European University Institute
DEPARTMENT OF LAW

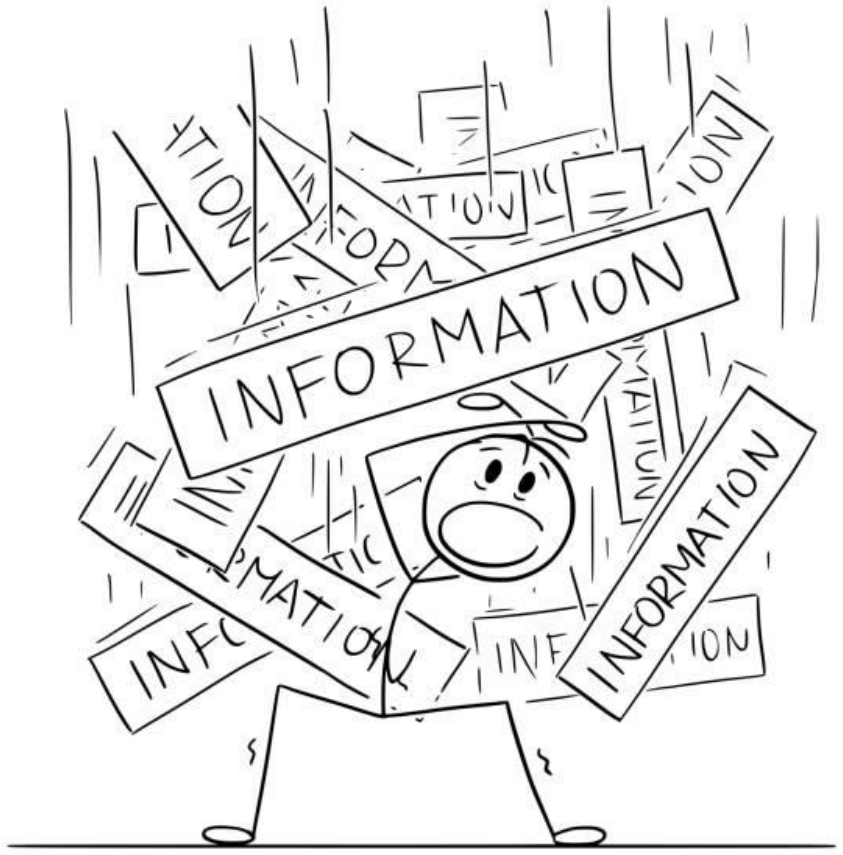
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LLMs and Transformer-based models for the analysis of Privacy Policies

Francesca Lagioia (francesca.lagioia@unibo.it)

Hey SKYNET, could you tell me the names of all corporations who will know that I just listened to Frankly, Mr. Shankly by The Smiths and list all the purposes for which they will use this data?

The Privacy Policy Landscape

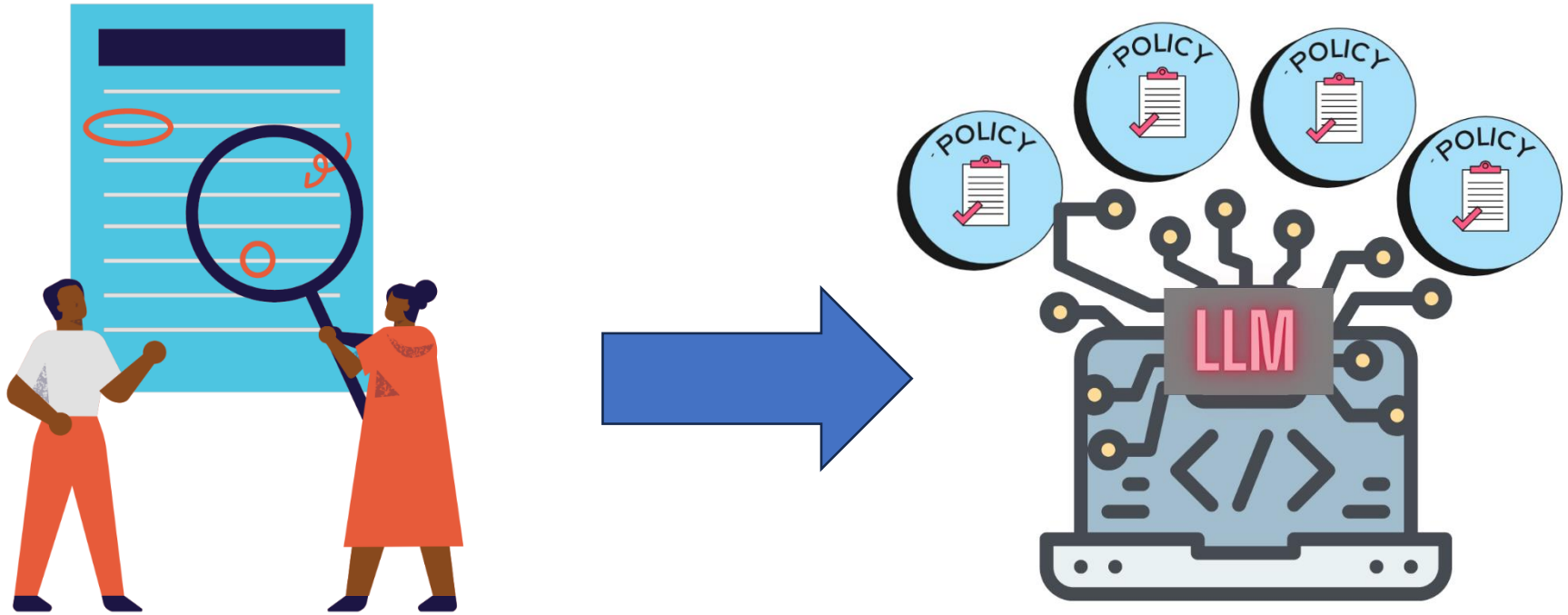


- Very long and complex
- Do not contain the necessary info
- Full of vague terms (“we collect data about your use of our service”) AND open-ended catalogs (“such as,” “including,” “for example”)
- Do not specify what categories of data are shared with whom (“we share your personal data with our marketing partners”)

- Consumers agree but don’t read
- NGOs have competence to control but lack resources
- Business keeps using unlawful clauses

Q_1

WHAT IF



Task1: Detecting and retrievieng information

Co-authors: Przemysław Pałka, Marco Lippi, Ruta Liepina, Giovanni Sartor

Legal Requirements for PPs

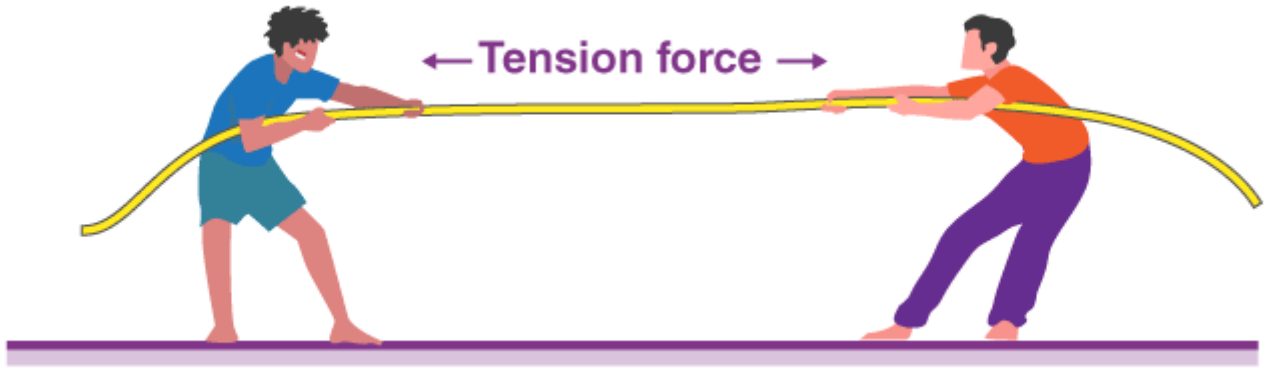
Comprehensiveness vs. Comprehensibility

Comprehensiveness:

A PP should contain all the info relevant for pondering whether to use a service.

Comprehensibility

A PP should be written in simple language and easy to understand, not excessively long.



Assessing Privacy Policies

9 simple questions a consumer should be able to get a clear answer to:

Q1: What data does the company process about me?

Q2: For what purposes does the company use my email address?

Q3: Who does the company share my geolocation with?

Q4: What types of data are processed on the basis of consent, and for what purposes?

Q5: What data does the company share with Facebook?

Q6: Does the company share my data with insurers?

Q7: What categories of data does the company collect about me automatically?

Q8: How can I contact the company if I want to exercise my rights?

Q9: How long does the company keep my delivery address?

Assessing Privacy Policies: Scenarios

- **Scenario 1:** Human Evaluation of Existing Privacy Policies
- **Scenario 2:** LLMs and Mock Privacy Policy
- **Scenario 3:** LLMs and Real Privacy Policies

Scenario 1: Human Evaluation of Existing Privacy Policies

- **Goal:** to what extent legal experts are able to answer the 9 questions from current PPs?
- 5 PPs from food delivery sector (Deliveroo, DoorDash, Glovo, Just Eat, and Wolt)
- Length: 4'379 words (Deliveroo)- 13'765 words (Glovo)
- Each policy independently evaluated by 2 legal experts using the 9 mentioned questions
- **PP Evaluation Criteria:**
  a) unambiguous (clear) information (i.e., definite answer could be given)
  b) ambiguous (impossible to know) information (i.e., no definite answer could be provided)
- Each expert were required to add justifications for the evaluations (perfect agreement)

Scenario 1:Results

Question	Just Eat	DoorDash	Wolt	Glovo	Deliveroo
Q1	✗	✗	✗	✗	✗
Q2	✗	✗	✗	✗	✗
Q3	✗	✗	✗	✗	✗
Q4	✗	✗	✗	✗	✗
Q5	✗	✗	✗	✗	✗
Q6	✓	✗	✓	✓	✓
Q7	✗	✗	✗	✗	✗
Q8	✓	✓	✓	✓	✓
Q9	✗	✗	✗	✗	✗

Table 1: Legal expert evaluation of privacy policies.

- Q6: Data sharing with insurers
- Q8: Contact information to exercise rights

The Legal Proposal

The Law should require Fully Comprehensive Privacy Policies

The Mock PP: Orderoo INC.

Each paragraph should contain:

A. data category

B. source of data

C. purpose of processing + explanation

D. legal basis + explanation

E. storage period

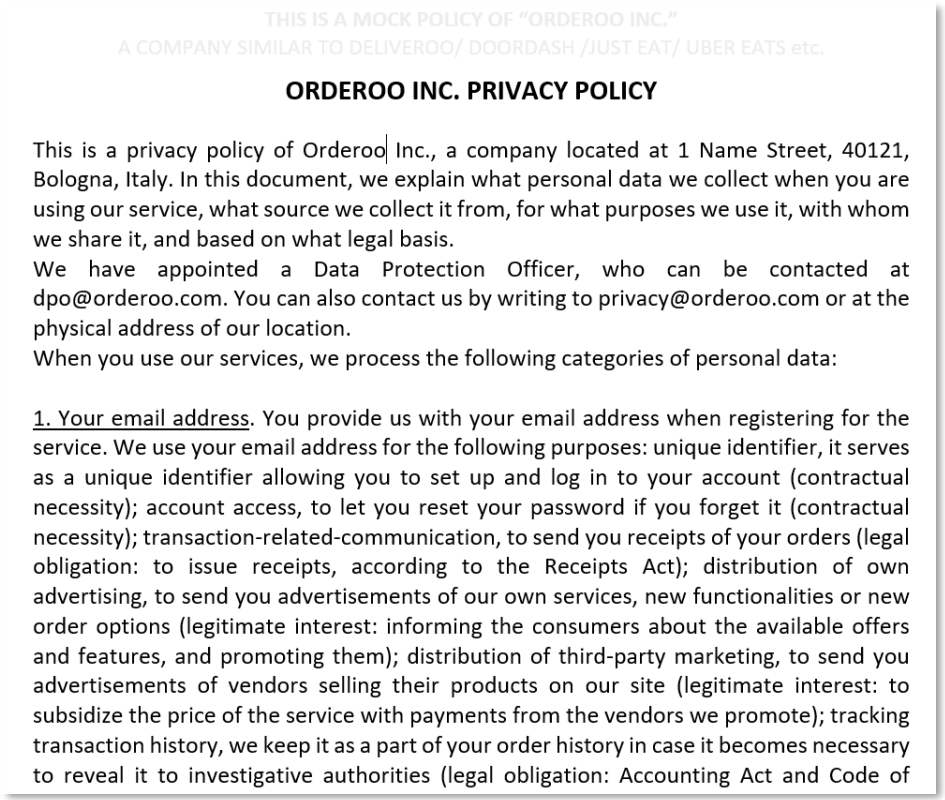
F. sharing

   i. the recipient's identity

   ii. their role (processor and controller)

   iii. the purpose of sharing

   iv. legal basis



THIS IS A MOCK POLICY OF "ORDEROO INC."
A COMPANY SIMILAR TO DELIVEROO/ DOORDASH /JUST EAT/ UBER EATS etc.

ORDEROO INC. PRIVACY POLICY

This is a privacy policy of Orderoo Inc., a company located at 1 Name Street, 40121, Bologna, Italy. In this document, we explain what personal data we collect when you are using our service, what source we collect it from, for what purposes we use it, with whom we share it, and based on what legal basis.
We have appointed a Data Protection Officer, who can be contacted at dpo@orderoo.com. You can also contact us by writing to privacy@orderoo.com or at the physical address of our location.
When you use our services, we process the following categories of personal data:

1. Your email address. You provide us with your email address when registering for the service. We use your email address for the following purposes: unique identifier, it serves as a unique identifier allowing you to set up and log in to your account (contractual necessity); account access, to let you reset your password if you forget it (contractual necessity); transaction-related-communication, to send you receipts of your orders (legal obligation: to issue receipts, according to the Receipts Act); distribution of own advertising, to send you advertisements of our own services, new functionalities or new order options (legitimate interest: informing the consumers about the available offers and features, and promoting them); distribution of third-party marketing, to send you advertisements of vendors selling their products on our site (legitimate interest: to subsidize the price of the service with payments from the vendors we promote); tracking transaction history, we keep it as a part of your order history in case it becomes necessary to reveal it to investigative authorities (legal obligation: Accounting Act and Code of

A sample from the fully comprehensive mock PP.

Scenario 2: LLMs and Mock Privacy Policy

- **Goal:** to what extent LLMs are able to answer the 9 questions from a fully comprehensive privacy policy?.
- **Tested LLMs:** GPT-4, Llama-7B, Llama-13B, Llama-70B, Mistral-7B, and Momo-70B (Issue: context window and number of tokens)
- **Selected LLMs:** GPT–4 and Lama2–7b
- **5 iterations** to assess variability across different runs

The experimental set up

Prompt: `In answering the questions please rely solely on the information included in the text and not your knowledge from other sources; please read the document carefully and mention everything, do not omit any information included in the text; please do not shorten or simplify the answers.`

Scenario 2: GPT–4 Results on Mock PP

$$= \frac{TP}{TP+FP}$$

$$= \frac{TP}{TP+FN}$$

$$= \frac{2PR}{P+R}$$

	Precision	Recall	F1	
Q1	100.0	100.0	100.0	
Q2	100.0	100.0	100.0	
Q3	100.0	100.0	100.0	
Q4	96.7	92.0	94.3	Data processed based on consent
Q5	94.4	90.0	92.1	Data shared with facebook
Q6	100.0	100.0	100.0	
Q7	100.0	78.2	87.8	Data collected automatically
Q8	100.0	100.0	100.0	
Q9	100.0	100.0	100.0	

- GPT answered the majority of questions correctly (33 of 45 questions)
- It followed the prompt instructions carefully, i.e., it used only the information available in the text and did not shorten or simplify the answers

Scenario 2:LLama2–7B Results on Mock PP

Llama2-7B provided only 1 correct answer out of 45 questions (9Q + 5 runs)

Error analysis:

- FN (Missing Info) = 91%
- HAL (Answers not based on the analysed policy) =62%
- SUM = 27%
- FP (information present in the policy but irrelevant for question) = 38%
- INF (inferences) =13%

Scenario 3:LLMs and Real Privacy Policies

- **Goal:** to what extent LLMs are able to answer the 9 questions from current privacy policies?
- 5 documents: Deliveroo, DoorDash, Glovo, Just Eat, and Wolt
- Selected LLMs: GPT-4 and Llama2-7B
- Same prompt as for the experiment on the mock policy
- Legal expert evaluation of the LLMs answers against the real PPs

Scenario 3: Results

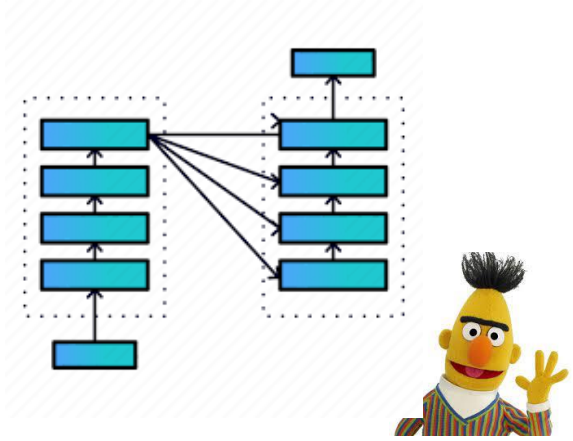
- GPT-4 provided 9 correct answers, while Llama2-7B only 1 correct answer out of 45 questions in total

- Error analysis:
  - **FN**: GPT-4 47% vs LLama2-7B 64%
  - **FP**: GPT-4 20% vs Llama2-7B 58%
  - **HAL**: GPT-4 0% vs Llama2-7B 38%
  - **SUM:** GPT-4 24% vs LLama2-7B 11%,
  - **INF**: GPT-4 33% vs Llama2-7B 31%

In a world of AI-readers it is possible to fully achieve both comprehensiveness and comprehensibility, IF privacy policies are clear and not vague!

Q_2

WHAT IF



Task2: Detection and Legal Assesment (GDPR violation)

Co-authors: Grundler G., Liepina R., Musicco M., Galassi A., Sartor G., Torroni P.

Comprehensiveness Requirement (art. 13-14 GDPR)

Privacy policies should contain all the information under art. 13-14 GDPR.

- **Focus**: the Categories of data collected (<cat> = 749)
- **Corpus**: 30 PPs in Eng
- **Selection criteria:**
  - Number of users
  - Service global relevance

- **Guidelines** for the assesment

1. Fully Informative: IF the <cat> are fully comprehensive and not vague (LEVEL=''1'') [exhaustive list and closed terms e.g. geo info] –181/749 clauses
2. Insuff. Informative: in all the other cases (LEVEL=''2'') [open list and open terms e.g., usage info]– 568/749 clauses

- **Hyerarchical levels of annotation**: 2 taggers 13/30 double marked
- **Agreemenet**: Cohen's κ= 0.97 <cat>– 0.72 (Kind)

Experimental Setting

6 TASKS

- CAT classification: given a sentence, classify it as CAT or non-CAT.
- LEVEL classification: given a CAT sentence, classify it as sufficiently informative (LEVEL= "1") or insufficiently informative (LEVEL="2").
- TYPE classification: given a CATEGORY or SUBCATEGORY, classify it as Open or Closed.
- KIND classification: given a CATEGORY or SUBCATEGORY, classify it as one of the 30 possible KINDs
- CATEGORY/SUBCATEGORY detection: given a CAT sentence, find the spans of text corresponding to CATEGORies and SUBCATEGORies.
- SPECIFICATION detection: given a CAT sentence, find the spans of text corresponding to SPECIFICATIONs.

Experimental Setting

For all tasks, we experimented 4 models:

- **Detection tasks:** evaluated in 2 settings
  1. the **BIO tagging format**: each token/word is classified as B-Class (begin), I-Class (inside) or O-Class (outside) for each class in question
  2. (ii) the IO tagging format, where the B tokens are tagged as I, resulting in I-Class and O-Class only.
- **Train-validation-test splits,** at document level => sentences of the same document belong to the same split.
- Splits manually created to
  1. balance their composition, (LEVEL and KIND)
  2. include as many double-tagged documents as possible in the test and validation sets, to increase their quality.

Experimental Setting

For all tasks, we experimented 4 models:

- ❖ 2 BERT-based: Distil-RoBERTa + LEGAL-BERT
  - Fine-tuned with a sequence classification head for the first 4 tasks (classifications) and a token classification head for the last 2 (detections).
  - Trained for 10 epochs in classification tasks and 20 epochs in detection tasks, with early stopping, a learning rate of 2e−5 and a batch size of 4 for LEVEL classification and 8 otherwise.
- ❖ 2 LLMs: Gemini 1.5 Flash + Meta Llama 3.1
  - Prompt-tuning based on the results on the validation set.
  - Prompts are mostly based on the definitions of the classes in the guidelines.
  - In few-shot mode, they contain all the examples of the training set (for CAT classification only the positive examples, i.e., the CAT tags)

Experimental Results

Model	Cat			Level			Type		
	cat	*non-cat*	Avg.	*1*	*2*	Avg.	*Closed*	*Open*	Avg.
Majority baseline	0.00	0.92	0.46	0.00	0.86	0.43	0.00	0.72	0.36
Random baseline	0.24	0.64	0.44	0.33	0.63	0.48	0.45	0.53	0.49
LEGAL-BERT	0.75	**0.96**	0.86	**0.77**	**0.92**	**0.84**	0.80	0.81	0.81
DistilRoBERTa	**0.77**	**0.96**	**0.87**	0.65	0.89	0.77	**0.82**	**0.82**	**0.82**
Gemini zero-shot	0.58	0.90	0.74	0.40	0.63	0.51	0.72	0.69	0.70
Gemini few-shot	0.72	0.95	0.84	0.58	0.86	0.72	**0.82**	0.81	**0.82**
Llama zero-shot	0.40	0.66	0.53	0.38	0.71	0.54	0.68	0.32	0.50
Llama few-shot	0.43	0.70	0.57	0.44	0.81	0.62	0.70	0.56	0.63

Table 4. Results for the CAT, LEVEL and TYPE classifications. We report the F1 score for the classes, along with their macro average.

Experimental Results

- **Gemini** few-shot reaches the best macro f1 score, closely followed by both variations of its zero-shot mode.
- **Llama** is the second- best model.
- **LEGAL-BERT and DistilRoBERTa** only get to 0.42 and 0.40 with good performance the more represented classes (e.g., DeviceInfo, Gen, GeoInfo, HealthFitness, UsageData and UserGenerated)
- They completely fail at classifying the classes with just a few or no training samples.

Kind	Maj	Rand	LB	DB	Gemini			Llama		
					$zero_l$	zero	few	$zero_l$	zero	few
AudioTyping	0.00	0.00	0.22	0.00	0.15	**0.29**	0.25	0.25	0.15	0.13
BasicAccountInfo	0.00	0.08	0.43	0.41	0.39	**0.51**	0.44	0.23	0.33	0.27
CommunicationProv	0.00	0.04	0.54	0.50	0.41	**0.67**	0.60	0.55	0.57	0.56
ContactInfo	0.00	0.00	0.56	0.54	0.45	**0.59**	0.48	0.45	0.46	0.47
ContactList	0.00	0.00	0.35	0.35	0.75	0.83	**0.97**	0.47	0.67	0.57
ContentPreferences	0.00	0.00	0.22	0.42	0.42	0.33	**0.50**	0.29	0.33	0.26
CriminalRecord	0.00	**0.06**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deidentified	0.00	0.00	0.29	0.00	**0.50**	0.29	0.00	0.00	0.00	0.00
Demographic	0.00	0.00	0.58	0.55	0.63	0.54	**0.66**	0.46	0.46	0.48
DeviceInfo	0.00	0.10	0.80	0.78	0.83	**0.88**	**0.88**	0.76	0.74	0.77
Financial	0.00	0.00	0.40	0.13	0.54	0.65	**0.80**	0.57	0.57	0.59
Gen	0.17	0.02	0.71	0.72	0.00	**0.76**	0.74	0.35	0.47	0.37
GeoInfo	0.00	0.05	0.79	0.69	0.79	0.77	**0.92**	0.82	0.90	0.90
Gov	0.00	0.00	**0.50**	0.33	0.25	0.25	0.18	0.00	0.33	**0.50**
HealthFitness	0.00	0.04	0.69	0.75	0.81	0.87	0.86	**0.90**	0.87	0.89
IdentityVerificationInfo	0.00	0.00	0.19	0.10	0.20	**0.43**	0.35	0.11	0.29	0.11
Images	0.00	0.00	0.56	0.76	0.77	0.69	**0.88**	0.65	0.73	0.76
InternetHistory	0.00	0.00	0.00	0.00	0.56	0.80	**0.84**	0.42	0.37	0.43
Languageanalysis	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
LicensingInfo	0.00	0.00	0.00	0.00	**1.00**	0.67	0.67	0.00	**1.00**	**1.00**
ListFriendsInfo	0.00	0.00	0.00	0.00	0.67	0.67	**1.00**	0.00	**1.00**	**1.00**
Metadata	0.00	0.00	0.65	**0.67**	0.23	0.54	**0.67**	0.26	0.44	0.22
Payment	0.00	0.08	0.65	0.59	0.70	0.78	**0.79**	0.72	0.75	0.69
Performance	0.00	0.05	0.56	0.56	0.59	**0.86**	**0.86**	0.67	0.38	0.15
Purchase	0.00	0.03	0.64	0.58	0.68	**0.69**	**0.69**	0.57	0.47	0.50
Settings	0.00	0.00	0.72	0.55	0.74	0.74	**0.83**	0.67	0.69	0.64
SocialInteraction	0.00	0.05	0.17	**0.35**	0.21	0.30	0.29	0.20	0.00	0.18
UsageData	0.00	0.12	**0.65**	0.60	0.58	0.63	0.64	0.53	0.55	0.59
UserGenerated	0.00	0.02	0.60	0.59	0.62	0.54	**0.76**	0.52	0.58	0.57
UserProfileInfo	0.00	0.11	0.33	0.43	0.34	**0.50**	0.43	0.39	0.44	0.45
Avg	0.01	0.03	0.42	0.40	0.53	0.57	**0.60**	0.39	0.48	0.47

Table 5. Results for the KIND classification task. We report the F1 score for the classes, along with their macro average. In the Gemini and Llama columns, $zero_l$ refers to the zero-shot experiment with class labels only, while the prompt for *zero* also contains a short description of each class.

Experimental Results

CATEGORY/SUBCATEGORY detection:

1. DistilRoBERTa in both settings
2. LEGAL-BERT.

Model	Category-Subcategory						Specification			
	B-C	*B-S*	*I-C*	*I-S*	*O*	Avg.	*B-Sp*	*I-Sp*	*O*	Avg.
Majority baseline	0.00	0.00	0.00	0.00	0.69	0.14	0.00	0.00	0.74	0.25
Random baseline	0.07	0.13	0.10	0.20	0.29	0.16	0.04	0.35	0.41	0.27
LEGAL-BERT	0.60	0.80	0.53	0.73	0.86	0.70	0.70	**0.92**	**0.93**	**0.85**
DistilRoBERTa	0.63	**0.81**	0.53	**0.74**	**0.88**	**0.72**	**0.73**	0.87	0.90	0.84
Gemini zero-shot	0.35	0.54	0.21	0.42	0.76	0.46	0.00	0.73	0.87	0.53
Gemini few-shot	**0.64**	0.75	**0.57**	0.65	0.85	0.69	0.42	0.81	0.89	0.71
Llama zero-shot	0.40	0.40	0.26	0.32	0.75	0.43	0.02	0.43	0.78	0.41
Llama few-shot	0.29	0.34	0.37	0.30	0.76	0.41	0.17	0.62	0.82	0.54
			I-C	*I-S*	*O*	Avg.		*I-Sp*	*O*	Avg.
Majority baseline			0.00	0.00	0.69	0.23		0.00	0.74	0.37
Random baseline			0.19	0.35	0.41	0.32		0.44	0.53	0.48
LEGAL-BERT			0.59	**0.83**	**0.87**	0.76		**0.92**	**0.93**	**0.92**
DistilRoBERTa			**0.62**	0.81	**0.87**	**0.77**		0.89	0.92	0.90
Gemini zero-shot			0.22	0.40	0.76	0.46		0.83	0.87	0.85
Gemini few-shot			0.50	0.56	0.85	0.64		0.82	0.89	0.86
Llama zero-shot			0.26	0.31	0.75	0.44		0.54	0.78	0.66
Llama few-shot			0.30	0.28	0.76	0.45		0.65	0.82	0.73

Table 6. Results for the detection tasks, both in BIO and IO formats. We report the F1 score for the classes, along with their macro average. In the name of the classes, we use C for CATEGORY, S for SUBCATEGORY and Sp for SPECIFICATION.

SPECIFICATION :

1. LEGAL-BERT
2. DistilRoBERTa, (best model in detecting the B- SPECIFICATION.
3. Gemini almost identical scores between zero and few-shot modes in IO setting, but unable to detect B-SPECIFICATION in zero-shot BIO setting.

Conclusions

Existing PPs fail to provide meaningful information (also) due to the trade-off between comprehensiveness and comprehensibility

- Legal experts are (often) unable to answer questions due to vague or missing info in PPs
- When given well-structured, fully informative PPs, LLMs can provide consumers with meaningful and precise answers.
- Fine-tuned models (LEGAL-BERT and DistilRoBERTa) obtained the best results in almost all tasks (not for kind classification: high number + few samples for some)
- For KIND classification LLMs' ability to learn with very few examples emerges (e.g., Gemini surpasses fine-tuned models even in zero-shot mode, with just the classes' names as task description

In a world of AI-readers it is possible to fully achieve both comprehensiveness and comprehensibility.

- **The law can and should be revised!!!**
- **NGoS and DPA should be supported by AI-empowering tech**

Aknowledgement

This work has been partially supported by the following projects: PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D230051300\\01); PRIN2022 EQUAL – EQUitableALgorithms (Ref. Prot n. 2022KFLF3E$_$001 - CUP J53D23005560001); CompuLaw – Computable Law – funded by the ERC under the Horizon 2020 (Grant Agreement N. 833647); CLAUDETTE IV, founded by the EUI Research Council for founding; ``FAIR - Future Artificial Intelligence Research'' -- Spoke 8 ``Pervasive AI'', under the European Commission's NextGeneration EU programme, PNRR -- M4C2 -- Investimento 1.3, Partenariato Esteso, (PE00000013).