1 Cat classification

Model	Macro F1			Precision			Recall		
	cat	*non-cat*	Avg.	*cat*	*non-cat*	Avg.	*cat*	*non-cat*	Avg.
Majority baseline	0.00	0.92	0.46	0.00	0.85	0.43	0.00	**1.00**	0.50
Random baseline	0.24	0.64	0.44	0.16	0.87	0.51	0.54	0.51	0.52
LEGAL-BERT	0.75	**0.96**	0.86	**0.84**	0.96	**0.90**	0.68	0.97	0.82
DistilRoBERTa	**0.77**	**0.96**	**0.87**	0.79	0.96	0.87	0.75	0.97	**0.86**
Gemini zero-shot	0.58	0.90	0.74	0.49	0.95	0.72	0.73	0.86	0.79
Gemini few-shot	0.72	0.95	0.84	0.72	0.95	0.84	0.72	0.95	0.83
Llama zero-shot	0.40	0.66	0.53	0.25	**0.99**	0.62	**0.98**	0.49	0.73
Llama few-shot	0.43	0.70	0.57	0.27	**0.99**	0.63	0.97	0.55	0.76

Table 1: Results for the Cat classification task. We report the F1 score, precision and recall for the classes, along with their macro average.

2 Level classification

Model	Macro F1			Precision			Recall		
	1	*2*	Avg.	*1*	*2*	Avg.	*1*	*2*	Avg.
Majority baseline	0.00	0.86	0.43	0.00	0.76	0.38	0.00	**1.00**	0.50
Random baseline	0.33	0.63	0.48	0.25	0.76	0.51	0.49	0.54	0.51
LEGAL-BERT	**0.77**	**0.92**	**0.84**	**0.74**	**0.93**	**0.84**	**0.79**	0.91	**0.85**
DistilRoBERTa	0.65	0.89	0.77	0.69	0.88	0.78	0.62	0.91	0.76
Gemini zero-shot	0.40	0.63	0.51	0.29	0.81	0.55	0.62	0.52	0.57
Gemini few-shot	0.58	0.86	0.72	0.57	0.87	0.72	0.59	0.86	0.72
Llama zero-shot	0.38	0.71	0.54	0.31	0.80	0.55	0.49	0.64	0.57
Llama few-shot	0.44	0.81	0.62	0.42	0.82	0.62	0.46	0.79	0.63

Table 2: Results for the Level classification task. We report the F1 score, precision and recall for the classes, along with their macro average.

3 Type classification

Model	Macro F1			Precision			Recall		
	Closed	*Open*	Avg.	*Closed*	*Open*	Avg.	*Closed*	*Open*	Avg.
Majority baseline	0.00	0.72	0.36	0.00	0.56	0.28	0.00	**1.00**	0.50
Random baseline	0.45	0.53	0.49	0.43	0.55	0.49	0.47	0.51	0.49
LEGAL-BERT	0.80	0.81	0.81	0.83	0.79	0.81	0.78	0.83	0.81
DistilRoBERTa	**0.82**	**0.82**	**0.82**	**0.84**	0.80	**0.82**	0.79	0.85	**0.82**
Gemini zero-shot	0.72	0.69	0.70	0.70	0.71	0.71	0.74	0.67	0.70
Gemini few-shot	**0.82**	0.81	**0.82**	0.81	**0.83**	**0.82**	0.84	0.79	**0.82**
Llama zero-shot	0.68	0.32	0.50	0.54	0.68	0.61	**0.91**	0.21	0.56
Llama few-shot	0.70	0.56	0.63	0.61	0.72	0.67	0.83	0.46	0.64

Table 3: Results for the Type classification task. We report the F1 score, precision and recall for the classes, along with their macro average.

4 Detection tasks

4.1 Macro F1

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 4: 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.

4.2 Precision and Recall

Model	Precision						Recall					
	B-C	*B-S*	*I-C*	*I-S*	*O*	Avg.	*B-C*	*B-S*	*I-C*	*I-S*	*O*	Avg.
Majority baseline	0.00	0.00	0.00	0.00	0.53	0.11	0.00	0.00	0.00	0.00	**1.00**	0.20
Random baseline	0.04	0.11	0.07	0.22	0.53	0.19	0.26	0.17	0.17	0.19	0.20	0.20
LEGAL-BERT	0.59	0.79	0.47	0.70	**0.90**	0.69	0.61	**0.81**	**0.62**	**0.75**	0.82	0.72
DistilRoBERTa	0.59	0.83	0.49	0.80	0.86	0.71	**0.66**	0.79	0.58	0.69	0.90	**0.73**
Gemini zero-shot	0.70	0.83	**0.81**	0.84	0.62	0.76	0.24	0.40	0.12	0.28	0.98	0.40
Gemini few-shot	**0.73**	0.80	0.71	**0.86**	0.76	**0.77**	0.56	0.70	0.47	0.52	0.96	0.64
Llama zero-shot	0.67	0.82	0.74	0.85	0.60	0.74	0.29	0.26	0.16	0.20	0.99	0.38
Llama few-shot	0.22	**0.93**	0.31	0.76	0.66	0.58	0.44	0.20	0.46	0.19	0.91	0.44
			I-C	*I-S*	*O*	Avg.			*I-C*	*I-S*	*O*	Avg.
Majority baseline			0.00	0.00	0.53	0.18			0.00	0.00	**1.00**	0.33
Random baseline			0.12	0.37	0.54	0.35			0.39	0.34	0.33	0.35
LEGAL-BERT			**0.55**	0.83	0.89	**0.76**			0.63	**0.83**	0.86	0.77
DistilRoBERTa			0.51	**0.84**	**0.90**	0.75			**0.80**	0.79	0.84	**0.81**
Gemini zero-shot			0.49	0.53	0.62	0.55			0.14	0.32	0.98	0.48
Gemini few-shot			0.50	0.54	0.76	0.60			0.50	0.58	0.96	0.68
Llama zero-shot			0.46	0.55	0.60	0.54			0.18	0.22	0.99	0.46
Llama few-shot			0.22	0.52	0.66	0.47			0.50	0.20	0.91	0.53

Table 5: Results for the Category-Subcategory detection task, both in BIO and IO formats. We report the Precision and Recall scores for the classes, along with their macro average. In the name of the classes, we use C for Category and S for SubCategory.

Model	Precision				Recall			
	B-Sp	*I-Sp*	*O*	Avg.	*B-Sp*	*I-Sp*	*O*	Avg.
Majority baseline	0.00	0.00	0.58	0.19	0.00	0.00	**1.00**	0.33
Random baseline	0.02	0.39	0.57	0.33	0.37	0.32	0.32	0.34
LEGAL-BERT	0.66	0.88	**0.96**	**0.83**	0.73	**0.96**	0.89	**0.86**
DistilRoBERTa	**0.69**	**0.90**	0.89	**0.83**	**0.78**	0.85	0.92	0.85
Gemini zero-shot	0.00	0.86	0.85	0.57	0.00	0.64	0.90	0.51
Gemini few-shot	0.35	**0.90**	0.85	0.70	0.53	0.74	0.94	0.74
Llama zero-shot	0.02	0.77	0.68	0.49	0.05	0.30	0.91	0.42
Llama few-shot	0.16	0.88	0.73	0.59	0.18	0.48	0.95	0.54
		I-Sp	*O*	Avg.		*I-Sp*	*O*	Avg.
Majority baseline		0.00	0.58	0.29		0.00	**1.00**	0.50
Random baseline		0.41	0.57	0.49		0.48	0.49	0.49
LEGAL-BERT		**0.96**	0.89	**0.92**		0.90	0.95	**0.93**
DistilRoBERTa		0.91	**0.90**	0.91		**0.92**	0.89	0.90
Gemini zero-shot		0.84	0.85	0.84		0.81	0.90	0.85
Gemini few-shot		0.87	0.85	0.86		0.78	0.94	0.86
Llama zero-shot		0.76	0.68	0.72		0.42	0.91	0.67
Llama few-shot		0.87	0.73	0.80		0.51	0.95	0.73

Table 6: Results for the Specification detection task, both in BIO and IO formats. We report the Precision and Recall scores for the classes, along with their macro average. In the name of the classes, we use Sp for Specification.