

Supplementary Information

Supplementary Methods

The evaluation used 100 aligned QuALITY questions. Runs used gpt-4o-mini with temperature 0.0 and seed 0. The Du implementation used three agents, fixed schedules of one and two rounds, and an adaptive maximum of six rounds with a judge-confidence threshold of 90. The Liang implementation used an affirmative side, a negative side, and a moderator; fixed schedules of one, two, and three rounds; and a strict adaptive maximum of six rounds with an early-stop threshold of 90. For Du adaptive stopping, the judge controlled stopping while the final scored answer remained the agent majority vote. Liang strict early stopping required a moderator preference, a non-empty debate answer, and confidence at or above the threshold.

Supplementary Results

Table S1 Final primary condition results

Condition	Correct	Accuracy	Total rounds	Average rounds	Schedule or stopping distribution
Single-agent baseline	77	0.77	100	1.0	Fixed 1 response
Du fixed 1 ensemble	80	0.8	100	1.0	Fixed 1
Du fixed 2	79	0.79	200	2.0	Fixed 2
Du adaptive (T90)	77	0.77	109	1.09	R1: 95; R2: 4; R6: 1
Liang fixed 2	73	0.73	200	2.0	Fixed 2
Liang fixed 2 prompt-matched	52	0.52	200	2.0	Fixed 2
Liang strict adaptive (T90)	53	0.53	270	2.7	R1: 21; R2: 52; R3: 4; R4: 1; R5: 3; R6: 19

Table S2 Du threshold sensitivity and Liang fixed round bracketing results

Condition	Correct	Accuracy	Total rounds	Average rounds	Schedule or stopping distribution
Du adaptive T70	79	0.79	100	1.0	R1: 100
Du adaptive T80	77	0.77	102	1.02	R1: 98; R2: 2
Du adaptive T90 revision rerun	78	0.78	104	1.04	R1: 97; R2: 2; R3: 1
Du adaptive T95	80	0.8	110	1.1	R1: 94; R2: 5; R6: 1
Liang fixed 1 original prompt	61	0.61	100	1.0	Fixed 1
Liang fixed 3 original prompt	68	0.68	300	3.0	Fixed 3

Table S3 Exact two sided McNemar comparisons

Comparison	A only	B only	Exact two sided McNemar p
Du fixed 1 vs. Du adaptive	6	3	0.5078125
Du fixed 1 vs. Du fixed 2	6	5	1
Du fixed 2 vs. Du adaptive	8	6	0.79052734375
Liang fixed 2 vs. Liang adaptive	30	10	0.00222143377323
Single agent vs. Du fixed 1	2	5	0.453125
Single agent vs. Du fixed 2	5	7	0.7744140625
Single agent vs. Du adaptive	4	4	1
Single agent vs. Liang fixed 2	11	7	0.480682373047
Single agent vs. Liang adaptive	35	11	0.0005355850721
Liang fixed 2 vs. prompt-matched fixed 2	29	8	0.000752897060011

Comparison	A only	B only	Exact two sided McNemar p
Liang prompt-matched fixed 2 vs. adaptive	10	11	1
Single agent vs. Liang prompt-matched fixed 2	36	11	0.000346001140315
Liang fixed 1 vs. fixed 2 original prompt	8	20	0.0356981381774
Liang fixed 2 vs. fixed 3 original prompt	13	8	0.383310317993

Table S4 Stopping round counts and accuracy

Family	Rounds used	Questions	Correct
Du	1	95	75
Du	2	4	1
Du	6	1	1
Liang	1	21	13
Liang	2	52	25
Liang	3	4	0
Liang	4	1	0
Liang	5	3	1
Liang	6	19	14

Table S5 Initial stopping answer to final scored answer

Family	Rounds used	Correct to correct	Correct to incorrect	Incorrect to correct	Incorrect to incorrect
Du	1	74	3	1	17
Du	2	0	1	1	2
Du	6	0	0	1	0
Du	All	74	4	3	19
Liang	1	13	0	0	8
Liang	2	24	0	1	27
Liang	3	0	0	0	4
Liang	4	0	1	0	0
Liang	5	1	1	0	1
Liang	6	5	3	9	2
Liang	All	43	5	10	42

Note. For Du, the initial stopping answer is the first judge answer and the final scored answer is the agent majority vote. For Liang, the initial stopping answer is the first moderator debate answer.

Table S6 Final recorded confidence values

Family	Final confidence	Questions	Correct
Du	70	1	1
Du	90	2	0
Du	95	71	54
Du	100	26	22
Liang	50	7	6
Liang	70	2	2

Family	Final confidence	Questions	Correct
Liang	75	3	2
Liang	80	6	4
Liang	90	82	39

Reproducibility Notes

All 13 reported condition files contained 100 aligned records and passed the supplied checks for question alignment, gold-answer alignment, accuracy totals, round totals, prediction labels, and stopping distributions. ALL 13 REPORTED CONDITION FILES PASSED.

Selected implementation logic

```
Du: if judge_confidence is not None and judge_confidence >= args.judge_threshold: break
Du: final_answer = majority_answer
Liang: moderator_has_preference(self.mod_ans) and confidence is not None and confidence >=
self.early_stop_threshold
```

Requirements. Du: numpy 1.22.4, openai 0.27.6, pandas 1.5.3, tqdm 4.64.1. Liang: openai, tiktoken 0.4.0, tqdm, langcodes, backoff, language_data.