Sequences are constructed using scenes from movies
- ◇ Semantically well-connected, centered around recurring characters
- ◇ Stories contain diverse entities, are longer, and coherent

Stories are provided by qualified workers through a crowdsourcing platform.

VGSG challenge[3]
- VWP dataset with 11778 training, 849 validation, and 586 test samples.
- Stories were evaluated using reference-based metrics such as METEOR and reference-free metrics such as CM (*character matching*).
- Two tracks: `open (O), strict (C)`—participants were constrained from using other visual feature extractors besides *SwinTransformer*.

In this work, we evaluate various models on the VGSG challenge dataset, using the recently proposed d_{HM}[4] method.

Modeling Approaches

We use 2 models that are shown to perform well on other visual storytelling datasets—TAPM (+LLAMA 2)[4], LLaVAv1.6[5]

We also compare two recent general-purpose VLMs that have demonstrated strong performance on various vision-language benchmarks.

Model	Vision Encoder	Language Decoder
General-purpose VLMs *(off-the-shelf)*		
LLaVAv1.6	CLIP-ViT-L-336px	Mistral-7B
Qwen2.5-VL[6]	ViT	Qwen2.5
DeepSeek-VL[7]	SigLIP	DeepSeek
Models specific to Visual Story Generation		
$TAPM_C$	SwinTransformer	LLAMA 2
$TAPM_O$	ResNet-101, FasterRCNN	LLAMA 2

Prompt: '[INST]<image>\nWrite a story using exactly [#images] sentences for this image sequence. Do not use more than [#images]sentences. [/INST]'

The general-purpose VLMs were not directly pre-trained using any of the visual storytelling datasets, including VWP.

References

[1]Visual Storytelling (Huang et al., NAACL 2016)
[2]Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences (Hong et al., TACL 2023)
[3]Visually Grounded Story Generation Challenge (Hong et al., INLG GenChal 2023)
[4]LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (Liu et al., 2024)
[5]Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition (Surikuchi et al., EMNLP 2024 *findings*)
[6]Qwen2.5-VL Technical Report (Bai et al., 2025)
[7]DeepSeek-VL: Towards Real-World Vision-Language Understanding (Lu et al., 2024)
[8]RoViST: Learning Robust Metrics for Visual Storytelling (Wang et al., NAACL 2022 *findings*)
[9]GROOViST: A Metric for Grounding Objects in Visual Storytelling (Surikuchi et al., EMNLP 2023)

Evaluation Methods

Evaluation is challenging. Plausibility of several creative stories for a single given image sequence, makes reference-based NLG metrics (e.g., METEOR) inappropriate for evaluating model-generated stories.

Coherence–RoViST-**C**[8]: average probability with which each sentence follows the preceding sentences (*entire prefix*) of the story; range $\in [0, 1]$

Visual grounding–GROOViST[9]: alignment scores between noun-phrases and image regions (*using CLIP*); penalization of low alignment scores and re-weighting using concreteness ratings; normalized and aggregated to range $\in [-1, 1]$

Repetition–RoViST-N**R**[8]: number of co-occurring words between two texts normalized by the total number of words in both texts (*Jaccard Similarity*); for every sentence average of inter- and intra-sentence repetition is computed; range $\in [0, 1]$

abs(**C**[human story] - **C**[model generated story]) = d_C
abs(**G**[human story] - **G**[model generated story]) = d_G
abs(**R**[human story] - **R**[model generated story]) = d_R
$avg(...) = \mathbf{d_{HM}}$

The lower the d_{HM}, the better.

Results

Coherence / Visual grounding / Repetition

Qwen2.5-VL 0.163
DeepSeek-VL 0.200
LLaVA v1.6 0.202
$TAPM_C$ 0.226
$TAPM_O$ 0.238

0 0.05 0.1 0.15 0.2 0.25
distance d_{HM}

Human-written story: It was a normal evening for [female0] when she entered her room. She saw her friend [female1] sitting in the bed. [female0] didn't know what was happening. [female1] looks like she has been crying. [female0] was confused by [female1]'s state. She asked [female1] what was happening as she didn't know why [female1] is crying like that. Then [female1] explained that she finally got a job after all the struggles she faced. Eve started to open her bag. [female0] was so happy for her friend as she finally able to live her life as she wishes. Eve took few money from her bag saying that it is her advanced payment for the job. They went to eat out to celebrate their happiness.

Qwen2.5-VL: In the cozy bedroom, Sarah noticed her friend, Emily, looking distressed. Emily sat on the bed, her head in her hands, clearly upset. Sarah approached her, concerned, and asked what was wrong. Emily, still in shock, told Sarah about a recent betrayal she had discovered. Sarah listened intently, offering comfort and support. As they talked, Sarah noticed a stack of money in Emily's handbag. Curious, she asked about it. Emily explained that the money belonged to someone else but had been left behind. Sarah suggested they return it. They decided to go to Pink's Chili Dogs, a local favorite.

$TAPM_C$: Her **husband [male0] was looking at her while she was on the floor**. The woman in pink asks the woman in brown where the earring might be. The woman in brown says that the man probably hid the earring inside a book. The woman in pink thinks about this. Then the woman in pink asks the woman in brown if she is certain about this. The woman in brown insists that the man would have hidden the earring inside a book. The woman checks the book by the computer, and sure enough, she finds her earring. She called her friend just to check on her. She kept thinking about how she would get out of this job. **It was a large and grand house with grand gardens.**

Takeaways

👍 The d_{HM} evaluation method is useful for differentiating 'good' stories in visually grounded story generation.

🤔 There are several other dimensions of a visual story that need to be considered for improving evaluation accuracy.

Topic consistency? overarching narrative?
Coherence
Repetition
Hallucinations?
Consistency of emotions?
Visual Grounding

Thanks for your attention! ✨

Using LLMs to Generate Patient Journeys in Portuguese: an Experiment

Tahsir Ahmed Munna[1,2], Ana Luísa Fernandes[1,3], Purificação Silvano[1,3,4], Nuno Guimarães[1,2] and Alípio Jorge[1,2]

[1]INESC TEC, Portugal, [2]Faculdade de Ciências da Universidade do Porto (FCUP), Portugal, [3]Faculdade de Letras da Universidade do Porto (FLUP), Portugal,
[4]Centro de Linguística da Universidade do Porto (CLUP), Portugal

ECIR 2025

Text2Story 2025
Lucca, Italy

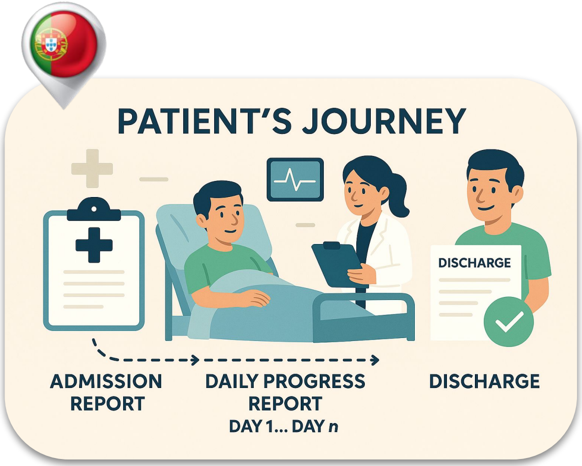
PROBLEM

1. Medical Data Scarcity
2. Patients Full Medical Journey
3. European Portuguese

PROBLEM

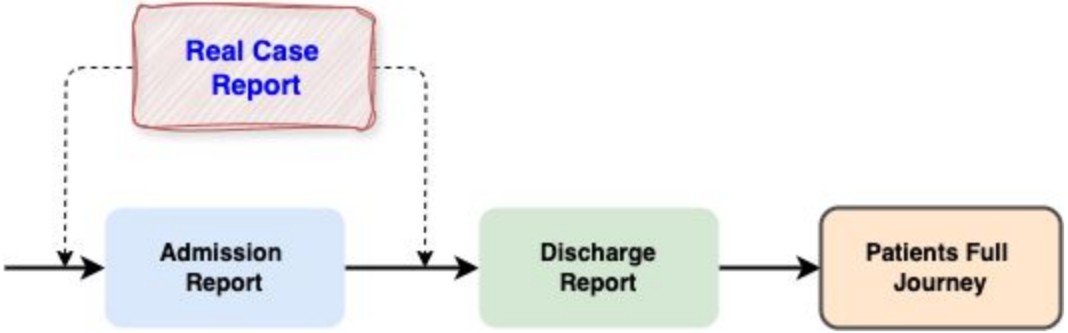
SOLUTION

1. Generating Synthetic Medical Dataset in PT-PT
2. Utilizing **Real** Medical Case Reports
3. **Focusing Patients Full Medical Journey**



Gemini 1.5 Flash

Using LLMs to Generate Patient Journey in Portuguese: An Experiment



Sequential Generation
(European Portuguese)

Evaluation Methods

Using LLMs to Generate Patient Journey in Portuguese: An Experiment