

Identifying GDPR-Critical Tasks in Business Process Descriptions

An Automatic Approach Trained on a Synthetic Data Set

Leonard Nake, Stephan Kuehnel, René Theuerkauf, Stefan Sackmann

Business & Information Systems Engineering (2026)

Appendix (available online via <http://link.springer.com>)

Appendix A – Prompts Used for the Generation of the Synthetic Data Set

Table 1. Main Prompt

Target Seed: Industry: <X> Please use the following business process descriptions as stylistic examples only. Generate a new business process description for a different organization operating in the specified target industry. The generated description must: - Not reuse tasks, actors, or data objects described in the examples. - Introduce novel process logic, context, and operational structure. - Not resemble the examples in terms of process type or task sequence. - Be realistic and organizationally plausible within the given industry. - Focus on a single coherent business process. - Be written as continuous text without bullet points or task lists. Business Process Description 1 ... Business Process Description n

Table 2. Revision Prompt

Please revise the following business process description to improve correctness and internal consistency. Specifically: - Ensure that all tasks, actors, and resources are named consistently throughout the text. - Resolve any contradictions or unrealistic procedural elements. - Correct hallucinated information or factual errors. - Do not introduce new tasks, actors, or process steps unless necessary to fix an inconsistency. - Preserve the original structure and intent as much as possible.
--

Appendix B – Guideline for Labeling

Table 3. Heuristics for Labeling Tasks as GDPR-Critical or GDPR-Uncritical

Business Process Element	GDPR-Criticality	Heuristics
Data Object	GDPR-Critical	A task uses, creates, modifies, or stores a data object that explicitly refers to personal data (e.g., customer data, patient file, student record, application form, etc.).
		A task's data object refers to individualized records tied to a natural person (e.g., case file, contract, email, invoice of a customer, etc.).
	GDPR-Uncritical	A task's data object explicitly describes aggregated, anonymized, or statistical data (e.g., monthly sales summary, anonymized evaluation, etc.).
Task Name	GDPR-Critical	A task's verb indicates GDPR-relevant actions (e.g., collect, receive, enter, store, transfer, archive, delete, etc.) in combination with personal data.
	GDPR-Uncritical	A task's verb indicates actions unrelated to transmitting, storing, or otherwise processing (e.g., wait, trigger, perceive, etc.).
Resource	GDPR-Critical	A task's resource gains potential access to personal data during execution, even if data is not explicitly transmitted or modified.
	GDPR-Uncritical	A task's resource does not require access to personal data (e.g., technical or manual roles). Only relevant if ambiguous wording is used in the task (e.g., data, information) and no further data is available.
Business Process Context	GDPR-Critical	A task's context involves locations, deliveries, appointments, or communication if the execution requires storing or using addresses or contact details tied to identifiable individuals.
		A task's context indicates that the tasks deals with personal data although the task itself is ambiguous (e.g., receive information in a process about customer orders)
	GDPR-Uncritical	A task's context indicates that the tasks deals exclusively with generalized, aggregated, technical, or organizational data although the task itself is ambiguous (e.g., receive information in a process inventory management)

Appendix C – Model Descriptions

Table 4. Selection of Models for the Evaluation of our Approach

Model Architecture	Model	Description
DistilBERT	distilbert/distilbert-base-uncased	A lighter and faster version of BERT. It is designed to maintain performance as far as possible while reducing computational costs.
ALBERT	albert/albert-base-v2	A more efficient and scalable version of BERT. It is optimized to reduce memory usage and improve performance.
BERT	google-bert/bert-base-uncased	Baseline model. Provides good results for various NLP tasks.
RoBERTa	FacebookAI/roberta-base	An optimized version of BERT. It is designed to enhance performance by refining the training strategy, providing high accuracy but extending computation.
DeBERTa	microsoft/deberta-v3-xsmall-mnli-fever-anli-ling-binary	Improved version of RoBERTa. It is designed to be an efficient model that provides strong performance despite fewer computation times.
LUKE	studio-ousia/luke-base	Alternative transformer model. It treats words and entities as independent tokens, enhancing performance for knowledge-intensive NLP tasks.
ELECTRA	google/electra-base-discriminator	Alternative transformer model. It is designed to improve efficiency and performance by introducing a sample-efficient pretraining method.
LegalBERT	nlpaueb/legalbert-base-uncased	Domain-specific variant of BERT. It is pre-trained on legal texts, improving the performance for NLP tasks in the legal domain.
LegalBERTEURLEX	nlpaueb/bert-base-uncased-eurlex	Domain-specific variant of BERT. Fine-tuned on legal texts published by the European Union, improving performance for NLP tasks related to EU law.
LegalBERTCHR	nlpaueb/bert-base-uncased-echr	Domain-specific variant of BERT. Fine-tuned on cases of the European Court of Human Rights (ECHR), improving performance for NLP tasks related to EU Human Rights law.
LegalBERTCONTRACTS	nlpaueb/bert-base-uncased-contracts	Domain-specific variant of BERT. Fine-tuned on contracts from the United States, improving performance for NLP tasks related to contractual information.
LegalBERTGDPR	AndreaSimeri/LegalBERT-GDPR	Domain-specific variant of BERT. Fine-tuned on legal texts of the GDPR, improving performance for NLP tasks related to privacy law and data protection.
FinancialBERT	ahmedrachid/FinancialBERT	Domain-specific variant of BERT. Fine-tuned on financial data, improving performance for NLP tasks related to finance.

Appendix D – Comparison of Transformer Model Performance

Table 5. Performance of Transformer Models Under a Process-Description-Level Split

Model	GDPR-Critical			GDPR-Uncritical			Macro Average				
	Prec.	Rec.	F1	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	AUC
DistilBERT	0.823	0.879	0.850	0.831	0.760	0.794	0.819	0.827	0.819	0.822	0.881
ALBERT	0.749	0.937	0.832	0.883	0.601	0.715	0.769	0.816	0.769	0.774	0.860
BERT	0.828	0.866	0.846	0.819	0.771	0.794	0.818	0.823	0.818	0.820	0.882
RoBERTa	0.870	0.898	0.884	0.865	0.829	0.847	0.864	0.867	0.864	0.865	0.921
DeBERTa	0.807	0.824	0.815	0.771	0.749	0.760	0.787	0.789	0.787	0.788	0.866
LUKE	0.875	0.852	0.864	0.819	0.845	0.832	0.849	0.847	0.849	0.848	0.925
ELECTRA	0.834	0.870	0.851	0.825	0.780	0.802	0.825	0.829	0.825	0.827	0.915
LegalBERT (LB)	0.731	0.868	0.734	0.780	0.596	0.675	0.731	0.756	0.731	0.734	0.829
LB EURLEX	0.869	0.835	0.852	0.800	0.840	0.820	0.838	0.835	0.838	0.836	0.886
LB ECHR	0.808	0.837	0.822	0.783	0.747	0.764	0.792	0.795	0.792	0.793	0.868
LB CONTRACTS	0.768	0.911	0.833	0.852	0.651	0.738	0.781	0.810	0.781	0.785	0.896
LB GDPR	0.853	0.831	0.842	0.792	0.818	0.805	0.824	0.822	0.824	0.823	0.905
FinancialBERT	0.816	0.859	0.837	0.808	0.755	0.781	0.807	0.812	0.807	0.809	0.893