ROBE: Reversed-Order-Biased-Experts for Extracting Extreme Long-tail Events from Historical Texts

Stella Verkijk[1,2,*], Piek Vossen[1]

[1]Vrije Universiteit Amsterdam, The Netherlands
[2]Huygens Instituut, Amsterdam, The Netherlands

Abstract

This paper proposes methods to extract over 50 types of events from a Dutch historical corpus spanning the 17th and 18th centuries. The methods we propose aim to tackle a very challenging scenario in Machine Learning: extracting the long-tail of the long-tail. Historic data from before the 19th century is in itself a niche domain not covered in the pre-training of Large Language Models, and we aim to extract events only scarcely annotated in the training data available for this domain. We propose creating expert classifiers for subgroups of the events present in the training data. We make these groupings based on similar frequency in the training data or on semantic relatedness. Experts trained on underrepresented events are assigned higher priority when predicting to avoid being dominated by frequency biases. We refer to this new way of combining classifiers, specifically tailored to protect the long-tail, as ROBE: Reversed-Order-Biased-Experts. We also propose a controlled method to create domain-specific synthetic data. Our two implementations of ROBE outperform a simple fine-tuned encoder model with a .16 increase in precision and a .05 increase in recall respectively. The best model achieves a .11 increase in f1 for a group of long-tail classes in our niche data set.

Keywords

Information Extraction, Computational History, Low-Resource NLP, Synthetic Data, Combining Classifiers

1. Introduction

Even in the age of huge models, extracting long-tail information from free text remains a massive problem. The term "long-tail" can be interpreted in different ways. Any information from a niche domain can be seen as long-tail because it is infrequent in comparison to information widely available on the web. Similarly, any low-resource language can be considered long-tail. Additionally, the less frequent entities (those persons, places, events, etc. that occur sporadically) in a data set are considered to be long-tail. In this study, we propose a solution to deal with the long-tail of the long-tail, as we extract events from the archives of the Dutch East India Company (VOC) (1602-1799), handwritten in early modern Dutch, that show a very skewed distribution in the training data set.

We formulate this task as a multiclass token classification problem (see Figure 1). Our dataset features 70 label types with the most frequent label occurring 794 times and the least frequent label once. The main problems we face are the skewed data distribution and the relatively large number of scarce data classes. We therefore build expert models for groups of event classes and apply these in a reversed order of bias: a Reversed Order of Biased Experts (ROBE) approach. We split the dataset into groups of labels that either have a similar frequency distribution in the training data or are semantically similar. We fine-tune a model for each group of labels, which we refer to as an expert. Each expert is also assigned a priority. Priority is assigned linearly where the expert trained on the labels with the smallest representation in the training data gets the highest priority and the expert trained on the labels with the biggest representation gets lowest priority. We then let each expert label our test dataset and resolve

The 25th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2026) - Satellite Events (Workshops, Tutorials, Demos, Posters)
*Corresponding author.
s.verkijk@vu.nl (S. Verkijk); p.t.j.m.vossen@vu.nl (P. Vossen)
0009-0000-9263-2272 (S. Verkijk); 0000-0002-6238-5941 (P. Vossen)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

CEUR-WS.org/Vol-4340/paper21.pdf

disagreements between event classes by choosing the label predicted by the model with the highest priority. A complete motivation and description of our ROBE approach is given in Section 4.1.

Previous research has shown that neither monolingual Dutch encoders nor larger multilingual encoders are capable of representing the semantics of these VOC archives, concluding that they are unable to capture relevant generalizations even after fine-tuning on domain-specific data [1, 2, 3]. In view of this, our ROBE approach consists of different fine-tuned implementations of GloBERTise, an encoder model pre-trained solely on the archives of the VOC[1].

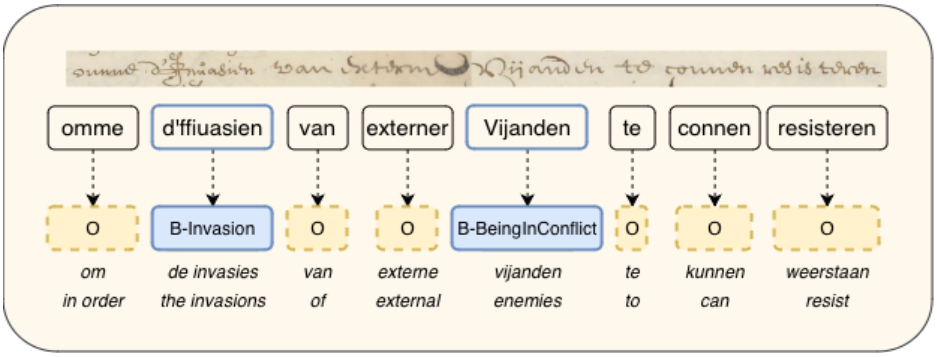


Figure 1: Example of event extraction task. In this example the nouns *invasions* and *enemies* (implicitly) refer to events, whereas the main verb *resist* does not refer to any event from GLOBALISE's event tag set and is therefore not annotated. Recognizing that any part of speech can refer to an event means widening the variation space of tokens that can refer to a single event class, making some instances of events even more long-tail.

We show that compared to a model fine-tuned on all classes simultaneously without any data manipulation, a ROBE approach drastically improves precision, while a ROBE combined with a general fine-tuned model (ROBE+) improves both recall and precision. We also employ a specialized method to create synthetic data. Incorporating the resulting synthetic data in the fine-tuning process improves precision, but not recall.

This research gives insight into how to deal with a skewed dataset in a highly domain-specific and low-resource setting. We offer a solution that can be adopted with little to no extra resources or complicated engineering. We show that creating synthetic data can help if you have a way to infuse your data with domain-specific samples, but that splitting up your dataset and training models on specific label groups increases performance even more. We highlight that underperformance on long-tail knowledge can be mitigated without access to additional domain-specific resources or large generative models. We hope it inspires other researchers to devise new ways to tackle extreme long-tail information extraction.

2. Related Work

Historical event extraction: There is some previous work related to event extraction in older data like ours [4, 5, 6]. Borrenstein et al. [4] aim to extract event information from historical newspaper adverts in different languages. However, they only extract attributes of events, and do not classify the events themselves. Lai et al. [5] present BRAD, an annotated dataset of nineteenth century African American periodicals published from 1827 to 1909. Their task does involve event classification, but only of 12 event types. The authors report that transformer-based models applied on BRAD's event extraction task score substantially worse than when applied on the modern event dataset ACE [7]. Albuquerque et al. [6] apply an event detection and classification system based on FrameNet for Portuguese [8] on 17th century letters, but do not report precision, recall or f1 scores. Other research that proposes automatic systems to extract events from historical texts was conducted on data from the 20th century [9, 10].

[1] https://huggingface.co/globalise/GloBERTise

A study investigating the performance of generative LLMs on event extraction reports that ChatGPT reaches only 51.04% of the performance of a task-specific model in long-tail and complex scenarios [11]. Kister & Schirmer conclude that open-source LLMs can summarize Apartheid Witness Reports but lag behind when extracting entities from them (low precision being the main culprit), even though this dataset is in English and from 1996 [12].

Combining classifiers: Combining classifiers to solve a machine learning task is nothing new [13]. A popular method in Machine Learning is ensembling: training different classifiers on varying features of the dataset and combining the classifier's representations, for example by averaging over their embeddings, or more simply, combining their predictions with a majority vote scheme or a weighted average. ROBE does not combine all classifiers equally, but rather pushes classifiers representing the strongest biases prevalent in the training data to the background by overriding their predictions. Our approach is closer to what is known as a *mixture of (competing) experts* [14, 15]. However, we do not use an extra classifier selection algorithm but rather use a simple rule-based approach to define the order of classifiers once. Also similar to our approach is *bagging*, where classifiers are trained on subsets of the training data. In contrast to our approach, these subsets tend to be randomly sampled.

Synthetic data generation: Creating synthetic data to enhance Machine Learning systems has been a common practice for years [16, 17]. By automatically creating data as training material, laborious human data annotation and curation can be avoided. Whereas earlier methods relied on, for example, Generative Adversarial Networks (GANs) [18, 19], the rise of high-performing generative Large Language Models has introduced a new era of synthetic data creation, especially for solving low-resource NLP problems [20, 21, 22, 23]. There are several advantages in using generative LLMs, such as the ease of prompting with natural language and, more importantly, the creativity of these probabilistic models, which makes it easier to spawn variety in the synthetic data. However, the inconsistency of LLMs, their largely unpredictable behavior, and their lack of knowledge of some specialized domains are downsides. Additionally, high frequency data and classes lead to generalizations adopted by a generative model that suppress long-tail interpretations.

Although synthetic data augmentation can be seen as an established method, the speed at which the wide offer of LLMs that are now used for this method are being published, updated or adapted means the research into this method is also ongoing. Where some research focuses on evaluating different LLMs on their generation abilities [24], or on how to actually evaluate the generated data [25], and thereby the algorithms producing them [26], others focus on enhancing methods in prompting, for example by thinking of ways to generate data from existing datasets [27], using knowledge graphs [28] or entity samples [29].

We draw on this last line of thinking by infusing our prompt templates with data points and (annotation) examples specific to our domain. By doing so, we ingrain domain-specific attributes into our synthetic data without the need to fine-tune or instruct a generative model. This allows us to steer the training towards long-tail phenomena. We describe this process in detail in Section 4.2.

3. Data

The data we work with was created and made publicly available by the GLOBALISE project. GLOBALISE uses their own event ontology as a backbone of their annotation process. This ontology organizes events taxonomically and makes relations between states (Static Events) and changes-of-state (Dynamic Events) explicit. We refer to the GLOBALISE papers on the ontology and the annotation process for further details [30, 31]. The training data consists of 280 annotated pages from 26 documents and the test data of 18 annotated pages from 3 documents.

The task is different from classical event extraction tasks that rely on locating the predicate of the sentence: Any part of speech can refer to an event (see Figure 1). The reasoning behind this task design is twofold: i) the archival texts contain very long sentences with unclear to no sentence boundaries, and ii) historically relevant information about events is often expressed implicitly, for example through nouns or adjectives, rather than only through verbs.

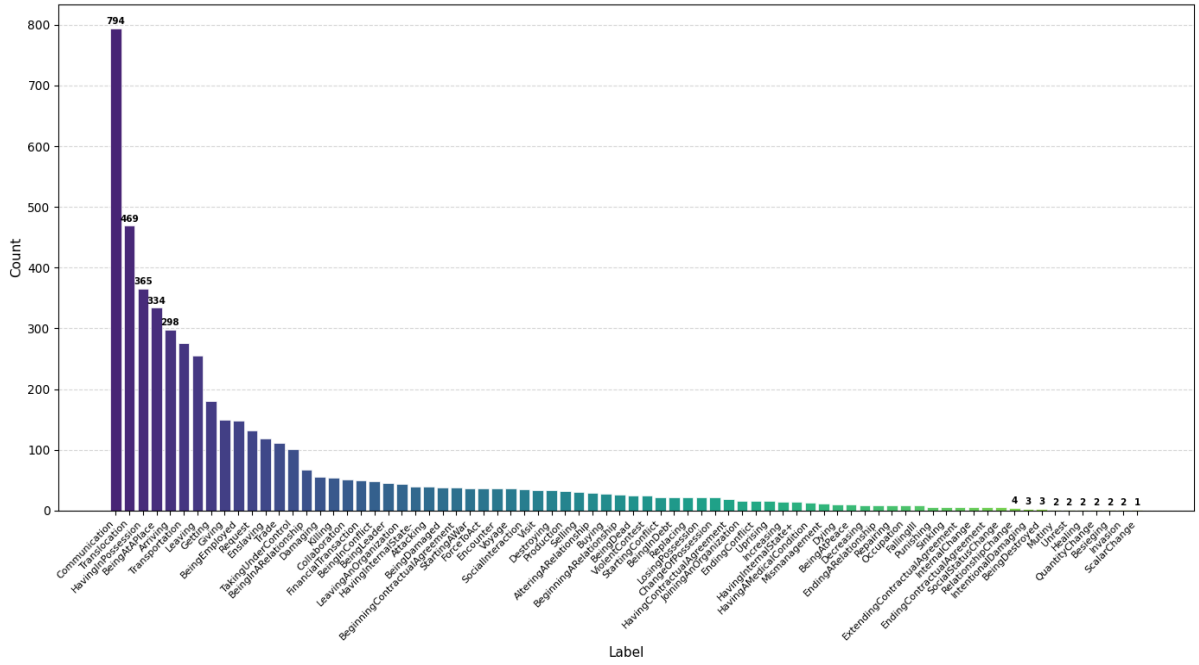


Figure 2: Label distribution in the expert annotated data of all event classes. Counts are per token.

As an intuitive way to illustrate that events and entities that are central to the world of the VOC are low-resource in the current data climate, we can compare the number of hits for certain Google searches (measured on September 19, 2026). The Dutch word 'boot' (*boat*) returns 1,4 billion hits, while the word 'fregat' (a common boat type used by the VOC) returns 1,5 million hits. The English word 'slavery' returns 150 million hits and the word 'enslaving' 2,3 million, while the word 'divorce' returns 217 million hits and 'wedding' 3 billion. Searching for 'Dutch East India Company' returns 196 million hits, while 'Apple' returns 6,7 billion hits.

Zooming in on the data distribution within our domain, Figure 2 shows the highly skewed label distribution in the training data. The frequencies follow a Zipfian distribution. The difference between the most frequent event (*Communication* with 794 tokens, 16% of all tokens with an event label) and the least frequent (*ScalarChange* with 1, 0.02%) is extreme. Also, the difference between the most frequent event and the runner-up (*Translocation*) is large: Translocation represents 9% of all annotated events. 92% of tokens do not refer to an event.

The VOC archives are known to describe a lot of trade and politics, but also violence. The VOC was a colonial enterprise and political entity that spread war and slave labor. However, describing the violence was not the main focus of the employees of the VOC, as writing letters was more often motivated by a need for more money and resources from the homeland, the obligation to report on evolving relationships with local rulers, and other economically relevant topics. As we can see in Figure 2, some of the events at the extreme end of the long-tail are *Mutiny*, *Besieging* and *Invasion*. Other events that are typical of VOC violence, such as *Enslaving*, *TakingUnderControl*, *Attacking* and *ForceToAct* are also much less frequent than classes related to translocation and trade. Hence, we observe a cultural bias reflected in the data distribution. A system that is only able to extract frequent classes will inaccurately represent the history of the VOC and the ability to extract long-tail events is therefore essential.

4. Method

4.1. The ROBE approach

The idea behind ROBE is to split the classification problem into smaller chunks (bootstrapping) by training different models on different groups of events. These groupings are based on their representation in the training data and on their semantics (putting semantically similar events in the same group). Grouping classes with similar representations avoids one class being pushed to the background by other classes with much more examples in the training data.

A typical thing you see when training models on little data is that after a small amount of epochs, the model is only able to recognize the most common classes it has seen in the training data and starts predicting those any time it encounters something that refers to an event. The intuition is thus that grouping classes with similar representations in one task makes the model converge for each class at a similar moment during training. All models were fine-tuned with learning rate 5e-5. For further parameter settings we refer to our repository with all code and data to fine-tune models[2].

When training a model for a single group, all annotated labels of classes that were not included in the group are relabeled as 'O', similar to a one-against-all training regime. While this introduces the downside that each group has an even bigger 'O' class, the impact of this is limited because the 'O' class is already extremely dominant, with 92% of tokens (see Figure 3 in the Appendix).

Event classes in group	#classes	Grouped on repr. window	Grouped on semantics	#e	Prio
Buying, BeginningARelationship, ViolentContest, BeingInDebt, Replacing, LosingPossession, ChangeOfPossession, JoiningAnOrganization, Unrest, HavingInternalState+, Mismanagement, HavingAMedicalCondition, BeingAtPeace, QuantityChange	14	10–30	n/a	20	**1**
TakingUnderControl, Enslaving, Unrest, Attacking, StartingAWar, Killing, BeingInConflict, ViolentContest, ForceToAct	9	n/a	Violence, repression, conflict	20	**2**
BeingInARelationship, Damaging, Killing, Collaboration, FinancialTransaction, BeingInConflict, BeingLeader, LeavingAnOrganization, HavingInternalState–, Attacking, BeingDamaged, HavingContractualAgreement, StartingAWar, ForceToAct, Encounter, SocialInteraction, Visit, BeingDestroyed, Production, Selling, BeingDead, RelationshipChange	22	30–100	n/a	18	**3**
Getting, Giving, BeingEmployed, Request, Enslaving, Trade, TakingUnderControl	7	100–200	n/a	12	**4**
HavingInPossession, Giving, Buying, Selling, ChangeOfPossession, Trade, FinancialTransaction, LosingPossession, Getting	9	n/a	Possession	20	**5**
HavingInPossession	1	300–400	n/a	20	**6**
Translocation, Transportation, Leaving, Arriving, Voyage, BeingAtAPlace	6	n/a	Translocation	20	**7**
Communication	1	>700	n/a	10	**8**

Table 1
Event classes grouped for training the experts of ROBE, given in order of priority. Groups with long-tail classes get higher priority. #e = number of epochs the expert was trained; Prio = Priority. Grouped on repr. window = range of label frequency in the training data given in amount of tokens, not mentions.

[2] https://github.com/globalise-huygens/nlp-event-finetuning-experts

See Table 1 for an overview of all groups with their representation in the training data. For the ROBE experts, some classes are merged in order to increase their representation. Classes are only considered for merge if it concerns i) a related Dynamic and Static Event, ii) events that are subclasses of the same class in the taxonomy, or iii) events that are one step away from each other in the taxonomy. Examples of the first are *Dying* and *BeingDead*; *Destroying* and *BeingDestroyed*. Examples of the second are *Besieging* and *Invasion*, where both are subclasses of *Attacking*. Examples of the third are *EndingARelationship* and *RelationshipChange*, where the first is a subtype of the second. Table 6 in the Appendix shows which classes are merged, what type of merge it is, and what their individual representation is.

4.2. Synthetic data creation

As explained in Section 2, previous studies on creating synthetic data have shown that infusing (domain-specific) knowledge from existing datasets can lead to more representative synthetic data for niche domains. Part of GLOBALISE's work involves creating structured data of persons, places, ships and commodities described in the archives (i.e., [32]). We use GLOBALISE's published dataset on the "Name-books" of the VOC to extract names of persons, their professions and the location they were mentioned [3]. Additionally, we use their published commodities thesaurus to extract names of commodities and their definitions[33][4]. Lastly, GLOBALISE provided a ship dataset[5] from which we extract names of ship types and their descriptions. We then create prompt templates that combine these data like so:

> *See the following example of an annotated sentence:* <example>. *Create a sentence in Early Modern Dutch in which* <commodity> *is transported. This is the definition of* <commodity>: <definition>. *Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the transportation in square brackets like in the example.*

We create prompts by randomly sampling from the aforementioned datasets and a small selection of 9 example annotations to create several combinations for each prompt template. Some prompt templates do not use an example annotation to promote diversity in the output. We use a total of 8 prompt templates. Table 7 in the Appendix contains all the prompt templates used and Table 8 all annotation examples. Using the created prompt templates between 15 to 30 times we create 375 annotated synthetic sentences for *Translocation* and *Transportation* with gpt-4o. We chose *Translocation* and *Transportation* because these are the most appropriate events to infuse with domain-specific entities like locations, ships and commodities. It serves as a proof-of-concept: If results show incorporating this type of synthetic data positively impacts performance on these classes, we can consider doing something similar for more extreme long-tail classes in the future[6].

4.3. Overview of evaluated methods

Table 2 gives an overview of all methods evaluated against the test set. The End to End model is simple: GloBERTise fine-tuned once on all available human-annotated training data without any data manipulation (EtE). We also fine-tune a version of this model on the training data augmented with our synthetic data. Apart from ROBE as described in Section 4.1, we also propose a version of ROBE where we include EtE as an expert: ROBE+. EtE gets lowest priority in ROBE+ (i.e., place 9).

The lexicon and the GenLLM serve as baselines: the first representing the extreme end of domain-specificity and human labor and the second representing the extreme end of general pre-trained models that do not require any human annotations. As per the previous research mentioned so far indicating general pre-trained models do not perform well on long-tail domains [2, 1, 22, 12] as well as research that shows that Generative LLMs do not fare well when applied to classification problems where the

[3] https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/MD59SC
[4] https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/5PZPPK
[5] This data set is still under development and will be published in the future
[6] All code and the resulting synthetic data can be found at https://github.com/globalise-huygens/nlp-event-synthetic-data-generation.

label set is large [34], we do not expect the GenLLM to perform well. From the lexicon we can expect more, as it was created in an iterative manner, using seeds from human annotations and expanding with a specialized Word2Vec model. Furthermore, contributions to the lexicon were given by annotators who also annotated the test set, so the possibility that the lexicon is overfit on the test set should be taken into account. Annotators added words to the lexicon over a time span of years[7].

Approach	Acronym	Design
Lexicon	Lexicon	Lexical Baseline provided by GLOBALISE
gpt-5.1	GenLLM	Generative baseline using gpt-5.1, zero-shot and few-shot setting
End to End	EtE	Fine-tuned GloBERTise on all data
End to End with augmented data	EtE+synth	EtE trained on all data plus additional synthetic data
Reversed-Order-Biased-Expert	ROBE	Eight fine-tuned GloBERTises on different label groups
ROBE with EtE	ROBE+	Nine fine-tuned GloBERTises, of which one EtE
ROBE with EtE and lexicon	ROBE+Lex	Nine fine-tuned GloBERTises, of which one EtE, and a lexicon

Table 2
Overview of evaluated methods

4.4. Evaluation

GLOBALISE released three documents annotated by at least four annotators as test data. We adjudicated these annotations by extracting the majority vote where possible, and otherwise consulting the annotation guidelines and consulting historians. We kept track of our progress and concluded that the percentage of adjudications that could be done through a majority vote was only 47%. Of those, the percentage where the majority vote (three out of four annotations) was 'O' was 52%. Additionally, the percentage of manual adjudications that could be made on ontological basis was only 12%. Hence, when annotators disagreed between event types, there was often no simple way to adjudicate them. These statistics show that the annotation is partially subjective, a phenomenon that is encountered often in NLP [35]. We can expect that evaluating strictly on the adjudicated labels may result in relatively low scores.

5. Results

5.1. Evaluation on test set

	Lexicon			EtE			EtE+synth			ROBE			ROBE+			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Supp.
overall	**0.69**	0.14	0.23	0.29	0.20	0.24	0.42	0.19	0.26	0.45	0.20	**0.28**	0.31	**0.25**	0.27	626
translocation	**0.77**	0.25	0.38	0.66	0.59	0.62	0.75	0.49	0.59	0.74	0.51	0.60	0.66	**0.65**	**0.65**	208
violence	**1.00**	0.03	0.05	0.20	0.04	0.07	0.40	0.08	0.14	0.38	0.08	0.13	0.31	**0.12**	**0.18**	73

Table 3
Micro average results on the adjudicated test data across event classes for EtE, EtE+synth, ROBE, ROBE+, and the lexical baseline. For the overall scores, we apply strict evaluation on BIO labels and event types (a prediction is only correct if it has the correct BIO label and the correct event type). For both the scores on translocation and violence, we count a prediction as correct if it falls within the set of labels that refer to either translocation or violence (as specified in Table 1. BIO labels do not have to overlap. To exemplify: if the gold label is *B-Translocation* and the model predicts *I-Transportation*, we count it as correct. We do not include the 'O' class in any metric.

[7]See GLOBALISE's blog post on their lexical approach: https://globalise.huygens.knaw.nl/yes-okay-but-what-were-they-doing/. We used the second release of the lexicon, which is available at https://github.com/globalise-huygens/nlp-event-lexical-approach/blob/main/lexicon_release2.csv.

As we can see in Table 3, the lexicon offers a strong baseline, scoring highest in precision across the board. ROBE scores .16 higher in precision over EtE overall and .18 for the more challenging group of labels referring to violence, repression and conflict. Recall also increases slightly for this group. Recall does not increase in the overall score and decreases for the translocation labels. Our combined model, ROBE+, scores highest in recall and f1 in most evaluation settings. The EtE model trained with additional synthetic data drastically improves precision over EtE, but declines in recall. We think this might be caused by the fact that the synthetic data contains unlabelled mentions of other events than *Transportation* and *Translocation*, providing the model with negative examples for classes it should recognize. It might also be that the generated output is limited in its variety in linguistic realizations of events. The EtE model trained on augmented data is outperformed in its f1 score for translocation classes by both ROBE and ROBE+. Training an expert translocation model on the augmented dataset, and replacing the original translocation expert with the synthetic expert in the ROBE model does not increase scores for ROBE or ROBE+, as can be seen in Table 9 in the Appendix[8]. Gpt5.1 performs poorly overall and for the isolated translocation labels. It scores on the higher end for violence classes, but not very reliably, as f1 scores differ while the temperature was set to 0, as shown in Table 10 in the Appendix.

Unfortunately, the classes in the lowest frequency group, which we trained the classifier with the highest priority on, do not have sufficient support in the test data to report reliable results on. The only class in this group with a support of more than 5 tokens in the adjudicated data is *HavingInternalState+* with a support of 11. The lexicon and EtE models have an f1 of 0 for this class across the board, while both ROBE and ROBE+ have an f1 of 0.15, but this score is only based on the same 2 predictions by each model.

To give our best performing model a little boost, we can include our lexicon as the lowest-priority classifier: the lexicon is used to label events missed by any of the classifiers in ROBE+ (i.e., ROBE+ labels a string as 'O' but the lexicon does not). This improves scores slightly.

ROBE+ with Lexicon				
	P	R	F1	Support
overall	**0.32**	**0.27**	**0.29**	626
translocation	0.65	**0.69**	**0.67**	208
violence	0.31	0.12	0.18	73

Table 4
Micro average results on the adjudicated test set across event classes for ROBE+ combined with the lexicon. Scores are in boldface if they surpass any of the scores presented in Table3. Metrics calculated as in Table 3.

The results indicate that assigning higher priority to classifiers that are experts on long-tail events does not hurt performance on events that are not long-tail. Experts on subsets of the dataset are more conservative and improve precision. Combining experts with a general model trained on the complete dataset balances out precision and recall[9].

5.2. External Validation

We also gathered external user validation of a simplified version of the ROBE model[10]. These external validation data were gathered during two iterations of a workshop for historians about Machine Learning.

[8]We make all trained experts and the EtE model publicly available at https://huggingface.co/collections/StellaVerkijk/robe-event-classifiers-for-globalise and both models trained on synthetic data at https://huggingface.co/collections/StellaVerkijk/event-classifiers-trained-on-synthetically-augmented-data

[9]Code and data to reproduce results can be found at https://github.com/globalise-huygens/nlp-event-classifier-evaluation/tree/main

[10]A ROBE without experts for possession and violence (i.e., missing models 2 and 5) and with a slight difference in priorities, namely 7-6-1-3-4, favoring the translocation model and the one built on the 300-400 token frequency window.

Participants were asked to assign one of two values to each prediction of ROBE: *useful* or *misleading*. See the results of the simplified ROBE on the part of the test set that validators judged in Table 5.

In the first iteration of the workshop, the model's predictions were labeled as useful **75%** of the time (average of six validators). This is based on a total of 329 validated (non unique) predictions. Of the validators, three had 1-5 years of experience reading archives, two had one year of experience and one had none. One validator was discarded because they did not pass a translation task made to assess their level of reading early modern Dutch. Interestingly, the participants labeled our adjudicated gold data as useful 68% of the time (based on a total of 78 validated gold labels). This reiterates that we are dealing with a complicated task where multiple interpretations of the text are possible. However, it should also be taken into account that the regions of the document in which we showed gold annotations were typically the regions where the model missed most labels, meaning those regions were more complex. Three out of six annotators did not validate any gold data (they did not reach that part of the assignment).

	Simplified ROBE			
	P	R	F1	Support
overall	0.38	0.13	0.20	260
translocation	0.75	0.55	0.64	78
violence	0.00	0.00	0.00	19

Table 5
Micro average results on the adjudicated test set for a simplified ROBE **on part of the test set**. Metrics calculated as in Table 3.

In the second iteration (9 participants), ROBE's predictions were labeled useful **79%** of the time, based on a total of 578 validated predictions. In this group, there was one person with 1-5 years of experience reading archives, one with less than half a year, and the rest had none. They labeled our adjudicated gold data as useful 84% of the time, which was based on a total of 190 validated gold labels. One person did not validate any gold data[11].

Since validators only judged events predicted by ROBE, they did not have any way to judge events ROBE might have missed, hence, we are only validating the model's precision, and we cannot say anything about the recall. However, the conclusion we can draw is that in the external validation, ROBE scores much higher on precision than in our classical evaluation, as can be seen in Table 5. We believe this finding contextualizes the somewhat low scores across the board in Table 3.

6. Conclusion

Our paper offers an accessible approach to enhance performance when classifying long-tail events in a low-resource language and niche domain. By combining classifiers and assigning a fixed picking order, where classifiers trained on underrepresented labels in the training data receive highest priority, it is possible to improve performance across classes in a skewed dataset and balance out precision and recall.

Our paper has a few limitations. An analysis of the synthetic data would give more insight into why it improves precision but not recall. Also, a more fine-grained error analysis where we compare predictions of different models on tokens that refer to specific long-tail events in the test set would provide a clearer analysis of the difference between model ensembling configurations.

[11]For the data see https://github.com/globalise-huygens/nlp-event-external-machine-output-validation and for workshop slides see https://zenodo.org/records/20118595

7. Declaration on Generative AI use

Generative AI was consulted for writing some of the software presented in this study. If a complete function was written by AI, we state this in the code's documentation. We did not use Generative AI for any of the categories described in CEUR's GenAI Usage Taxonomy.

Acknowledgments

This research was funded by the Dutch Research Council (NWO) under project number 175.2019.003 and is part of the Globalise project. All fine-tuning took place on the National Supercomputer Snellius, which we were granted access to by the Dutch national e-infrastructure with the support of the SURF Cooperative using grantS no. EINF-11854 and EINF-18858. We want to thank Kay Pepping for his help conceptualising and setting up the synthetic data creation pipeline and Lodewijk Petram for proofreading. Most importantly, we want to thank Globalise's annotators: Brecht Nijman, Femke Brink, Renate Smit, Henrike Vellinga, Kay Pepping, Philipp Huber, Pascal Konings, Lodewijk Petram and Manjusha Kuruppath.

References

[1] S. Arnoult, B. Nijman, L. v. Wissen, Fine-grained named-entity recognition for the east-india company domain, Anthology of Computers and the Humanities 3 (2025) 953–967.

[2] S. Verkijk, P. Vossen, P. Sommerauer, Language models lack temporal generalization and bigger is not better, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 20629–20637.

[3] S. Verkijk, P. Vossen, Out-of-tune rather than fine-tuned: How pre-training, fine-tuning and tokenization affect semantic similarity in a historical, non-standardized domain, in: Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), 2026, pp. 515–531.

[4] N. Borenstein, N. da Silva Perez, I. Augenstein, Multilingual event extraction from historical newspaper adverts, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 10304–10325.

[5] V. D. Lai, M. Van Nguyen, H. Kaufman, T. H. Nguyen, Event extraction from historical texts: A new dataset for black rebellions, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2390–2400.

[6] G. C. Albuquerque, M. Souza, R. Vieira, A. S. Ribeiro, Applying event classification to reveal the estado da índia, in: Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1, 2024, pp. 247–254.

[7] C. Walker, S. Strassel, J. Medero, K. Maeda, Ace 2005 multilingual training corpus, (No Title) (2006).

[8] A. d. S. B. Sacramento, M. Souza, Joint event extraction with contextualized word embeddings for the portuguese language, in: Brazilian Conference on Intelligent Systems, Springer, 2021, pp. 496–510.

[9] R. Sprugnoli, S. Tonelli, Novel event detection and classification for historical texts, Computational Linguistics 45 (2019) 229–265.

[10] A. Cybulska, P. Vossen, Historical event extraction from text, in: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011, pp. 39–43.

[11] J. Gao, H. Zhao, C. Yu, R. Xu, Exploring the feasibility of chatgpt for event extraction, arXiv preprint arXiv:2303.03836 (2023).

[12] P. Kister, M. Schirmer, Evaluating open-source LLMs for text summarization and named entity recognition in long, unstructured text, in: S. Hamilton, E. Öhman, R. M. M. Hicke, Y. Bizzoni,

A. Bax, J. A. Matthews, M. Hämäläinen (Eds.), Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, Association for Computational Linguistics, San Diego, USA, 2026, pp. 390–410. URL: https://aclanthology.org/2026.nlp4dh-1.35/. doi:10.18653/v1/2026.nlp4dh-1.35.

[13] S. Abimannan, E.-S. M. El-Alfy, Y.-S. Chang, S. Hussain, S. Shukla, D. Satheesh, Ensemble multifeatured deep learning models and applications: A survey, IEEE Access 11 (2023) 107194–107217.

[14] A. A. Aburomman, M. B. I. Reaz, A survey of intrusion detection systems based on ensemble and hybrid classifiers, Computers & security 65 (2017) 135–152.

[15] S. Masoudnia, R. Ebrahimpour, Mixture of experts: a literature survey, Artificial Intelligence Review 42 (2014) 275–293.

[16] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 international interdisciplinary PhD workshop (IIPhDW), IEEE, 2018, pp. 117–122.

[17] K. Maharana, S. Mondal, B. Nemade, A review: Data pre-processing and data augmentation techniques, Global Transitions Proceedings 3 (2022) 91–99.

[18] F. H. K. dos Santos Tanaka, C. Aranha, Data augmentation using gans, Proceedings of Machine Learning Research XXX 1 (2019) 16.

[19] Y. Akkem, S. K. Biswas, A. Varanasi, A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network, Engineering Applications of Artificial Intelligence 131 (2024) 107881.

[20] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, Advances in Neural Information Processing Systems 35 (2022) 462–477.

[21] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, H. Liu, Large language models for data annotation and synthesis: A survey, arXiv preprint arXiv:2402.13446 (2024).

[22] J. Gao, R. Pi, Y. Lin, H. Xu, J. Ye, Z. Wu, W. Zhang, X. Liang, Z. Li, L. Kong, Self-guided noise-free data generation for efficient zero-shot learning, arXiv preprint arXiv:2205.12679 (2022).

[23] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, L. Kong, Zerogen: Efficient zero-shot learning via dataset generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 11653–11669.

[24] S. Kim, J. Suk, X. Yue, V. Viswanathan, S. Lee, Y. Wang, K. Gashteovski, C. Lawrence, S. Welleck, G. Neubig, Evaluating language models as synthetic data generators, CoRR (2024).

[25] K. Ramesh, D. Smolyak, Z. Zhao, N. Gandhi, R. Agarwal, M. Bjarnadóttir, A. Field, Synthtext-teval: Synthetic text data generation and evaluation for high-stakes domains, arXiv preprint arXiv:2507.07229 (2025).

[26] A. Havrilla, A. Dai, L. O'Mahony, K. Oostermeijer, V. Zisler, A. Albalak, F. Milo, S. C. Raparthy, K. Gandhi, B. Abbasi, et al., Surveying the effects of quality, diversity, and complexity in synthetic data from large language models, arXiv preprint arXiv:2412.02980 (2024).

[27] S. Gandhi, R. Gala, V. Viswanathan, T. Wu, G. Neubig, Better synthetic data by retrieving and transforming existing datasets, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 6453–6466.

[28] H. Kim, J. Hessel, L. Jiang, P. West, X. Lu, Y. Yu, P. Zhou, R. Bras, M. Alikhani, G. Kim, et al., Soda: Million-scale dialogue distillation with social commonsense contextualization, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12930–12949.

[29] Z. Yang, N. Band, S. Li, E. Candes, T. Hashimoto, Synthetic continued pretraining, arXiv preprint arXiv:2409.07431 (2024).

[30] S. Verkijk, P. Vossen, Sunken ships shan't sail: Ontology design for reconstructing events in the dutch east india company archives., in: CHR, 2023, pp. 320–332.

[31] S. Verkijk, P. Sommerauer, P. Vossen, Studying language variation considering the re-usability of modern theories, tools and resources for annotating explicit and implicit events in centuries old text, in: Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and

Dialects (VarDial 2024), 2024, pp. 174–187.
[32] R. Mourits, K. Pepping, T. van Oort, P. Konings, B. van Duijvenvoorde, Modelling the enslaved as historical persons: Extending the persons in context (pico) model to fit enslaved individuals, in: Digital Humanities Benelux 2025, 2025.
[33] K. Pepping, Reflections on encoding languages in historical data: Working with the multilingual dimension of the dutch east india company archives, Journal of Open Humanities Data 10 (2024).
[34] O. Khrapunova, Bridging zero-shot and fine-tuned performance in text classification throughretrieval-augmented prompting, American Academic Scientific Research Journal for Engineering, Technology, and Sciences (2025).
[35] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

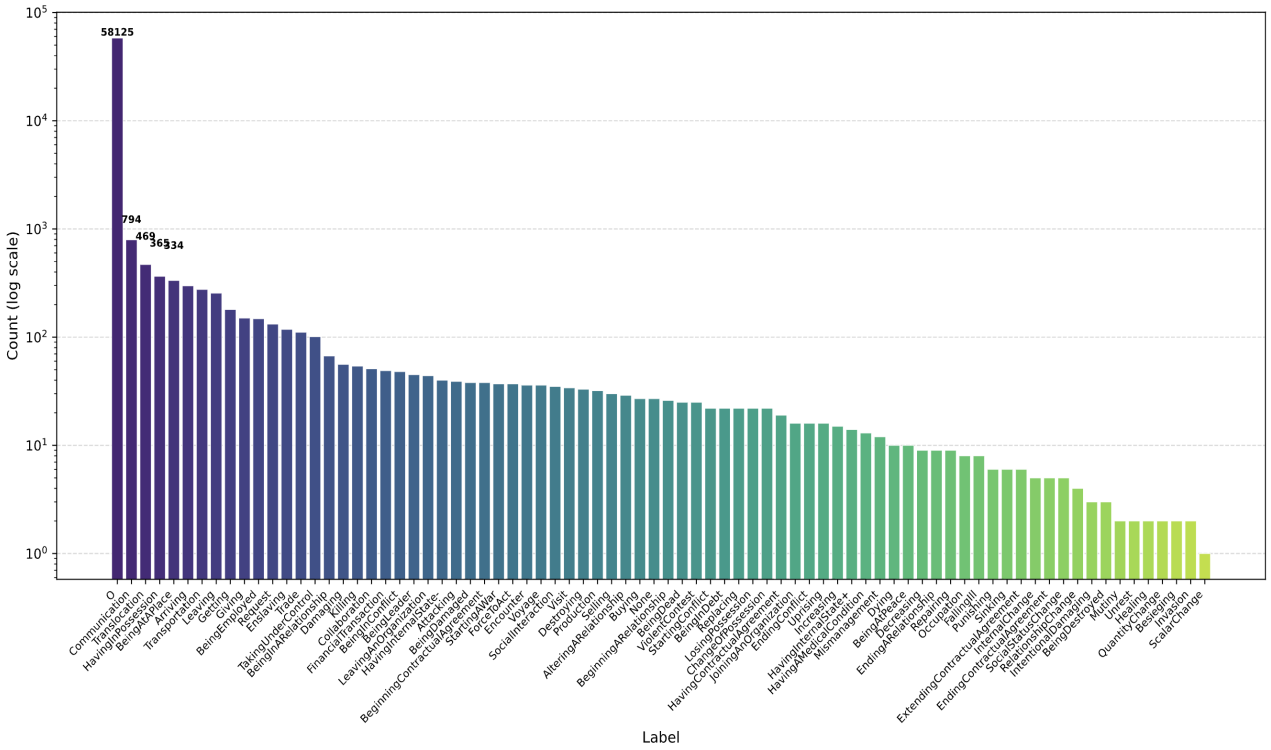


Figure 3: Label distribution in the expert annotated data of all classes (including 'O').

Classes merged	Rep	Merge type	New name	New rep
TakingUnderControl, Occupation	101, 8	Stat-Dyn	TakingUnderControl	109
Attacking, Besieging, Invasion	39, 2, 2	Taxonomic	Attacking	43
Destroying, BeingDestroyed	33, 3	Stat-Dyn	BeingDestroyed	36
BeginningContractualAgreement, ExtendingContractualAgreement, EndingContractualAgreement, HavingContractualAgreement	38, 6, 5, 19	Stat-Dyn	HavingContractualAgreement	68
Dying, BeingDead	10, 25	Stat-Dyn	BeingDead	35
AlteringARelationship, EndingARelationship, RelationshipChange	29, 9, 4	Taxonomic	RelationshipChange	42
StartingConflict, EndingConflict, BeingInConflict	22, 16, 48	Stat-Dyn	BeingInConflict	86
Uprising, Mutiny, Unrest	16, 2, 2	Stat-Dyn	Unrest	20
Repairing, Healing, HavingInternalState+	9, 2, 14	Stat-Dyn	HavingInternalState+	25
FallingIll, HavingAMedicalCondition	8, 13	Stat-Dyn	HavingAMedicalCondition	21

Table 6
Merged event classes. Rep = representation.

See the following example of an annotated sentence: example. Create a sentence in Early Modern Dutch in which label is transported. This is the definition of label: definition. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the transportation in square brackets like in the example.
See the following example of an annotated sentence: example. Create a sentence in Early Modern Dutch in which label is transported. This is the definition of label: definition. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the transportation in square brackets like in the example. Do not repeat the definition as given in the prompt, only use it for your own context.
See the following example of an annotated sentence: example Create a sentence in Early Modern Dutch in which a translocation is taking place with a label. This is the definition of label: definition. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets like in the example. Make sure to vary in syntax.
See the following example of an annotated sentence: example Create a sentence in Early Modern Dutch in which a translocation is taking place with a label. This is the definition of label: definition. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets like in the example. A translocatiion could for example be an arrival, a departure, a voyage or a transport. Make sure to vary in syntax.
Create a sentence in Early Modern Dutch in which a translocation is taking place with a label. This is the definition of label: definition. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets. Do not start your sentence with a time indication.
See the following example of an annotated sentence: example. Create a similar sentence, that differs in syntax and vocabulary, in Early Modern Dutch in which a translocation is taking place. A translocation could for example be an arrival, a departure, a voyage or a transport. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets like in the example.
Create a sentence in Early Modern Dutch in which name is travelling to or from location. The way name is travelling can be inspired by their profession, which is function. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets.
Create a sentence in Early Modern Dutch in which name is travelling to or from location. The way name is travelling can be inspired by their profession, which is function. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets. Do not start the sentence with name.

Table 7
All prompts used to create synthetic data

wordende door de onderkooplieden Jevers en van Bijland aan de vrlreep gerecibieert, en bij de hand naar den Heer Commissaris [geleid] bestaande in Aria soeradikarsa, Aria wangsanagara Aria ong Jaija, en den Ingabeij Ratoe Bagoes Pringalaija
Den 27=en Januarij 1625 sijn dan hier in sloote naer t vaderslandt [vertrocken], ondert Commandement, vanden E: Cornelis Geuertss de Scheepen hollandia Goudaende Middelburch.
Vant vaderslandt sijn hier Godtloff alle de Schepen uijt gesondert alckmaer int Gexel behouden [aengecomen] Namentlijck . T Jacht de Hase den 4 Junij 1625 T schip de Cameel den 19: septemb = r
Is bij onsen scheeps Raet goet gevonden om [voort te seijlen] naer de soute Eijlande om hem aldaer te vinden ofte een dach off drie te verwachten
deen 25 . november sijn wij [gepasseert] het Eijlant palma ende vernomen dat ons Esels hooft gans veermolsent (was soo dat wij L. tgroote marsseyjl niet dorsten voeren den 10 december sin wij [gearrineert] opde rede voor Jlijo de maijo alwaer wij een ander Esels hooft hebben toe gestelt
man die binnen boord bleeven , in de boot aan land [gevlugt] , en daar soolange verbleven waren tot tijd enwijle sagen dat hetselve bij hare chialoup — aan boord gelegen , daar eenige uuren aandoende geweest dog die nae dato weder [verlaten] , en ['t zee gat gekosen] hadde , als wanneer hun ook weder naar derwaarts [begeven] , en aan boord komende door de twee aldaar verblevene maats te verstaan gegeven was
t beste benevens diverse provi sien , Equipagie goederen & t „ a tesamen omtrend ter waardije van 400 : thailen daar [uijtgeligt] envoor een goede buijt [mede gevoert] hadden
t beste benevens diverse provi sien , Equipagie goederen & t „ a tesamen omtrend ter waardije van 400 : thailen daar [uijtgeligt] envoor een goede buijt [mede gevoert] hadden waar nae demallabaren met d „ o chialoup ook [terug gekeert] , en weder op de atchinse rheede [gelopen]

Table 8
All example annotations used to fill the prompt templates' example slots

	ROBE with synth			ROBE+ with synth			
	P	R	F1	P	R	F1	Support
overall	0.45	0.18	0.26	0.30	0.24	0.27	626
translocation	0.77	0.45	0.57	0.65	0.63	0.64	208
violence	0.38	0.08	0.13	0.32	0.12	0.18	73

Table 9
Results across event classes for ROBE with a classifier for translocation classes trained on synthetically augmented data and ROBE+ with the same classifier for translocation classes

	Zero-shot			**Few-shot**									
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Support
overall	0.07	0.04	0.05	0.07	0.05	0.06	0.03	0.04	0.04	0.03	0.04	0.04	626
translocation	0.21	0.12	0.15	0.20	0.11	0.14	0.14	0.13	0.13	0.15	0.13	0.14	208
violence	0.42	0.11	0.17	0.42	0.07	0.12	0.14	0.08	0.10	0.18	0.05	0.08	73

Table 10
Macro results across event classes for two runs of gpt-5.1 in zero-shot setting and two runs in few-shot setting with temperature=0. Metric scores are calculated in the same way as in Table 3.