Online Resource 2

Supplementary results

Allocating cognitive work across people, language models and deterministic tools: A task-analytic method

Simon Vajda

Applied Artificial Intelligence Institute, Deakin University, Burwood, Victoria, Australia

simonvajda@gmail.com

Cognition, Technology & Work

Study context

This supplement reports the task scores and decisions underlying the comparisons in Section 5 of the article. The four conditions used the same 64 synthetic tanker-join cases, with five repetitions per case. Tables S1–S5 report accuracy, failures and the decisions supplied by models, tools and the scripted person, a program that stood in for the exercise supervisor.

The article measures accuracy within each case before averaging across cases. Tables S1 and S3 instead pool required decisions across runs, giving greater weight to cases with more decisions. These pooled scores show where errors occurred; their proportions can differ from the article's means.

Table S1 names the eight scored tasks. The separate decision whether to deviate from the command (DEV) was normative and excluded from scoring. The scripted person supplied choices from the exercise plan and reference answers when machine tiers could not decide. No human participants took part. The small model was Qwen3-4B-Instruct-2507 and the reasoning model Qwen3.8-27B.

Completion counts, token use, paired comparisons and policy-sequence means come from the study's saved analysis. Acceptance thresholds were selected on the held-out cases. The remaining descriptive counts were calculated from the saved responses and execution records; Table S13 identifies the source files.

Accuracy by task

Table S1 Correct answers to required decisions, by task and condition

Task	Single model	Decomposed, single tier	Routed	Allocated
ACK: Acknowledge the command	259/320 (80.9)	235/320 (73.4)	204/320 (63.8)	282/320 (88.1)
LOC: Locate the tanker	179/240 (74.6)	219/240 (91.2)	217/240 (90.4)	240/240 (100.0)
SEL: Select aircraft 2	179/240 (74.6)	219/240 (91.2)	217/240 (90.4)	240/240 (100.0)
HDG: Adjust heading to intercept the tanker	169/240 (70.4)	215/240 (89.6)	217/240 (90.4)	240/240 (100.0)
CFG: Configure the distance tool	177/240 (73.8)	185/240 (77.1)	217/240 (90.4)	240/240 (100.0)
WAIT: Wait for the reporting distance	173/240 (72.1)	20/240 (8.3)	217/240 (90.4)	240/240 (100.0)
CALL: Contact the control agency	129/180 (71.7)	148/180 (82.2)	151/180 (83.9)	178/180 (98.9)
FLY: Fly to the tanker and respond on radio (parent)	124/260 (47.7)	111/260 (42.7)	201/260 (77.3)	226/260 (86.9)
All required decisions	1389/1960 (70.9)	1352/1960 (69.0)	1641/1960 (83.7)	1886/1960 (96.2)

Note. Cells give correct/required (%). A call addressed to aircraft 3 required only ACK and FLY, including under a scheduled jamming inject. For an acknowledged join instruction under scheduled jamming, CALL and FLY were excluded because their outcome depended on a training choice. A decision not reached because its run failed counts as incorrect. In the routed and allocated conditions, the counts include decisions the scripted person supplied (Table S4).

In the decomposed condition, the reasoning model gave 220 incorrect WAIT decisions. In 137 of them it reported the current distance as not yet at range, with no elapsed time, where the decision rule asked it to predict when the aircraft would reach the reporting distance. In 39 it gave no usable answer.

Completed and failed runs

Table S2 Runs completed and runs that ended with an unusable candidate or action

Condition	Runs	Completed	Failed
Single model	320	164	156
Decomposed, single tier	320	295	25
Routed	320	289	31
Allocated	320	288	32

Note. Failed runs remain in the analysis, with required decisions not reached counted as incorrect. All model requests returned a response; no request attempt failed or was retried. A failed run therefore differs from a failed model request.

The single-model reply was marked failed if it named a field outside the task list, contained an invalid answer or omitted ACK or FLY. Of its 156 failed runs, 87 failed because only the unscored deviation choice (DEV) was malformed, 60 because the reply contained fields outside the task list and 9 because a scored answer was malformed or missing. Every scored decision was correct in 78 of these runs. The completion flag and scored accuracy thus describe different aspects of the response.

The decomposed condition required the reasoning model to answer each task. All 25 failures followed an unusable answer: 13 at WAIT, 5 at CFG, 4 at ACK, 2 at CALL and 1 at FLY. The routed condition also required its chosen model to answer; one run ended with an unusable ACK answer.

The other 30 routed failures began when a model wrongly acknowledged a call meant for aircraft 3. The selection tool then rejected that addressee, and the scripted person had no reference selection for a call meant for another aircraft. All 32 allocated failures arose in the same way.

Sources of scored decisions

Table S3 Required decisions by the tier that supplied them, with the number correct

Tier	Decomposed, single tier	Routed	Allocated
Deterministic tools	–	1176/1204	1391/1393
Small model	–	171/211	171/208
Reasoning model	1352/1781	198/283	68/71
Scripted person	–	96/96	256/256
No decision recorded (run ended first)	0/179	0/166	0/32

Note. Cells give correct/supplied, pooled over all runs. The single-model condition is omitted: one reply covered every task, so its records have no per-task deciding event. The scripted person's answers were correct by construction.

Table S4 Scored decisions supplied by the scripted person, by task

Condition	ACK	CALL	SEL
Routed	0	0	96
Allocated	103	121	32

Note. ACK and CALL reached the person after both models had abstained or given unusable output. SEL reached the person after the selection tool rejected the addressee passed on from ACK and no model tier then decided. The tool requires an exact entity name, whereas scoring ignores letter case.

The selection tool rejected “Aircraft 2”, which differed from its required entity name only in letter case, 47 times in the routed condition and 29 times in the allocated condition. In the routed condition the scripted person supplied 36 subsequent selections and the reasoning model supplied 11. In the allocated condition the scripted person supplied 26, while the reasoning model supplied 3.

In the routed condition the scripted person supplied 96 SEL decisions. In 46 of them, ACK had named aircraft 2 alone in a form the tool did not accept: “Aircraft 2” (36), “Aircraft two” (7), “aircraft two” (2) and “Aircraft Two” (1). In the other 50, ACK had named two aircraft, most often as “aircraft two or three” (34).

In the allocated condition the scripted person supplied 32 SEL decisions. In 26 of them, ACK had named aircraft 2 alone in a form the tool did not accept: “Aircraft 2” (26). In the other 6, ACK had named two aircraft: “aircraft 2 or 3” (5) and “Aircraft two or three” (1).

Table S5 Who supplied the response to a scheduled jamming inject

Condition	Small model	Reasoning model	Scripted person	Other runs
Single model	–	–	–	46 answered; 34 no usable answer
Decomposed, single tier	–	47	–	33 not reached
Routed	56	–	–	24 not reached
Allocated	–	25	35	20 not reached

Note. Each condition had 80 such runs (16 cases, five repetitions). The response is a training choice, which the method assigns to people, and it was excluded from scoring. The frozen task specification classed CALL as epistemic in every case. In the allocated condition the runtime therefore sent the response to the scripted person only after both models had abstained, and a request to the person was made in exactly those runs. In the decomposed and routed conditions the models had to answer. In the single-model condition one reply covered every task, so the table counts usable CALL answers in that reply; CALL was optional there and unscored.

Requests, notices and approvals

Table S6 Mean events per run by condition

Condition	Requests to people	Notices	Approval requests	Abstentions	Escalations
Single model	0.00	0.00	0.00	0.00	0.00
Decomposed, single tier	0.68	1.31	0.00	0.00	0.00
Routed	1.17	1.36	0.00	0.83	0.83
Allocated	1.86	1.01	0.00	2.32	2.32

Note. Requests to people include the normative choices the scripted person made and the decisions it supplied. A notice reports an action to people, who may veto it.

Token use by tier

Table S7 Mean tokens per run by model tier

Condition	Small model, read	Small model, written	Reasoning model, read	Reasoning model, written
Single model	0	0	4,763	10,713
Decomposed, single tier	0	0	3,045	10,761
Routed	446	84	562	778
Allocated	846	176	709	1,124

Note. Token counts are those the model server reported for each completed request, including the reasoning model's reasoning tokens.

Paired differences

Table S8 Paired differences between conditions (95% CI)

Comparison	Accuracy (points)	Tokens read	Tokens written	Largest context (tokens)	Time per run (s)
Decomposed minus single	7.5 (−0.3 to 15.3)	−1,718 (−2,044 to −1,416)	48 (−551 to 632)	−4,137 (−4,160 to −4,116)	−2 (−31 to 27)
Routed minus decomposed	5.3 (−4.2 to 13.7)	−2,036 (−2,291 to −1,761)	−9,899 (−11,331 to −8,394)	−100 (−115 to −85)	−500 (−571 to −424)
Allocated minus routed	9.6 (3.2 to 15.7)	546 (477 to 616)	438 (337 to 543)	50 (41 to 60)	18 (13 to 24)
Allocated minus decomposed	14.9 (4.7 to 24.0)	−1,490 (−1,725 to −1,242)	−9,461 (−10,851 to −7,999)	−50 (−61 to −39)	−481 (−551 to −408)
Allocated minus single	22.4 (15.6 to 29.0)	−3,208 (−3,315 to −3,104)	−9,413 (−10,566 to −8,176)	−4,187 (−4,203 to −4,172)	−483 (−541 to −420)

Note. We calculated each difference between paired repetitions, averaged within cases and then across cases. Intervals use 10,000 percentile bootstrap resamples of cases (seed 20261010). Accuracy differences are in percentage points; time was measured under shared load.

Acceptance thresholds and withheld candidates

Table S9 Confidence thresholds selected on held-out cases

Task	Small model	Reasoning model
ACK	0	0
CALL	disabled (no threshold met the target)	0

Note. Each threshold is the lowest at which at least 10 held-out responses were accepted and no more than 5% of accepted candidates were wrong. A threshold of 0 accepts any candidate the model does not itself withhold, subject to the output checks. The router threshold was 0.4892, which sent 50% of held-out calls to the reasoning model. The held-out router scores themselves were not saved. On the test runs the router chose the small model 267 times and the reasoning model 270 times.

Table S10 Tier abstention in the paired diagnostics, by cause

Tier	Candidate	Candidates	Withheld	Explicit refusal	Malformed reply	Failed check	Below threshold
Small model	incorrect	115	112	102	0	0	10
Small model	correct	385	245	75	0	0	170
Reasoning model	incorrect	97	90	85	5	0	0
Reasoning model	correct	403	27	27	0	0	0

Note. Table 8 of the article reports withheld candidates. An explicit refusal is a reply that set its decision flag to false. A malformed reply lacked a valid decision flag; the frozen records count it with refusals as model abstention. A failed check is an invalid confidence value or hand-off. Every candidate withheld below threshold was a small-model CALL candidate, because no CALL threshold met the target for that model.

Policy sequences

Table S11 Judgment requests per case over the first and last ten cases of each policy order

Order	Seed	First ten	Last ten	Change
1	20261101	1.6	0.7	−0.9
2	20261102	1.0	1.1	+0.1
3	20261103	1.8	1.0	−0.8
4	20261104	1.7	0.6	−1.1
5	20261105	1.6	1.1	−0.5

Note. Means over the five orders were 1.54 and 0.90. The orders reuse the same 64 cases, so they show sensitivity to order and are not independent replications. Across the 320 policy-sequence runs, the scripted person stored 184 policies, and stored policies were applied 234 times.

Model requests and output limits

Table S12 Model requests and responses that reached the output limit

Phase or condition	Model	Requests	Reached limit
Held-out calibration	Small model	250	0
Held-out calibration	Reasoning model	250	0
Single model	Reasoning model	320	48
Decomposed, single tier	Reasoning model	1,907	4
Routed	Small model	267	0
Routed	Reasoning model	410	0
Allocated	Small model	560	0
Allocated	Reasoning model	419	0
Test diagnostics	Small model	500	0
Test diagnostics	Reasoning model	500	0
Policy sequences	Small model	560	0
Policy sequences	Reasoning model	420	0

Note. All 6,363 requests returned a response. The output limit was 4,096 tokens for the small model and 16,384 for the reasoning model, including its reasoning.

Source files

Table S13 identifies the analysis, calibration and descriptive-count files used for these tables. Paths are relative to collect-3910969, except for the two files under publication/. SHA-256 checksums identify the exact source versions.

Table S13 Source files and SHA-256 checksums

File	SHA-256
study/analysis/table7.csv	f09f8c0b35744ce9a5466c2174c2a2886381a4da54c638361af37dc4e8d06adf
study/analysis/table8.csv	e3560a71bf1b9be7486a75ee0e94a367f33bed37727ee46e1d1dd4b76a5aa454
study/analysis/paired-differences.csv	dc6a3818f4a2d943a0cd90bf37e999beed451637e7c05deeeb15af3da88e406a
study/analysis/tokens-by-tier.csv	43e39bc953705e0783ea13f659fc851a13b83dabcb21cac1f02b862285c4f0c5
study/analysis/policy-sequences.csv	1ed6025836c8fb7a569bbb27e9bcf97048b04bfbb46de134cd1fa2bf32fc91f7
study/analysis/run-completion.csv	697b21483bd158c56be4696aa12c8ceaf2f4a6a1710697f8711d3b09a0f8711e
study/analysis/retry-summary.csv	0771cee92d7c1e12bd9435e1fccae1ee4bb23e610c9e2fc959241967336669bc
study/analysis/results.json	4c4a46edf1420d22ad556df1bf941fdc8e6ac141afdccac69efba197c7962dce
study/calibration.json	742449e66ff8194f2f0b7582707b1bbdf2ed23c56012a754bf63cc0db8435288
publication/supplementary-counts.json	eb111425ccc16fdc0f45b10acf7c39a0be030fac132774e8e3866e2ca085c0ef
publication/supplementary_counts.py	79bc0c36ddb68b83dcaad2142e8067a01eb4903250a4fcaa75754261883b3b2e