Generating a Synthetic Dataset for Named Entity Recognition Using Large Language Models

Savkin Maksim

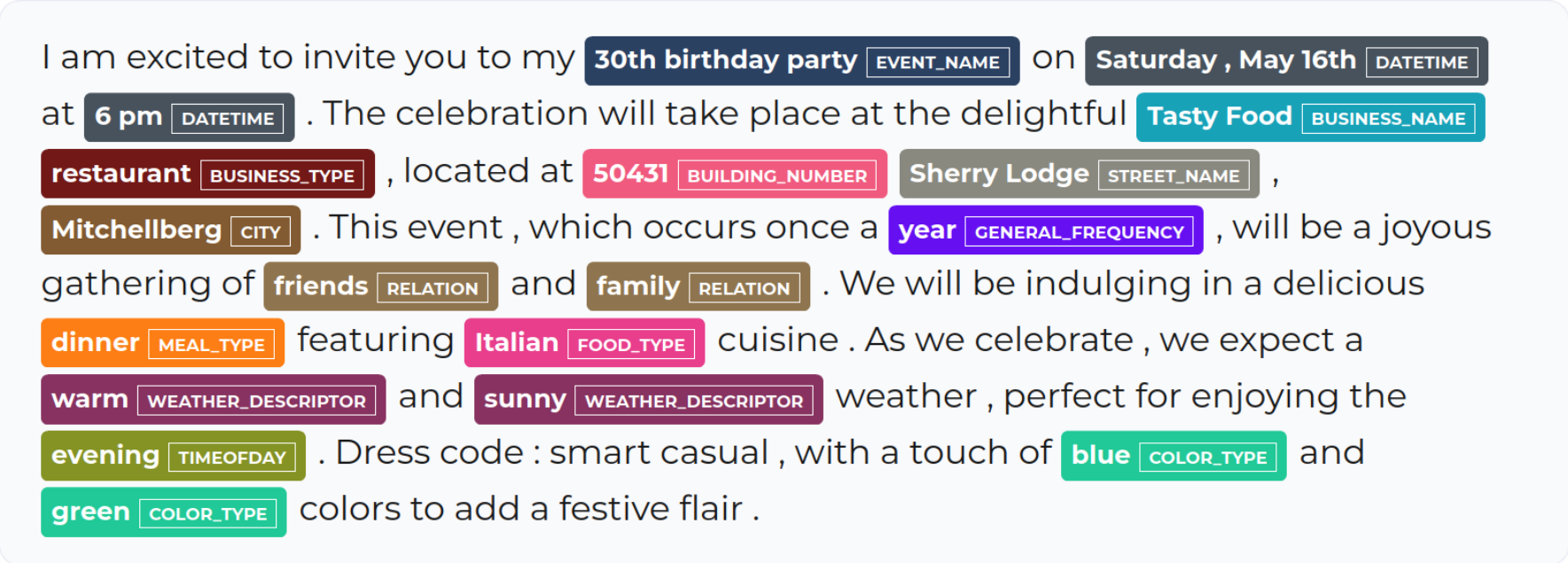
Center for Cognitive Modeling, MIPT

2024 – Scientific research seminar

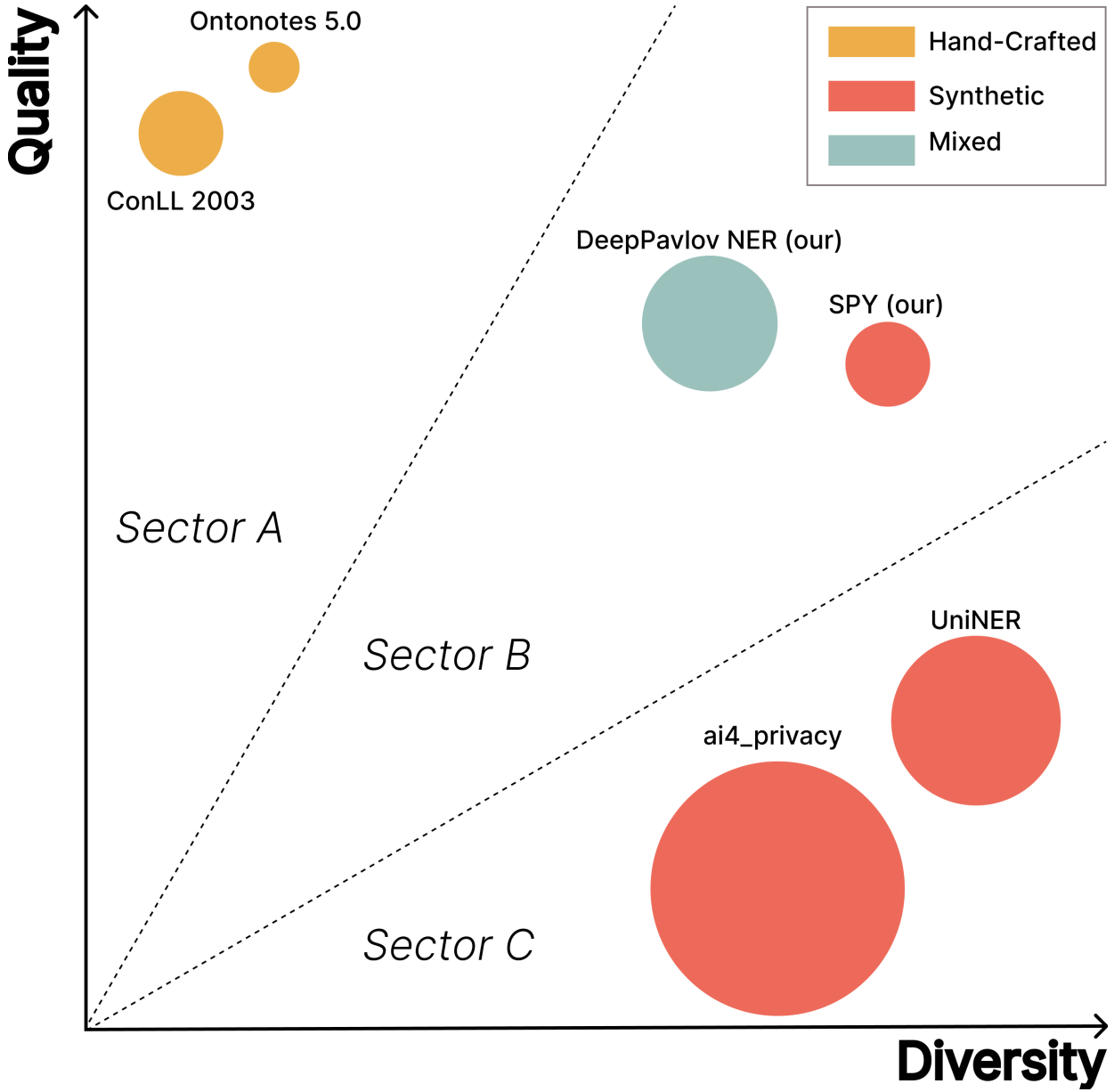
Introduction

Named Entity Recognition (NER) is a task of identifying named entities in the text and classifying them into types. Application areas are:

- Personal Information Identification (PII)
- Datasets Anonymization
- Dialogue Assistants



Motivation: Dataset Quality vs Diversity Tradeoff



Problem statement

The **objective** is to develop a domain-agnostic methodology for generating a synthetic dataset $(\hat{X}, \hat{Y})$, which would maximize the F1-score on target domain test set:

$$F1\left(M_{\hat{W}}(X_{test}(D)), Y_{test}(D)\right) \rightarrow max \tag{1}$$

where model M_W is trained by minimizing the cross-entropy loss $\forall x \in \hat{X}_{train}, y \in \hat{Y}_{train}$:

$$\hat{W} = \arg\min_{W} \left(\sum_{t_{id}=0}^{N} \left[- \sum_{l_{id}=0}^{K} y_{l_{id}}^{t_{id}} \log[M_W(x^{t_{id}})_{l_{id}}] \right] \right) \tag{2}$$

- $\mathcal{D}$ - target domain
- $X_{test}(\mathcal{D}) = \{\vec{x}^0, \vec{x}^1, ..., \vec{x}^M | \vec{x}^i \in \mathcal{T}\}$, $Y_{test}(\mathcal{D}) = \{\vec{l}^0, .., \vec{l}^M | \vec{l}^i_j \in \mathcal{C}^K\} = \{\vec{y}^0, \vec{y}^1, ...\vec{y}^M | \vec{y}^i_j \in [0,1]^K\}$ - test set of domain D
- $\mathcal{V}$ - vocabulary
- $\mathcal{T} = \{\vec{\mathcal{T}}_i\}$ - input texts, where $\vec{\mathcal{T}}_i = [t_0, t_1, t_2, ..., t_N], t_i \in \mathcal{V}$
- $\mathcal{C} = \{c_i\}_{i=0}^K$ - set of entity labels
- $M_W(x) : \mathcal{T} \rightarrow ([0,1]^K)^N$ - model with trainable weights

Task: Personal Information Identification

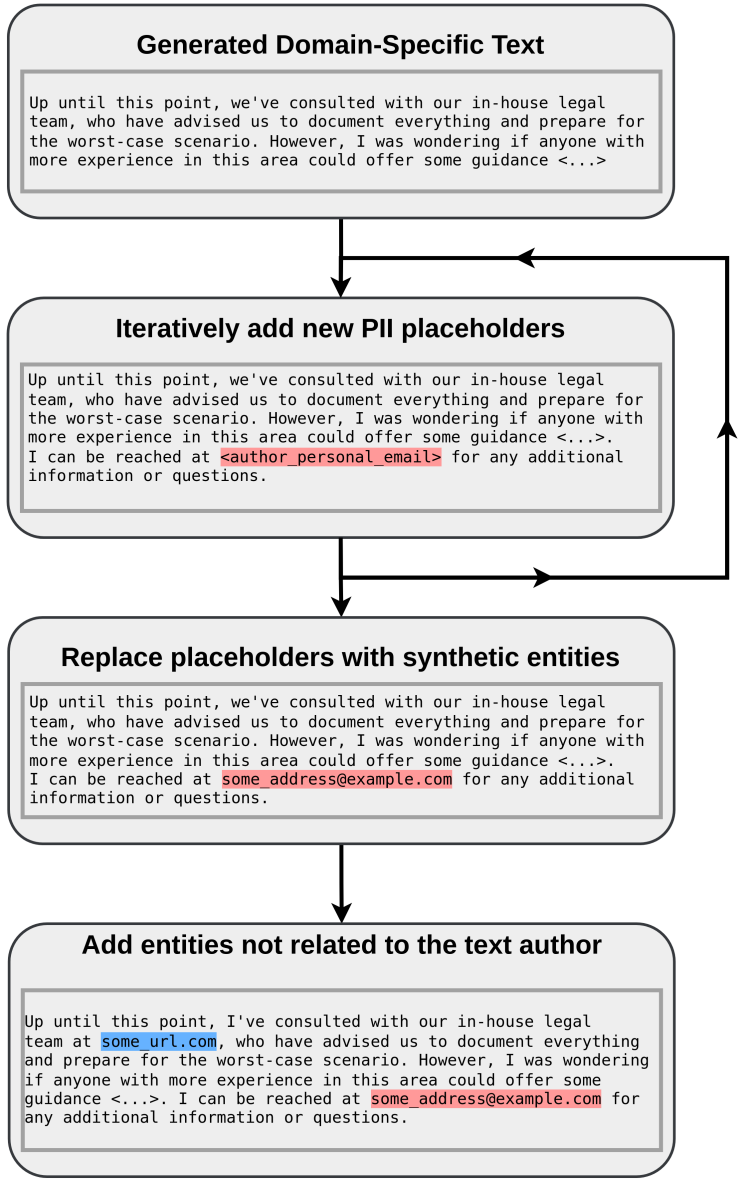
- NER subtask
- Lack of open datasets due to privacy reasons
- Limitations of automatic PII markup:
  - PII is rare in open datasets due to initial filtering
  - Sharing sensitive data with LLM providers may violate user agreement

PII vs NER	
a)	Apple technical support for education customers: 1-800-800-2775. Satya Nadella is CEO of Microsoft Corp.
b)	Mark Lee lives at 80081 Hardy Crescent Suite 420.

Table: Examples of NER entities - section **a)** and PII entities - section **b)**.

Methodology

- Domain-Specific Text Generation
  - Author BIO generation for increased diversity
  - Instructions to avoid generating PII
  - Domain-specific text generation with Llama-3-70B
- Iterative text updates
  - Increases entities density (similar to CoD[a] prompting)
  - Placeholders reduce entity paraphrasings
- Placeholders replacement:
  - PII entities are generated with Faker[b]
- Non-PII entities
  - Update text with non-PII entities



[a] Adams et al., "From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting".

[b] Faraglia, *Faker*.

Methodology: Domain-Specific Text Generation

Generate a biography of a fictional man named ***generated-name-goes-here***.

Occupation: any average job you can come up with
Personality: describe in 5 sentences

Present results in json format with fields ”occupation”: str, ”personality”: str

Figure: Prompt for biography generation.

Step 1) Look through the personality of the text author and pretend to be that person.

occupation: ***generated-occupation***
personality: ***generated-personality***

Step 2) Use the following instructions to generate a text:

domain-specific-instructions

Requirements:
- At any circumstance do not include any personal information in generated text.

Respond only with generated text with no commentary. Here goes your text:

Figure: Prompt for generating texts without personal information.

Methodology: Iterative text updates

Text: {}

Task: You are an author of the above Text. Your task is to add new placeholders in the Text from the list below. You will be penalized for mentioning any placeholders other than what is listed below!

Here is the list of placeholders representing your personal information:
author_personal_name - A full or partial name of the text author
author_personal_email - An author's email address
author_personal_username - An author's username on any website, social media etc.
author_personal_phone_number - A phone number associated with the author or his relatives
author_personal_url - A link to author's social media page or personal website
author_personal_address - A full or partial street address that is associated with the author, such as home address
author_personal_identifier - A number or sequence of characters that could be used to identify an author, such as a social security number or medical policy number

Requirements:
- Do NOT change existing placeholders
- Distribute placeholders evenly throughout your text, do not stack them all in one place
- New text must be more entity-dense than the previous one

Respond only with updated text with no commentary. Here goes an updated text:

Figure: Prompt for adding PII placeholders into the text.

Methodology: Non-PII entities

Text: {}

Task: You are given a Text, which contains author's personal information. Your task is to add new entities, which are not related to the text author. Generate entities using the following classes: name, email, username, phone number, url, address, identifier.

Requirements:
- At any circumstance DO NOT change author's personal information in the above text
- Newly generated entities should not disclose the personal information of the author of the text

Respond only with updated text with no commentary. Here goes an updated text:

Figure: Prompt for adding entities with personal information that are not related to text author.

Methodology: Iterative text updates

	Legal questions			Medical Consultations		
	upd pii 1	upd pii 2	final	upd pii 1	upd pii 2	final
name	0.58	1.06 (+0.48)	0.91 (+0.33)	0.69	1.12 (+0.43)	0.99 (+0.3)
email	1.03	1.15 (+0.12)	0.86 (-0.17)	1.01	1.12 (+0.11)	0.93 (-0.08)
username	0.91	1.14 (+0.23)	1.30 (+0.39)	0.80	1.16 (+0.36)	1.33 (+0.53)
phone number	0.87	1.1 (+0.23)	0.75 (-0.12)	0.88	1.12 (+0.24)	0.89 (+0.01)
personal url	1.07	1.34 (+0.27)	0.87 (-0.2)	1.03	1.32 (+0.29)	0.88 (-0.15)
personal address	0.71	1.19 (+0.48)	0.87 (+0.16)	0.73	1.28 (+0.55)	1.06 (+0.33)
personal identifier	0.39	0.98 (+0.59)	0.69 (+0.3)	0.53	1.05 (+0.52)	0.89 (+0.36)
average	0.79	1.14 (+0.35)	0.89 (+0.1)	0.81	1.17 (+0.36)	0.99 (+0.18)

Table: Average number of PII entities detected in texts generated by SPY prompting pipeline. Each entity type is counted separately. *upd pii* $\{k\}$ refers to the average number of PII entities in texts after k iterative updates of PII placeholders; *final* represents the average number of PII entities in texts that have completed all stages of the pipeline.

Experimental results

Entity	Metric	Medical Consultations			Legal Questions		
		Llama-3-70B	Presidio	DeBERTa (our)	Llama-3-70B	Presidio	DeBERTa (our)
Name	P	73.0	17.1	**86.9**	64.7	17.9	**87.4**
	R	62.9	80.4	**88.7**	68.9	79.4	**93.2**
	F1	67.6	28.2	**87.8**	66.7	29.2	**90.2**
Email	P	92.7	37.6	**97.6**	91.8	33.7	**92.1**
	R	90.9	92.2	**99.5**	88.5	91.8	**99.1**
	F1	91.8	53.4	**98.5**	90.1	49.3	**95.5**
Username	P	68.8	-	**92.1**	66.1	-	**90.3**
	R	70.4	-	**95.4**	59.7	-	**98.0**
	F1	69.6	-	**93.8**	62.7	-	**94.0**
URL	P	83.6	6.9	**97.5**	84.5	7.9	**94.4**
	R	91.9	19.4	**98.9**	92.5	21.3	**99.0**
	F1	87.5	10.2	**98.2**	88.3	11.5	**96.7**
ID	P	91.7	26.1	**96.7**	91.9	20.6	**93.0**
	R	75.1	38.9	**98.3**	62.2	34.4	**96.6**
	F1	82.6	31.2	**97.5**	74.2	25.8	**94.8**
Phone	P	89.8	37.4	**93.3**	85.7	34.1	**87.5**
	R	90.0	65.5	**96.9**	92.8	68.1	**98.7**
	F1	89.9	47.6	**95.0**	89.1	45.4	**92.8**
Address	P	**96.2**	-	89.3	**93.7**	-	88.3
	R	90.4	-	**95.1**	81.3	-	**94.5**
	F1	**93.2**	-	92.1	87.1	-	**91.3**

Table: Performance metrics. **Presidio**[1] is a Microsoft SDK for fast PII detection using NER, regex, rule-based logic. **LLaMA-3-70B-zero-shot**[2] is a zero-shot prompted LLM for PII task. **DeBERTa**[3] is a model cross-validated on different domains of SPY dataset. Blanks mean that entity class is not supported.

[1] Microsoft, *Presidio*.

[2] AI@Meta, "Llama 3 Model Card".

[3] He et al., *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*.

Ablation study: PII vs NER

Legal questions			Medical Consultations		
P	R	F1	P	R	F1
21.5	89.5	34.7	21.7	80.4	34.1

Table: UniNER[4] evaluation results on the SPY dataset. Metrics are calculated specifically for entities of the *NAME* type, using prompts from the original UniNER paper

UniNER "NAME" Class Prediction Example
Hi all, I 'm Nuria Batista, reaching out because I 'm in a bit of a tricky situation and I 'm hoping someone with legal expertise can offer some guidance. I 'm a marketing coordinator at an advertising agency, and one of our clients is accusing us of breach of contract. My team and I have reviewed the contract thoroughly, and we 're confident that we 've met all of the requirements. However, the client is still pushing for a refund and is threatening to take legal action against me, specifically at the office of attorney Emily Brown, located at 123 Main St, San Francisco, CA 94105.

Table: Name Nuria Batista is correctly classified as PII, while Emily Brown is misclassified due to the fact that UniNER doen't differentiate between PII and non-PII.

[4] Zhou et al., *UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition*.

Future work: PII Detection via LLM Probing

- PII as a hallucination
  - Uncertainty estimation for hallucination detection[5]
  - PII neurons localization[6] & Logits/Tuned Lens[7]

Method	Attribution $R_i^c(x)$
Gradient * Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$
Integrated Gradient	$(x_i - \bar{x_i}) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x_i})}\Big\|_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$
ϵ-LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x_i}) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$
Occlusion-1	$S_c(x) - S_c(x_{[x_i=0]})$

Figure: Mathematical formulation of five gradient-based attribution methods and of Occlusion-1.[8]

[5] Fadeeva et al., "LM-Polygraph: Uncertainty Estimation for Language Models".

[6] Chen et al., "Learnable Privacy Neurons Localization in Language Models".

[7] Belrose et al., "Eliciting Latent Predictions from Transformers with the Tuned Lens".

[8] Ancona et al., "Towards better understanding of gradient-based attribution methods for Deep Neural Networks".

Thank you for your attention!