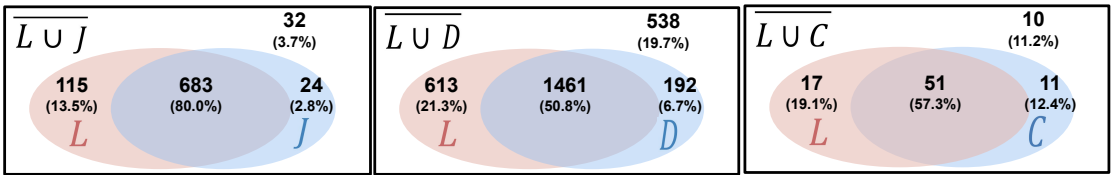
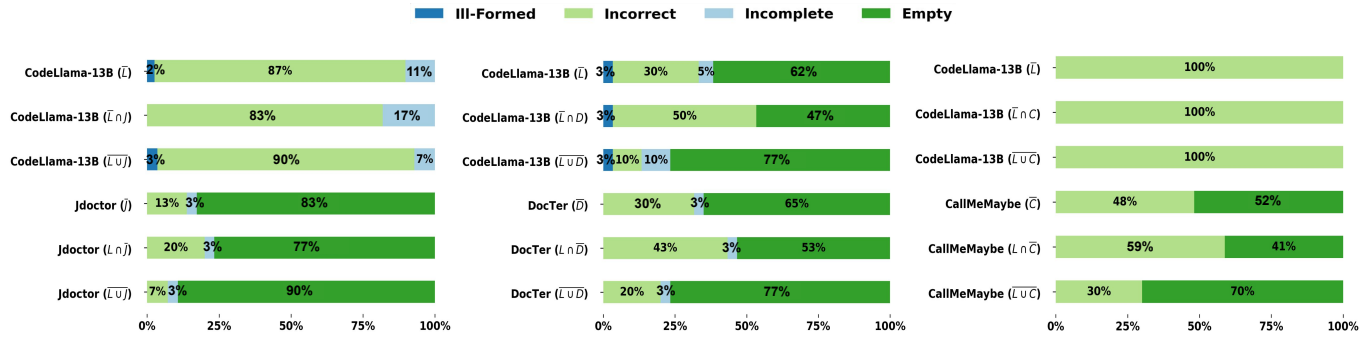


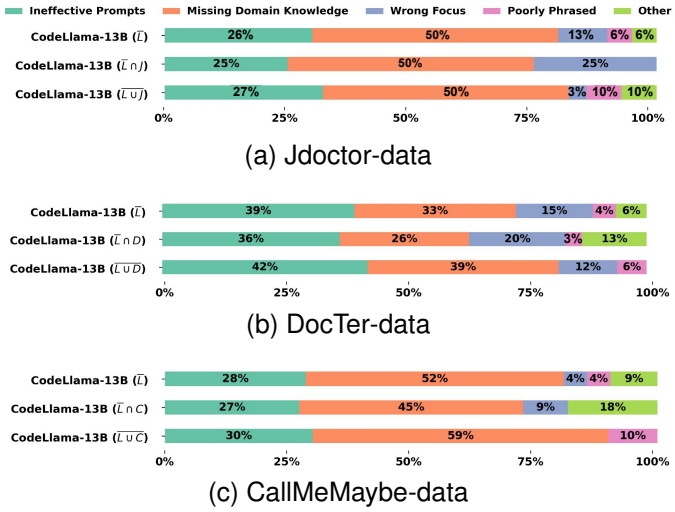
(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 1: Venn diagrams of specification generation. L: CodeLlama-13B; J: Jdoctor; D: DocTer; C: CallMeMaybe.



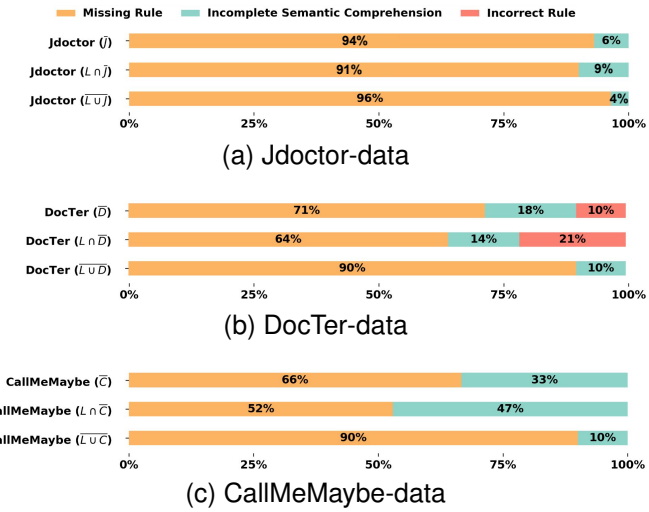
(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 2: Distributions of symptoms in failing cases across approaches and datasets. The y-axis is "approach (section)".



(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 3: Distributions of root causes of CodeLlama-13B's failures.



(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 4: Distributions of root causes for baselines' failures.

APPENDIX

RQ3: COMPARATIVE FAILURE DIAGNOSIS

Fig. 1 presents the comparative performance of the CodeLlama-13B-based LLM method and baseline methods as Venn diagrams. Fig. 1a, 1b, and 1c respectively illustrate the number of unique and shared successes and failures of specifications for CodeLlama-13B (L) against Jdoctor (J), DocTer (D), and CallMeMaybe (C). The numbers in the intersections ($L \cap J$ and $L \cap D$, and $L \cap C$) denote cases where both methods are correct, while numbers in sections $\overline{L \cup J}$, $\overline{L \cup D}$, and $\overline{L \cup C}$ indicate cases they both fail. The presented results are derived from the experiment using CodeLlama-13B with SR and $K = 60$ in RQ2 (Section V.B).

A. Failure Symptom Analysis

We conduct further analysis on the distributions of failure symptoms in various sections of the Venn diagrams (Fig. 1), and present our results in Fig. 2.

The failure symptoms are classified into four categories, "ill-formed", "incorrect", "incomplete", and "empty". The bars illustrate the distributions of failure symptoms for these methods (CodeLlama-13B, Jdoctor, DocTer, and CallMeMaybe) from different sections of Fig. 1. For example, bar "CodeLlama-13B ($\overline{L}$)" in Fig. 6a denotes the distributions of CodeLlama-13B's failure symptoms on Jdoctor-data cases where it fails. Bar "CodeLlama-13B ($\overline{L} \cap J$)" indicates the distributions of CodeLlama-13B's failure symptoms within the $\overline{L} \cap J$ section (CodeLlama-13B fails, and Jdoctor succeeds). When both fail, as in $\overline{L \cup J}$, "CodeLlama-13B ($\overline{L \cup J}$)" and "Jdoctor ($\overline{L \cup J}$)"

shows the failure symptoms categories for CodeLlama-13B and Jdoctor outputs, respectively.

Category "*ill-formed*" refers to invalidly formed generated specifications. For Jdoctor-data, this denotes syntactical errors or grammatical mistakes. For DocTer-data, it refers to improper specification forms, such as generating a boolean instead of an expected numerical range. "*Incorrect*" indicates specification errors, while "*Incomplete*" denotes the specification is a strict subset of the ground truth. For DocTer-data, if the ground-truth specifications for the *dtype* category include both `int32` and `int64`, generating only `int32` is incomplete, whereas generating `bool` is incorrect. "*Empty*" denotes a missing specification. For DocTer-data, each specification contains four categories (Section III.A.(2)). If the specification for one of the categories is empty while the ground truth is not, it is considered an empty type of failure.

a) LLMs are more likely to generate ill-formed and incomplete specifications: LLMs are more likely (by 5%) to generate incomplete specifications. For instance, when given the description "tuple/list of 1 int", CodeLlama-13B only generates `tuple` for the *structure* specification and misses the specification `list`. Such errors are unlikely to occur with rule-based methods, as they match the entire sequence of types and extract them directly from the document as specifications, while LLMs use sampling to decode outputs from a distribution, which may occasionally miss tokens.

b) Traditional techniques are much more likely to generate empty specifications than LLMs: For CodeLlama-13B, none of the failing cases for Jdoctor-data are empty, while only 61% of the failing cases for DocTer-data are empty. The reason is that the Jdoctor-data and CallMeMaybe-data examples provided to the LLMs are never empty and hence the LLMs always generate some results for queries for Jdoctor-data and CallMeMaybe-data. In contrast, some of the specification categories (e.g., `dtype`) of DocTer-data examples (provided to the LLMs) may be empty. The LLMs learn that empty is a possible result for DocTer-data.

The results are in line with the precision and recall in Table VII, that CodeLlama-13B has a much higher recall than DocTer (e.g., 84.4% versus 78.2%) with a comparable or slightly lower precision (e.g., 84.1% versus 85.4%).

B. Generalizibility

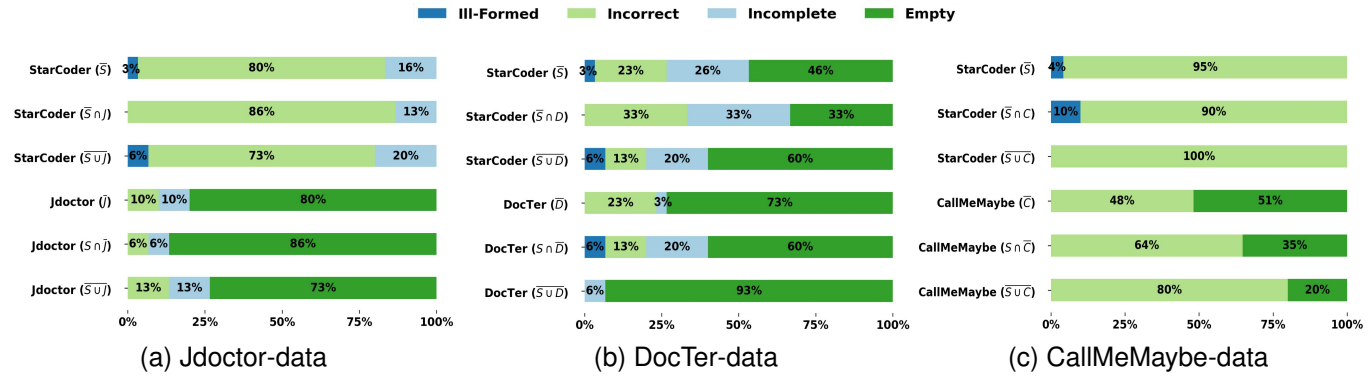
To validate the generalizability of our analysis and conclusions in RQ3 (Section VI), we extended our evaluation to include two additional models, StarCoder and GPT-3.5. We follow the same procedure to sample cases and manually investigate them. The results of the analysis of StarCoder's failure cases are shown in Figure 5 and Figure 7. The results of the analysis over StarCoder's failure cases are shown in Figure 6 and Figure 8.

RQ4: Model Comparison

We provide the complete results of RQ4 in Tables A, B, and C.

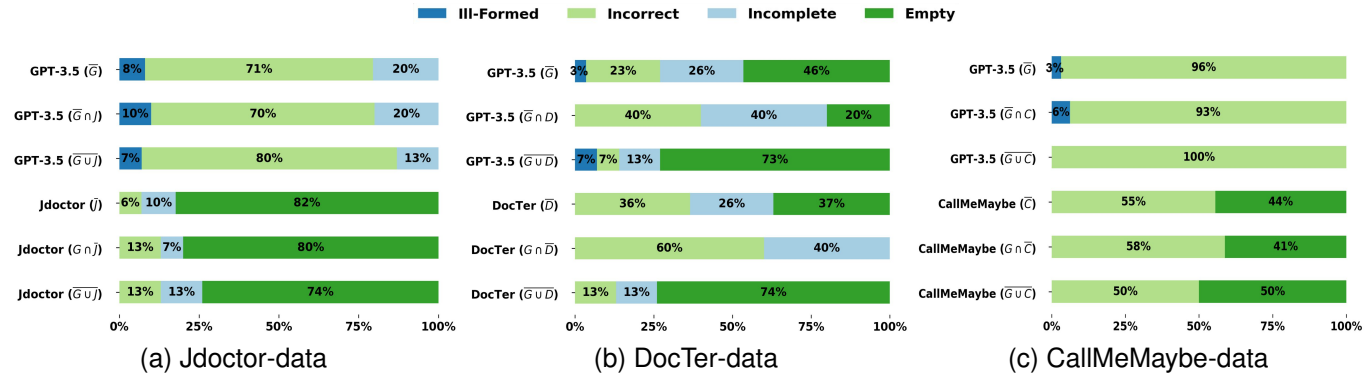
TABLE A: Comparison of different LLMs with SR on Jdoctor-data: Accuracy (%) and Cost.

Approach/ Model (+SR)		K	`@param`		`@return`		`@throws`		Overall		Cost
			Raw	Processed	Raw	Processed	Raw	Processed	Raw	Processed	($)
Jdoctor		-	97.0	97.0	69.0	69.0	79.0	79.0	83.0	83.0	-
GPT	4	10	88.1	95.9	53.2	72.7	84.1	91.5	80.2	**89.7**	19.5
		20	89.7	95.9	60.4	77.7	86.2	92.8	83.0	**91.2**	39.0
		40	89.3	95.9	58.3	75.5	88.8	94.5	84.0	**91.8**	78
		60	90.1	95.5	63.3	77.7	90.0	94.9	85.7	**92.3**	117
	3.5	10	86.9	93.5	60.8	70.9	81.6	91.0	80.6	**89.3**	1.3
		20	82.0	89.1	60.8	72.2	83.5	90.3	80.4	**87.9**	2.6
		40	83.6	89.6	59.5	74.7	84.5	88.8	81.3	**87.4**	5.2
		60	86.3	89.1	59.5	72.2	80.1	84.7	79.4	**84.4**	7.8
Codellama	13B	10	93.0	96.7	62.6	71.2	84.5	90.3	83.4	**89.0**	-
		20	94.7	**97.5**	63.3	73.4	86.4	91.3	85.0	**90.2**	-
		40	95.9	97.9	62.6	72.7	89.6	94.7	87.0	**82.0**	-
		60	95.9	98.8	65.5	75.5	90.3	96.0	87.9	**93.5**	-
	7B	10	92.2	96.3	60.4	71.2	84.5	90.5	82.8	**89.0**	-
		20	93.8	97.9	64.0	74.8	86.2	91.9	84.7	**90.8**	-
		40	93.8	97.5	61.9	75.5	89.2	92.6	86.1	**91.2**	-
		60	95.5	98.4	60.4	72.7	90.9	95.3	87.2	**92.5**	-
deepseek-coder	6.7B	10	90.9	95.9	61.9	74.1	84.3	91.1	82.5	**89.7**	-
		20	93.8	97.5	64.0	67.5	86.4	92.6	84.9	**90.0**	-
		40	93.4	96.7	60.4	65.5	90.5	94.9	86.4	**90.6**	-
		60	93.8	96.3	60.4	71.2	91.5	96.0	87.1	**92.0**	-
Llama3	8B	10	90.5	95.5	63.3	66.9	83.3	90.3	82.1	**88.0**	-
		20	93.8	97.5	66.2	69.8	86.7	92.6	85.4	**90.3**	-
		40	95.1	97.1	60.4	63.3	89.0	94.1	86.1	**89.9**	-
		60	95.9	96.4	65.5	77.7	90.0	94.5	87.1	**92.9**	-
Llama2	13B	10	91.4	96.3	58.3	66.9	81.6	85.8	80.6	**85.7**	-
		20	93.0	96.7	60.4	66.9	83.5	87.9	82.4	**87.0**	-
		40	94.2	96.7	59.0	65.5	86.9	91.9	84.4	**89.0**	-
		60	95.1	98.4	56.8	68.3	88.6	93.0	85.3	**90.5**	-
	7B	10	88.9	93.0	56.8	64.7	77.5	82.0	77.4	82.3	-
		20	91.4	95.9	56.1	66.9	82.8	86.7	80.9	**86.1**	-
		40	94.2	95.9	58.3	67.6	86.2	91.7	83.9	**89.0**	-
		60	95.1	97.5	56.1	67.6	86.9	91.7	84.2	**89.4**	-
StarCoder2	15B	10	91.8	97.1	61.9	71.9	83.1	89.8	82.1	**89.0**	-
		20	94.7	97.5	64.0	73.4	87.1	92.6	85.5	**90.9**	-
		40	95.1	97.5	64.0	78.4	89.4	94.3	86.9	**92.6**	-
		60	96.3	98.4 98.4	65.0	77.7	91.7	95.6	88.7	**93.5**	-
	7B	10	92.2	93.2	59.0	67.6	83.7	90.3	82.1	**87.4**	-
		20	93.8	93.2	63.3	71.9	85.8	91.7	84.4	**88.9**	-
		40	95.1	97.0	61.2	74.8	88.8	93.2	86.1	**91.3**	-
		60	95.9	97.9	64.0	76.3	90.5	94.5	87.7	**92.5**	-
StarCoder	16B	10	91.8	95.4	63.3	74.1	83.9	90.0	82.8	**88.9**	-
		20	94.2	97.9	65.5	78.4	86.2	91.1	85.1	**91.0**	-
		40	95.1	97.9	63.3	74.8	88.8	93.4	86.4	**91.7**	-
		60	95.9	98.4	64.7	77.7	90.3	94.7	87.7	**93.0**	-
CodeGen2	16B	10	93.4	97.8	59.5	68.4	81.6	85.0	82.2	**86.5**	-
		20	93.4	98.9	60.8	70.9	84.2	88.1	84.0	**89.0**	-
	7B	10	89.6	95.1	64.6	70.9	77.9	80.8	79.5	**83.5**	-
		20	91.8	96.7	64.6	72.2	84.5	87.6	84.2	**88.3**	-



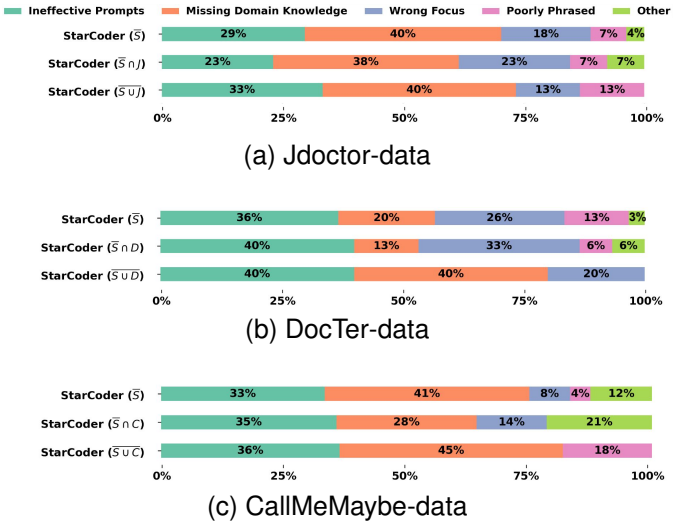
(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 5: Distributions of symptoms in failing cases (of StarCoder and baseline methods) across approaches and datasets. The y-axis is "approach (section)".



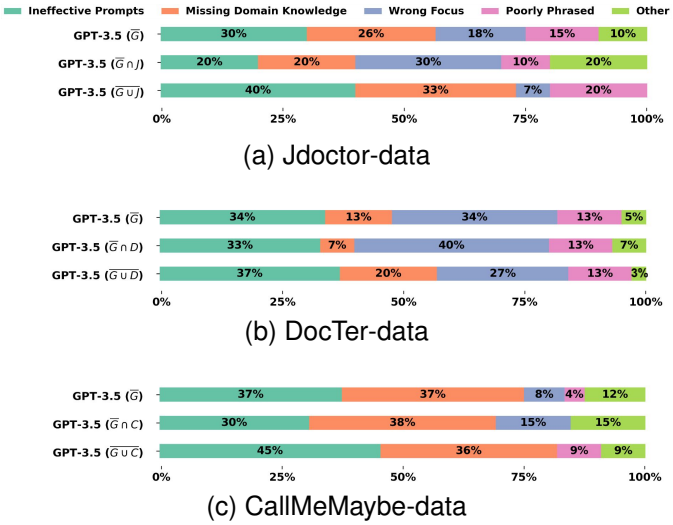
(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 6: Distributions of symptoms in failing cases (of GPT-3.5 and baseline methods) across approaches and datasets. The y-axis is "approach (section)".



(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 7: Distributions of root causes of StarCoder's failures.



(a) Jdoctor-data (b) DocTer-data (c) CallMeMaybe-data

Fig. 8: Distributions of root causes of GPT-3.5's failures.

TABLE B: Comparison of different LLMs with SR on DocTer-data: Precision/Recall/F1 (%), and Cost.

Approach / Model (+SR)		K	TensorFlow	PyTorch	MXNet	Overall	Cost ($)
DocTer		-	90.0/74.8/81.7	78.4/77.4/77.9	87.9/82.4/85.1	85.4/78.2/81.6	-
GPT	4	10	80.2/83.4/81.8	77.1/81.2/79.1	83.3/86.2/84.8	81.2/84.4/**82.8**	13
		20	82.4/85.6/84.0	80.5/84.2/82.3	84.8/88.1/86.4	83.2/86.6/**84.9**	26
		40	83.9/87.1/85.5	81.2/85.7/83.4	85.6/88.9/87.2	84.3/87.7/**86.0**	52
		60	84.7/87.5/86.1	82.8/87.3/85.0	86.8/89.9/88.3	85.4/88.6/**87.0**	78
	3.5	10	79.4/81.7/80.5	73.6/79.8/76.6	81.2/85.6/83.3	79.3/83.3/81.2	2.6
		20	81.9/83.7/82.8	77.3/81.2/79.2	84.3/87.3/85.8	82.3/85.0/**83.6**	5.2
CodeLlama	13B	10	82.8/82.4/82.6	81.6/80.2/80.9	86.0/87.4/86.7	84.1/84.4/**84.3**	-
		20	84.1/84.4/84.2	83.6/82.7/83.1	86.6/88.4/87.5	85.2/86.0/**85.6**	-
		40	85.9/85.8/85.9	83.4/83.4/83.4	87.5/88.4/87.9	86.2/86.6/**86.4**	-
		60	85.0/86.7/85.8	83.4/84.6/84.0	87.4/89.2/88.3	85.9/87.5/**87.2**	-
	7B	10	83.9/77.3/80.5	83.0/74.7/78.6	87.2/84.4/85.8	85.3/80.3/**82.7**	-
		20	86.5/81.1/83.7	83.3/76.0/79.5	87.5/86.6/87.0	86.4/82.9/**84.6**	-
		40	86.7/82.8/84.7	83.1/77.1/80.0	88.2/87.7/87.9	86.8/84.2/**85.4**	-
		60	86.4/83.6/85.0	84.4/80.4/82.4	88.2/88.8/88.5	86.9/85.6/**86.2**	-
deepseek-coder	6.7B	10	82.4/80.9/81.6	81.8/79.7/80.8	85.3/86.2/85.8	83.7/83.2/**83.5**	-
		20	84.4/83.5/83.9	83.1/80.2/81.6	86.4/87.9/87.2	85.1/85.1/**85.1**	-
		40	85.5/85.4/85.5	83.7/81.9/82.8	87.4/88.1/87.8	86.1/86.1/**86.2**	-
		60	85.9/85.4/85.7	85.6/84.6/85.1	87.5/89.1/88.3	86.6/87.0/**86.9**	-
Llama3	8B	10	84.3/76.1/80.0	83.8/75.6/79.5	87.0/85.3/86.2	85.5/80.4/**82.9**	-
		20	85.3/81.4/83.3	82.8/78.5/80.6	87.7/87.7/87.7	86.0/83.9/**85.0**	-
		40	85.9/82.2/84.0	83.9/80.6/82.2	87.8/88.3/88.0	86.5/84.9/**85.6**	-
		60	86.1/82.8/84.4	83.3/83.0/83.1	87.7/85.6/86.7	86.4/84.2/**85.3**	-
Llama2	13B	10	82.3/76.4/79.3	81.1/73.4/77.1	85.9/82.4/84.1	83.8/78.8/81.2	-
		20	83.9/80.0/81.9	80.4/74.2/77.2	84.4/81.3/82.8	83.6/79.6/81.5	-
	7B	10	78.2/77.7/78.0	75.2/73.5/74.3	82.6/83.2/82.9	79.8/79.6/79.7	-
		20	80.5/79.2/79.8	77.7/75.5/76.6	84.0/80.6/82.3	81.7/79.3/80.5	-
StarCoder2	15B	10	85.4/82.4/83.9	83.4/79.6/81.4	86.6/86.4/86.5	85.6/83.9/**84.7**	-
		20	86.6/85.4/86.0	84.2/81.8/83.0	87.5/88.4/88.0	86.6/86.2/**86.5**	-
		40	87.4/86.7/87.1	84.6/84.1/84.3	87.6/88.7/88.2	87.0/87.2/**87.2**	-
		60	87.9/87.2/87.6	85.8/85.4/85.6	88.6/89.4/89.0	87.9/88.0/**87.9**	-
	7B	10	83.6/81.0/82.3	82.7/79.4/81.0	86.6/86.6/86.6	84.9/83.4/**84.2**	-
		20	85.5/84.1/84.8	83.2/82.6/82.9	87.1/88.4/87.7	85.9/85.9/**85.9**	-
		40	86.2/85.8/86.0	83.7/84.1/83.9	87.4/89.0/88.2	86.4/87.1/**86.7**	-
		60	86.8/86.7/86.7	84.8/85.4/85.1	87.7/89.1/88.4	86.9/87.6/**87.2**	-
StarCoder	16B	10	81.9/81.7/81.8	79.7/77.1/78.4	86.3/86.1/86.2	83.6/83.0/**83.3**	-
		20	83.2/84.4/83.8	81.3/80.7/81.0	86.8/88.1/87.5	84.6/85.6/**85.1**	-
		40	84.8/86.1/85.4	82.6/83.1/82.8	87.6/89.0/88.3	85.8/87.0/**86.4**	-
		60	85.0/86.7/85.8	83.4/84.6/84.0	87.4/89.2/88.3	85.9/87.5/**86.7**	-
CodeGen2	16B	10	79.7/81.4/80.6	77.6/79.3/78.5	85.0/86.3/85.7	81.9/83.4/**82.7**	-
	7B	10	77.7/77.3/77.5	76.6/77.2/76.9	82.2/84.0/83.1	79.7/80.5/80.1	-

TABLE C: Comparison of different LLMs with SR on CallMeMaybe-data: Accuracy (%), and Cost ($).

Approach / Model (+SR)		K	Accuracy	Cost ($)
CallMeMaybe		-	70.0	-
GPT	4	10	67.4	2.5
		20	66.3	5.1
		40	65.2	10.1
		60	**70.8**	15.15
	3.5	10	**70.8**	0.25
		20	**71.9**	0.51
		40	**73.0**	1.01
Codellama	13B	10	**71.9**	-
		20	**70.8**	-
		40	**75.3**	-
		60	**76.4**	-
	7B	10	**73.0**	-
		20	**75.3**	-
		40	**74.2**	-
		60	**75.3**	-
deepseek-coder	6.7B	10	67.4	-
		20	66.3	-
		40	66.3	-
		60	66.3	-
Llama3	8B	10	69.7	-
		20	67.4	-
		40	**70.8**	-
		60	**71.9**	-
Llama2	13B	10	69.7	-
		20	**74.2**	-
		40	**73.0**	-
	7B	10	66.3	-
		20	68.5	-
		40	69.7	-
StarCoder2	15B	10	**75.3**	-
		20	**76.4**	-
		40	**76.4**	-
		60	**75.3**	-
	7B	10	**74.2**	-
		20	**75.3**	-
		40	**76.4**	-
		60	**76.4**	-
StarCoder	16B	10	62.9	-
		20	65.2	-
		40	**74.2**	-
		60	**73.0**	-
CodeGen2	16B	10	67.4	-
		20	68.5	-
	7B	10	68.5	-
		20	67.4	-