Table A1: Average performance when the query budget is increased to 500 iterations. The performance increased as compared to Table 2 of the original paper. The gaps between EASE and baselines are still significant.

Tasks (10% noise)	Evo	Best-of-N	EASE
LR	70.0±0.0	70.0±0.0	**78.3±1.7**
LP-variant	73.3±1.7	66.7±1.7	**78.3±1.7**
AG News Remap	40.0±0.0	50.0±0.0	**55.0±0.0**
SST5 Reverse	40.0±0.0	41.7±1.7	**51.7±1.7**

Table A2: Average performance for tasks involving reasoning chains. The advantages of EASE are significant.

Tasks	Evo	Best-of-N	EASE
MATH (Hendrycks et al., 2021)	61.7±1.7	60.0±0.0	**66.7±4.4**
GSM8K (Cobbe et al., 2021)	70.0±2.9	68.3±1.7	**75.0±2.9**
AQuA-RAT (Ling et al., 2017)	46.7±1.7	43.3±1.7	**50.0±2.9**

Table A3: Extension of Appendix D.3 ablation for NeuralUCB and OT on real tasks. Evaluations are done on the below three open-ended real tasks.

Tasks	OT	NeuralUCB	EASE
auto_categorization	5.0±0.0	**30.0±0.0**	**30.0±0.0**
taxonomy_animal	58.3±1.7	86.7±1.7	**88.3±2.7**
word_sorting	85.0±0.0	90.0±0.0	**91.7±1.4**

Table A4: Average performance on benchmarks used in the TEMPERA paper.

Tasks	Evo	Best-of-N	EASE
MR	95.0±0.0	95.0±0.0	**96.7±1.7**
Yelp P.	**95.0±0.0**	**95.0±0.0**	**95.0±0.0**
CR	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**
MNLI	**81.7±1.7**	80.0±0.0	**81.7±1.7**
MRPC	**81.7±1.7**	**81.7±1.7**	**81.7±1.7**

Table A5: Average performance on Meta's newest open-weight model Llama-3.1-8B-Instruct.

Tasks	Evo	Best-of-N	EASE
LR	11.7±1.7	15.0±2.9	**30.0±2.9**
LP-variant	11.7±1.7	15.0±0.0	**23.3±1.7**
AG News Remap	31.7±1.7	31.7±1.7	**33.3±3.3**
SST5 Reverse	31.7±3.3	30.0±0.0	**33.3±3.3**

Table A6: Results for the setting that allows different number exemplars to be selected (i.e., allowing a range instead of a fixed k. EASE outperforms Best-of-N. We also observe high-performing prompts usually consist of more exemplars (i.e., larger average k).

Tasks	n	Max k	Best-of-N		EASE	
			Acc	Avg. best k	Acc	Avg. best k
LR	100	5	60.0±2.9	5.0±0.0	**76.7±1.7**	4.7±0.3
LP-variant	100	5	60.0±0.0	2.7±0.7	**75.0±2.9**	5.0±0.0
AG News Remap	100	5	31.7±1.7	3.0±0.0	**48.3±1.7**	4.7±0.3
SST5 Reverse	100	5	31.7±1.7	4.3±0.7	**43.3±1.7**	5.0±0.0
AG News Remap	1000	50	45.0±2.9	25.0±0.0	**51.7±1.7**	49.3±0.7
SST5 Reverse	1000	50	60.0±0.0	38.3±8.8	**61.7±1.7**	41.3±2.7

Table A7: The average performance of random exemplar in-context prompting in "out-of-distribution" tasks. It only achieves 17.6% accuracy on average, indicating that the model has little knowledge about the task. The performance gain from EASE is relatively large.

Task	Noise	Random	EASE	Gap
LR	0%	34.4±0.0	81.7±3.6	47.3
	10%	28.0±0.1	73.3±3.6	45.3
	30%	18.2±0.1	78.3±1.4	60.1
	50%	10.8±0.1	71.7±2.7	60.9
	70%	4.2±0.1	66.7±3.6	62.5
	90%	0.5±0.0	53.3±2.7	52.8
LP-variant	0%	46.3±0.1	75.0±0.0	28.7
	10%	44.1±0.2	75.0±2.4	30.9
	30%	36.0±0.2	73.3±1.4	37.3
	50%	27.5±0.1	76.7±2.7	49.2
	70%	16.8±0.1	75.0±0.0	58.2
	90%	3.6±0.1	63.3±1.4	59.7
AG News Remap	0%	9.2±0.1	53.3±3.6	44.1
	10%	8.6±0.1	56.7±2.7	48.1
	30%	8.1±0.1	51.7±1.4	43.6
	50%	7.5±0.0	56.7±1.4	49.2
	70%	7.0±0.1	51.7±1.4	44.7
	90%	6.8±0.1	55.0±2.4	48.2
SST5 Reverse	0%	18.8±0.0	50.0±0.0	31.2
	10%	18.2±0.1	50.0±0.0	31.8
	30%	17.5±0.1	41.7±3.6	24.2
	50%	16.9±0.1	43.3±1.4	26.4
	70%	16.8±0.1	45.0±2.4	28.2
	90%	17.3±0.0	31.7±1.4	14.4
Mean		17.6	60.4	42.8

Table A8: The average performance of random exemplar in-context prompting in Instruction Induction (II) tasks. It achieves 64.7% accuracy on average, indicating that the model has much knowledge about the task. The performance gain from EASE is relatively small.

Task	Random	EASE	Gap
active_to_passive	100.0±0.0	100.0±0.0	0.0
antonyms	79.3±0.0	90.0±0.0	10.7
auto_categorization	5.4±0.1	30.0±0.0	24.6
diff	6.5±0.0	100.0±0.0	93.5
first_word_letter	100.0±0.0	100.0±0.0	0.0
larger_animal	83.3±0.1	100.0±0.0	16.7
letters_list	100.0±0.0	100.0±0.0	0.0
negation	94.7±0.0	95.0±0.0	0.3
num_to_verbal	100.0±0.0	100.0±0.0	0.0
object_counting	52.9±0.1	73.3±1.4	20.4
orthography_starts_with	50.7±0.1	78.3±1.4	27.6
rhymes	56.7±0.2	100.0±0.0	43.3
second_word_letter	23.8±0.1	50.0±0.0	26.2
sentence_similarity	20.5±0.2	56.7±1.4	36.2
sentiment	89.8±0.2	100.0±0.0	10.2
singular_to_plural	100.0±0.0	100.0±0.0	0.0
sum	12.5±0.0	100.0±0.0	87.5
synonyms	21.3±0.1	30.0±0.0	8.7
taxonomy_animal	51.5±0.2	88.3±2.7	36.8
translation_en-de	84.4±0.0	90.0±0.0	5.6
translation_en-es	94.1±0.1	100.0±0.0	5.9
translation_en-fr	78.7±0.0	88.3±1.4	9.6
word_sorting	84.5±0.1	91.7±1.4	7.2
word_unscrambling	61.5±0.1	78.3±2.7	16.8
Mean	64.7	85.0	20.3