Table 1: Performance of the service LLMs and local LLMs fine-tuned on 128K synthetic dataset produced by GPT4o, evaluated by GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash as judges on test subsets of coverage dataset. Each entry is presented as mean score / coverage percentage (%) with 50 score / coverage percentage (%) with 70 score. The best results from service and local LLMs are highlighted in **bold**. "**P-Match**" represents performance matching, which is defined as the best performance of the local LLM divided by the best performance of the service LLM, with the best results highlighted in **bold** across different judges.

Task	**Model**	**GPT4o**		**Claude 3 Sonnet**		**Gemini 1.5 Flash**	
		Precision↑	**Similarity↑**	**Precision↑**	**Similarity↑**	**Precision↑**	**Similarity↑**
Summarization	GPT4o	**90.71** / **97 %** / **96%**	**82.00** / **95%** / **89%**	93.25 / **100%** / **100%**	**86.60** / **100%** / **95%**	**87.10** / **100%** / 92%	**67.45** / 85% / **48%**
	Claude 3 Sonnet	88.04 / **97%** / 92%	78.18 / **95%** / 78%	**93.39** / **100%** / 99%	85.55 / **100%** / **95%**	86.70 / **100%** / 92%	64.10 / 80% / 36%
	Gemini 1.5 Flash	87.90 / 96% / **96%**	79.14 / **95%** / 88%	91.95 / **100%** / 98%	85.05 / **100%** / **95%**	85.65 / 98% / **96%**	66.45 / **89%** / 40%
	Gemma 2B	57.60 / 64% / 35%	54.49 / 61% / 35%	74.89 / 86% / 69%	64.09 / 73% / 50%	61.90 / 78% / 40%	42.15 / 38% / 12%
	Gemma 7B	73.54 / 85% / 65%	68.58 / 85% / 59%	86.19 / **99%** / 93%	77.41 / 94% / 77%	**74.59** / **95%** / **69%**	**53.92** / **65%** / 22%
	Mistral 7B	**76.38** / **93%** / 70%	69.65 / **88%** / 56%	86.20 / **99%** / 92%	**78.44** / **96%** / 80%	72.74 / **95%** / 62%	50.15 / 54% / 14%
	LLaMA3 8B	75.67 / 88% / **75%**	**70.54** / 86% / **69%**	**87.02** / **99%** / **94%**	78.42 / 93% / **86%**	72.74 / 91% / 64%	52.23 / 64% / **25%**
	P-Match↑	84.20% / 95.88% / 78.13%	86.02% / 92.63% / 77.53%	**93.18%** / **99%** / **94%**	**90.58%** / **96%** / **90.53%**	85.64% / 95% / 71.88%	79.94% / 73.03% / 52.08%
Classification	GPT4o	83.62 / **94%** / 81%	**74.45** / 80% / 66%	87.50 / 92% / 92%	72.28 / 72% / 66%	82.68 / 94% / 80%	63.06 / 67% / 44%
	Claude 3 Sonnet	82.32 / 92% / 78%	71.53 / **81%** / 70%	**92.89** / **100%** / **100%**	75.07 / **81%** / 73%	**87.34** / **97%** / **97%**	**67.18** / **80%** / 45%
	Gemini 1.5 Flash	**85.43** / **94%** / **91%**	72.73 / **81%** / **75%**	89.03 / 94% / 89%	**77.96** / **81%** / **81%**	83.35 / 94% / 84%	64.25 / 78% / **47%**
	Gemma 2B	58.47 / 58% / 42%	52.76 / 50% / 39%	69.98 / 73% / 62%	56.31 / 58% / 47%	62.17 / 62% / 48%	48.54 / 50% / 39%
	Gemma 7B	70.73 / 69% / 55%	64.67 / 62% / 53%	78.78 / 81% / 75%	67.76 / 69% / 62%	70.73 / 75% / 61%	59.77 / 59% / 52%
	Mistral 7B	67.53 / 70% / 53%	61.65 / 67% / 47%	76.01 / 80% / 72%	64.43 / 70% / 52%	67.90 / 73% / 53%	54.27 / 53% / 45%
	LLaMA3 8B	**81.64** / **88%** / **73%**	**78.02** / **77%** / **67%**	**89.20** / **94%** / **94%**	**82.18** / **88%** / **75%**	**83.63** / **94%** / **77%**	**72.54** / **73%** / **64%**
	P-Match↑	95.56% / 93.62% / 80.22%	104.80% / 95.06% / 89.33%	**96.03%** / 94% / **94%**	105.41% / **108.64%** / 92.59%	95.75% / **96.91%** / 79.38%	**107.98%** / 91.25% / **136.17%**
Coding	GPT4o	**90.31** / **100%** / **98%**	75.18 / 92% / 70%	**94.57** / **100%** / **100%**	86.32 / **100%** / 91%	**90.78** / **100%** / **100%**	58.43 / 62% / 25%
	Claude 3 Sonnet	88.76 / **100%** / 92%	75.23 / **94%** / 67%	93.82 / **100%** / **100%**	**87.42** / **100%** / **100%**	89.84 / **100%** / **100%**	60.46 / 69% / 31%
	Gemini 1.5 Flash	88.51 / 98% / 94%	**75.62** / 91% / **73%**	93.59 / **100%** / **100%**	82.92 / 97% / 84%	90.62 / **100%** / 98%	**64.21** / **84%** / **41%**
	Gemma 2B	62.31 / 70% / 44%	56.48 / 66% / 41%	80.92 / 89% / 84%	67.24 / 78% / 48%	72.98 / 89% / 66%	44.08 / 50% / 8%
	Gemma 7B	**80.56** / **92%** / **80%**	**71.92** / **89%** / **70%**	**90.47** / **100%** / **98%**	**80.26** / **92%** / **84%**	**84.66** / **100%** / **88%**	**61.23** / **72%** / **36%**
	Mistral 7B	68.32 / 77% / 56%	61.01 / 69% / 45%	81.25 / 92% / 81%	69.10 / 83% / 55%	72.39 / 86% / 69%	45.25 / 50% / 8%
	LLaMA3 8B	77.47 / 88% / 72%	69.46 / 88% / 61%	83.97 / 94% / 83%	73.51 / 88% / 67%	75.55 / 89% / 73%	51.10 / 58% / 17%
	P-Match↑	89.20% / 92% / 81.63%	95.11% / **94.68%** / 95.89%	**95.66%** / **100%** / **98%**	91.81% / 92% / 84%	93.26% / **100%** / 88%	**95.36%** / 85.71% / **97.80%**
Closed QA	GPT4o	**95.45** / **100%** / **100%**	84.23 / 93% / 80%	97.21 / **100%** / **100%**	92.56 / **100%** / 97%	93.58 / **100%** / **100%**	75.58 / 85% / 63%
	Claude 3 Sonnet	94.03 / **100%** / 98%	85.28 / **100%** / 82%	97.60 / **100%** / **100%**	93.95 / **100%** / **100%**	93.66 / **100%** / **100%**	76.33 / 92% / 65%
	Gemini 1.5 Flash	94.63 / **100%** / 97%	**87.43** / 95% / **87%**	**98.25** / **100%** / **100%**	**97.41** / **100%** / **100%**	**95.00** / **100%** / **100%**	**85.66** / **97%** / **83%**
	Gemma 2B	67.25 / 65% / 53%	67.41 / 67% / 57%	80.22 / 85% / 78%	70.20 / 73% / 72%	70.33 / 73% / 60%	59.68 / 62% / 50%
	Gemma 7B	81.85 / **88%** / **83%**	79.02 / **85%** / 78%	**88.83** / **93%** / **93%**	83.95 / 87% / **83%**	**82.51** / **93%** / **80%**	72.24 / 75% / 67%
	Mistral 7B	**83.63** / 87% / 82%	**81.36** / **85%** / **83%**	88.25 / **93%** / 85%	**84.77** / **88%** / **83%**	82.04 / 85% / 78%	**73.95** / **78%** / **70%**
	LLaMA3 8B	75.55 / 78% / 75%	72.62 / 77% / 72%	86.03 / 88% / 85%	77.64 / 80% / 80%	79.09 / 85% / 77%	68.78 / 75% / 65%
	P-Match↑	87.62% / 88% / 83%	**93.06%** / 85% / **95.40%**	**90.41%** / **93%** / **93%**	87.02% / **88%** / 83%	86.85% / **93%** / 80%	86.33% / 80.41% / 84.34%

cluding summarization, classification, coding, and closed QA. We utilize GPT4o, Claude 3 Sonnet, and Gemini 1.5 Flash as judges to evaluate the fine-tuned model performance on test subsets of the coverage dataset. As demonstrated in Table 1, the fine-tuned local LLMs, despite their significantly smaller scale, achieve comparable performance on diverse tasks compared to much larger service LLMs. For instance, in the summarization task, LLaMA3 8B achieved a comparable precision score of 87.02 / 99% / 94%, compared to GPT4o's score of 93.25 / 100% / 100%, Claude 3 Sonnet's score of 93.39 / 100% / 99%, and Gemini 1.5 Flash's score of 91.95 / 100% / 98%, with Claude 3 Sonnet serving as judge. These results underscore the efficacy of LlamaDuo in seamlessly transferring knowledge and capabilities from service LLMs to smaller local LLMs without a substantial decrease in performance.

In Table 1, we observe distinct performance across four fine-tuned models when applied to different tasks. Specifically, Mistral 7B stands out in summarization tasks, achieving the best performance in 7 out of 12 cases. Moreover, LLaMA3 8B consistently outperforms competitors across all metrics and evaluators in the classification task. Conversely, in coding tasks, Gemma 7B is identified as the leading model, excelling across all metrics and evaluations. Mistral 7B shows superior performance in the closed QA task, leading in 8 out of 12 cases. Within the realm of service LLMs, Claude 3 Sonnet and Gemini 1.5 Flash demonstrate exceptional performance in classification and closed QA tasks, securing the best results in 8 and 10 out of 12 cases, respectively. Lastly, GPT4o emerges as the leading model in summarization and coding tasks, achieving the best performance in 10 and 7 out of 12 cases, respectively. Notably, although Gemma 2B exhibits inferior performance compared to larger 7B models overall, the disparity in results is not markedly substantial, with Gemma 2B attaining closely comparable performance in certain tasks. For example, in closed QA tasks, Gemma 2B secures a mean precision score of 80.22, while Gemma 7B achieves 88.83, Mistral 7B reaches 88.25, and LLaMA3 8B obtains 86.03, as evaluated by Claude 3 Sonnet. This observation lends further support to the notion that through the strategic fine-tuning of smaller local LLMs on synthetic datasets via the LlamaDuo, it is possible to closely approximate the performance of their larger counterparts. Consequently, it offers increased flexibility and solutions for users and scenarios with budgetary considerations. More experimental results are presented in Appendix C.