Appendix

Table 1: Standard prompting number of successful functions per task category out of 50 by number of choices with gpt-3.5-turbo. Categories Padded Fill, Move 2 Towards, Flip, Mirror, Denoise, Denoise Multicolor, Pattern Copy, Pattern Copy Multicolor, Recolor by Odd Even, Recolor by Size, Recolor by Size Comparison, Scaling were omitted as none of the tasks were successful.

Choice	1	2	3	4	5	6	7	8	9	10
Task category										
Move 1	1	1	1	2	2	2	2	2	2	3
Move 2	0	0	0	0	0	0	0	1	2	2
Move 3	0	1	2	2	3	3	3	4	4	4
Move Dynamic	0	0	0	0	0	1	1	1	2	2
Fill	0	0	0	1	2	2	2	2	2	2
Hollow	0	0	0	0	0	0	0	0	1	1
Mean success rate (%)	**0.11**	**0.22**	**0.33**	**0.56**	**0.78**	**0.89**	**0.89**	**1.11**	**1.44**	**1.56**
Standard deviation	**0.471**	**0.647**	**1.029**	**1.338**	**1.833**	**1.844**	**1.844**	**2.193**	**2.357**	**2.526**

Table 2: Direct feedback number of successful functions per task category out of 50 by number of choices with gpt-3.5-turbo. Categories Move 2, Move 2 Towards, Move Dynamic, Fill, Padded Fill, Flip, Mirror, Denoise, Denoise Multicolor, Pattern Copy, Pattern Copy Multicolor, Recolor by Odd Even, Recolor by Size, Recolor by Size Comparison, Scaling were omitted since none of those tasks were successful.

Choice	1	2	3	4	5
Task category					
Move 1	0	1	2	2	2
Move 3	2	2	2	2	3
Hollow	1	1	2	2	2
Mean success rate (%)	**0.33**	**0.44**	**0.67**	**0.67**	**0.78**
Standard deviation	**1.029**	**1.097**	**1.534**	**1.534**	**1.833**

Table 3: CoT number of successful functions per task category out of 50 by number of choices with gpt-3.5-turbo. Categories Pattern Copy, Mirror, Pattern Copy Multicolor, Recolor by Odd Even, Recolor by Size, Recolor by Size Comparison, Scaling were omitted since none of those tasks were successful.

Choice	1	2	3	4	5
Task category					
Move 1	3	5	6	6	6
Move 2	0	1	1	1	1
Move 3	0	1	1	1	1
Move Dynamic	1	1	1	1	1
Move 2 Towards	0	0	0	0	1
Fill	1	2	2	2	2
Padded Fill	0	0	0	0	0
Hollow	1	2	3	5	7
Flip	0	1	1	1	1
Denoise	0	1	1	1	1
Denoise Multicolor	0	0	1	1	1
Mean success rate (%)	**0.67**	**1.56**	**1.89**	**2.11**	**2.44**
Standard deviation	**1.534**	**2.526**	**3.027**	**3.462**	**4.033**

Table 4: Standard prompting used tokens, success rate and costs per number of choices with gpt-3.5-turbo

Number of choices	1	2	3	4	5
Input tokens	**129607**	**129607**	**129607**	**129607**	**129607**
Output tokens	**45289**	**90578**	**135867**	**181156**	**226446**
Mean success rate (%)	**0.111**	**0.222**	**0.333**	**0.556**	**0.778**
Cost in $	**0.133**	**0.201**	**0.269**	**0.337**	**0.404**
Benefit-cost ratio	**0.837**	**1.107**	**1.241**	**1.651**	**1.923**
Number of choices	6	7	8	9	10
Input tokens	**129607**	**129607**	**129607**	**129607**	**129607**
Output tokens	**271735**	**317024**	**362313**	**407602**	**452891**
Mean success rate (%)	**0.889**	**0.889**	**1.111**	**1.444**	**1.556**
Cost in $	**0.472**	**0.54**	**0.608**	**0.676**	**0.744**
Benefit-cost ratio	**1.882**	**1.645**	**1.827**	**2.136**	**2.09**

Table 5: CoT used tokens, success rate and costs per number of choices with gpt-3.5-turbo

Number of choices	1	2	3	4	5
Input tokens	**143122**	**143122**	**143122**	**143122**	**143122**
Output tokens	**165003**	**330007**	**495010**	**660014**	**825017**
Mean success rate (%)	**0.667**	**1.556**	**1.889**	**2.111**	**2.444**
Cost in $	**0.319**	**0.567**	**0.814**	**1.062**	**1.309**
Benefit-cost ratio	**2.089**	**2.746**	**2.32**	**1.989**	**1.867**

Table 6: Direct feedback used tokens, success rate and costs per number of choices with gpt-3.5-turbo

Number of choices	1	2	3	4	5
Input tokens	**129964**	**308737**	**528124**	**789198**	**1092075**
Output tokens	**45080**	**88221**	**132960**	**177407**	**223007**
Mean success rate (%)	**0.333**	**0.444**	**0.667**	**0.667**	**0.778**
Cost in $	**0.133**	**0.287**	**0.464**	**0.661**	**0.881**
Benefit-cost ratio	**2.514**	**1.55**	**1.438**	**1.009**	**0.883**

Table 7: Standard prompting number of successful functions per task category out of 50 by number of choices with GPT-4. Categories Pattern Copy, Pattern Copy Multicolor, Mirror, Recolor by Odd Even, Recolor by Size, Recolor by Size Comparison were omitted since non of those tasks were successful.

Choice	1	2	3	4	5	6	7	8	9	10
Task category										
Move 1	7	11	13	14	17	19	20	21	21	25
Move 2	2	4	4	5	5	5	5	7	8	9
Move 3	2	8	12	13	13	14	16	18	18	18
Move Dynamic	2	2	2	3	4	4	4	6	6	6
Move 2 Towards	0	0	0	0	0	1	1	2	2	3
Fill	5	13	19	22	22	24	26	27	28	30
Padded Fill	0	0	0	0	0	1	1	1	1	1
Hollow	6	10	14	18	22	26	29	30	30	32
Flip	5	7	9	13	19	21	22	24	24	27
Denoise	0	4	4	5	5	5	6	6	6	7
Denoise Multicolor	1	1	1	4	5	6	8	10	12	13
Scaling	1	2	4	6	6	7	7	7	7	7
Mean success rate (%)	**3.44**	**6.89**	**9.11**	**11.44**	**13.11**	**14.78**	**16.11**	**17.67**	**18.11**	**19.78**
Standard deviation	**4.743**	**8.818**	**12.179**	**14.255**	**16.381**	**18.281**	**19.888**	**20.81**	**21.049**	**22.936**

Table 8: CoT number of successful functions per task category out of 50 by number of choices with gpt-4. Categories Pattern Copy, Pattern Copy Multicolor, Mirror were omitted since non of those tasks were successful.

Choice	1	2	3	4	5
Task category					
Move 1	7	10	12	14	18
Move 2	5	15	17	22	27
Move 3	6	8	10	15	19
Move Dynamic	2	2	2	4	6
Move 2 Towards	1	2	3	3	4
Fill	5	11	14	17	23
Padded Fill	1	1	1	1	1
Hollow	16	25	28	37	39
Flip	9	11	20	21	24
Denoise	5	8	12	13	16
Denoise Multicolor	15	21	24	31	34
Recolor by Size	1	2	2	2	2
Recolor by Size Comparison	0	0	1	1	1
Scaling	4	4	7	10	12
Mean success rate (%)	**8.56**	**13.33**	**17.0**	**21.22**	**25.11**
Standard deviation	**9.889**	**15.262**	**18.153**	**22.949**	**25.743**

Table 9: Direct feedback number of successful functions per task category out of 50 by number of choices with GPT-4. Categories Mirror, Pattern Copy, Pattern Copy Multicolor, Padded Fill, Recolor by Odd Even, Recolor by Size, Recolor by Size Comparison were omitted since non of those tasks were successful.

Choice	1	2	3	4	5
Task category					
Move 1	5	10	13	14	17
Move 2	1	2	4	4	4
Move 3	4	5	5	6	6
Move Dynamic	1	1	1	3	3
Move 2 Towards	1	1	2	3	3
Fill	9	13	15	15	16
Hollow	10	15	20	23	23
Flip	5	8	9	11	11
Denoise	1	1	2	4	5
Denoise Multicolor	3	4	5	6	6
Scaling	1	4	4	4	7
Mean success rate (%)	**4.56**	**7.11**	**8.89**	**10.33**	**11.22**
Standard deviation	**6.28**	**9.634**	**11.985**	**13.092**	**13.791**

Table 10: Standard prompting used tokens, success rate and costs per number of choices with gpt-4

Number of choices	1	2	3	4	5
Input tokens	**129607**	**129607**	**129607**	**129607**	**129607**
Output tokens	**50767**	**101533**	**152300**	**203067**	**253834**
Mean success rate (%)	**3.444**	**6.889**	**9.111**	**11.444**	**13.111**
Cost in $	**6.934**	**9.98**	**13.026**	**16.072**	**19.118**
Benefit-cost ratio	**0.497**	**0.69**	**0.699**	**0.712**	**0.686**
Number of choices	6	7	8	9	10
Input tokens	**129607**	**129607**	**129607**	**129607**	**129607**
Output tokens	**304600**	**355367**	**406134**	**456900**	**507667**
Mean success rate (%)	**14.778**	**16.111**	**17.667**	**18.111**	**19.778**
Cost in $	**22.164**	**25.21**	**28.256**	**31.302**	**34.348**
Benefit-cost ratio	**0.667**	**0.639**	**0.625**	**0.579**	**0.576**

Table 11: CoT used tokens, success rate and costs per number of choices with gpt-4

Number of choices	1	2	3	4	5
Input tokens	**143122**	**143122**	**143122**	**143122**	**143122**
Output tokens	**276249**	**552498**	**828748**	**1104997**	**1381246**
Mean success rate (%)	**8.556**	**13.333**	**17.0**	**21.222**	**25.111**
Cost in $	**20.869**	**37.444**	**54.019**	**70.593**	**87.168**
Benefit-cost ratio	**0.41**	**0.356**	**0.315**	**0.301**	**0.288**

Table 12: Direct feedback used tokens, success rate and costs per number of choices with gpt-4

Number of choices	1	2	3	4	5
Input tokens	**129607**	**309496**	**541517**	**828076**	**1168956**
Output tokens	**10624**	**71260**	**180380**	**335254**	**542357**
Mean success rate (%)	**4.556**	**7.111**	**8.889**	**10.333**	**11.222**
Cost in $	**4.526**	**13.56**	**27.068**	**44.957**	**67.61**
Benefit-cost ratio	**1.007**	**0.524**	**0.328**	**0.23**	**0.166**