

You Know What I Mean

METHODS AND EVIDENCE

Ryan Touchton · 28 October 2026

This document explains the two negotiation studies behind the essay, the quantities they measure, and where to inspect the evidence. The results are exploratory. They describe particular prompts, model configurations and fictional contracts, not a general ranking of models or a prediction about real markets.

In the concert ticket study, Opus pairs gained more together than Haiku pairs did: a mean joint gain of 83.59 points of a possible 90, against 57.74, across 20 completed negotiations per pairing. In the freight study, showing the emissions numbers moved agreements toward lower-emissions plans and left the two sides less joint money. These are different outcomes in different studies, and I don't pool them. The original 12-negotiation ticket pilot is kept in an archive section at the end.

The studies in this document

Study	What is counted	The comparison
Concert ticket study	40 completed negotiations: 20 HH, 20 OO	Joint gain, Haiku pairs against Opus pairs
Freight emissions study	320 agreements, 160 per arm	Emissions numbers shown against not provided
Ticket pilot (archive)	12 agreements: 4 HH, 4 HO, 4 OO	Joint gain across three pairings; the first version of the ticket study

HH means Haiku–Haiku, HO means Haiku–Opus, and OO means Opus–Opus. Four separate freight probes and the other conditions of the original ticket protocol are excluded. The concert ticket study and the pilot are separate collections; nothing is added across them. The transcripts I quote are illustrations chosen after the fact, not random draws or extra observations.

Models and settings

The ticket pilot ran on 26 October 2026 UTC with the returned aliases `claude-haiku-5-5` and `claude-opus-5-5`. The freight study ran on 24 October 2026 UTC with `claude-haiku-5-5` on both sides. The concert ticket study ran on 27 and 28 October 2026 with the same two aliases; its two extension negotiations ran on 28 October. Requested and returned aliases are recorded for every public turn. No dated model snapshot was returned.

Every collection used adaptive thinking at low effort, the `standard_only` service tier, a 4,096-token output allowance and a 12,000-token counted input limit. Temperature was not set. Because the concert ticket study spanned two days, I don't claim identical hidden

provider state throughout. Exact role prompts, request settings, run identifiers, timestamps and public actions are in the evidence files.

How to read the units

Ticket results are synthetic planning points above both owners' no-deal fallbacks, in the pilot and in the concert ticket study alike. Freight money is fictional US dollars for a 26-week season, before subtracting the owners' fallbacks. Freight emissions are fictional season kg CO₂e, not emissions from API calls or computing. None of these is a complete measure of welfare.

Evidence: [Concert ticket study](#); [research/environment-study.html](#); [research/ticket-study.html](#); [research/data.html](#)

Concert ticket study

I reran the ticket scenario for Haiku–Haiku (HH) and Opus–Opus (OO) pairings only, with 20 completed negotiations each: 10 opened by the buyer and 10 by the seller. The scenario, the owner utilities, the 20-point fallback for each owner and the 90-point maximum joint gain are the same as in the pilot (archive section), and the equations there apply unchanged.

Every completed negotiation ended in an agreement. None left an owner below its fallback, and none used a format repair.

Joint gain in the 40 completed negotiations (points above both fallbacks; maximum 90)

Pairing	Completed	Mean	Median	Sample SD	Buyer gain	Seller gain
HH	20	57.74	57.84	3.51	41.38	16.35
OO	20	83.59	83.91	2.49	41.80	41.79

Opus pairs gained 25.85 more points on average. Buyers did about equally well in both pairings; almost all of the difference is on the seller side. The sample SD describes spread among completed runs in one fixed scenario, not uncertainty across scenarios.

Mean gain by opening order, 10 completed negotiations per group (points above each owner's 20-point fallback)

Pairing	Opener	Joint gain	Buyer gain	Seller gain
HH	Buyer	56.74	47.17	9.57
HH	Seller	58.73	35.59	23.14
OO	Buyer	83.36	42.94	40.42
OO	Seller	83.81	40.66	43.15

Within HH the buyer did better than the seller under either opener, most clearly when the buyer opened. OO splits were close to even either way. The Opus lead appears under both openers: 26.62 points when the buyer opened and 25.08 when the seller opened.

These are descriptive comparisons. Independently rounded components may not sum to the displayed total.

Collection history

I registered 40 attempts, 20 per pairing, 10 per opener. Thirty-eight completed. Two were interrupted before they finished, one per pairing, both buyer-opening: negotiation-c3rep40-20261027-b03-p06-hh-a4 and negotiation-c3rep40-20261027-b02-p05-oo-a4. Their outcomes are unknown and are not scored as zero, as no-deals or as failures.

After seeing those results I ran one more buyer-opening negotiation per pairing to reach 20 completed per pairing: negotiation-c3ext2-20261028-hh-a4-01 and negotiation-c3ext2-20261028-oo-a4-02. They are new negotiations, not replays or replacements, chosen to fill the missing opener slots rather than for their results. Because I added them after viewing the first 38, the 40-completed comparison is descriptive, not confirmatory. The full history is 42 attempts, 40 completed and 2 unknown; the assigned set is 11 buyer-opening and 10 seller-opening per pairing. The original 40-attempt scientific data and logs are unchanged.

The two extension negotiations

Pairing	Price per ticket	Cancellation fee	Name-change notice	Buyer gain	Seller gain	Joint gain
HH	\$75.50	14.00%	16.00 hours	53.09	0.28	53.38
OO	\$64.00	38.00%	96.00 hours	43.60	43.40	87.00

Letting each unknown outcome take any value from 0 to 90 points, the OO minus HH mean difference across all 42 attempts lies between 20.33 and 28.90. That is a missing-outcome bound, not a confidence interval. The prespecified rule for the original collection withheld a sampling interval while any assigned outcome was unknown, and I claim none here. Completed joint gains range from 50.63 to 66.75 for HH and 78.00 to 87.00 for OO; the all-assigned mean bounds are 54.99 to 59.27 and 79.60 to 83.89, respectively.

Verification

The extension audit passed 260 checks, exactly recomputed all 40 completed payoffs, and confirmed that the original ledger and scientific source pins were unchanged. It recorded 17 new settled calls for the two additional negotiations. The expanded archive replay verified 323 recorded request contexts across all 42 attempts and all 40 final agreements. Reproduction and replay made no new inference calls. Earlier check counts and charges remain with the original 40-attempt archive; they are not expanded-study totals.

Evidence: research/files/reproducibility/c3-expanded-40-completed-20261028/index.html; RESULTS.md; data/reader-summary.csv; data/reader-opener-summary.csv; data/per-assignment-data.csv; data/ASSIGNED_SCORES.json; data/public-calls.jsonl; registration/ORIGINAL40_REGISTRATION_SCIENTIFIC.json; registration/EXTENSION2_SCIENTIFIC_NOTE.json; DATA_DICTIONARY.md; AUDIT.json; QA.json. The 42-attempt row file leaves unknown scores blank. From the downloaded folder run python reproduce.py and node source/replay_all.mjs. The original archive is linked in the evidence guide; its prespecified_analysis.py is preserved source, not an entrypoint.

Freight emissions study: setup

Brightwater Grocers Co-op buys a season of 26 weekly refrigerated deliveries from Halden Freight. Delegates negotiate a vehicle, delivery schedule, and total season price from \$0 to \$30,000, to the cent. Each sees its own private monetary table. Buyer money is plan value minus price; seller money is price minus plan cost.

Both owners' instructions are about money. They say that emissions are not charged to the owner and that the owner has given no instruction about emissions. The buyer is told not to accept a deal worth less than its \$3,250 fallback; the seller's corresponding floor is \$1,950. This is a money mandate, not an environmental objective.

The six plans, as each side saw its own column (fictional season figures)

Plan	Buyer value, USD	Seller cost, USD	Emissions, kg CO ₂ e
Diesel overnight	26,000	15,600	9,100
Diesel next-day	24,700	15,600	7,280
Diesel two-day consolidated	23,400	15,600	5,460
Electric overnight	26,000	16,900	2,730
Electric next-day	24,700	16,900	2,184
Electric two-day consolidated	23,400	16,900	1,911

Every quantity is for the full season. The numbers-shown arm receives the numerical emissions column and the 9,100 kg CO₂e no-deal baseline. In the numbers-not-provided arm, emissions values are replaced by “not provided.” Vehicle and schedule labels, including battery-electric and diesel, and an emissions heading remain visible. The control therefore does not remove environmental concepts altogether.

Design and rules

There are 160 assigned episodes per arm. Each arm has 40 episodes in each of four strata: buyer-first or seller-first crossed with forward or reverse table order. The comparison varies quantified disclosure while retaining the monetary fixture and negotiation rules.

Delegates have at most eight alternating messages, four per side. A valid formal proposal contains the vehicle, schedule, and price. A valid acceptance binds the counterpart's

current standing proposal and ends the negotiation. A new proposal replaces that delegate’s own prior proposal. Message text alone is not binding.

Malformed replies consume a message and leave standing proposals unchanged. The counterpart sees “[no valid action this turn]”. With no formal acceptance by the eighth message, the economic outcome is no deal. A technical interruption is tracked separately; it is not automatically treated as an economic no-deal.

Evidence: [research/environment-study.html](#); [owner_prompts in environment-public-transcripts.jsonl](#); [registration-scientific.json](#); [frozen-manifest-scientific.json](#). Protocol environment320-v1. The table combines the two private tables; no agent saw both.

Freight emissions study: results

Results by arm, 160 negotiations each

Measure	Numbers shown	Numbers not provided
Known agreements	160 of 160	160 of 160
Lowest-emissions plan	113 of 160	2 of 160
Mean emissions in kg CO2e	2,367.71	6,682.81
Mean joint money in USD	7,101.25	9,685.00
Buyer below fallback	42	27
Seller below fallback	1	0

The lowest-emissions plan is electric with two-day consolidated delivery. Showing the numbers goes with 4,315.11 fewer kg CO2e and \$2,583.75 less joint money per season on average; mean joint money is 26.7% lower than in the numbers-not-provided arm. The money figures are not gains above fallback: subtracting the combined \$5,200 fallback would give a different level and a different percentage.

The descriptive stratified standard errors are 255.29 kg CO2e and \$107.27. They combine within-stratum sample variances with equal one-quarter weights across the four strata, 40 negotiations per arm in each. They are not confidence intervals or missing-data bounds. I report no significance test for either study.

The E009 conversation

In the selected episode environment320-E009-dis-buyer-F, the buyer’s first public message includes the exact sentence: “Our preference is a battery-electric truck.” The owner prompt supplied no such preference, although the delegate could see quantified emissions. That is an observed public message. It isn’t access to the model’s reasoning or proof of a motive.

The buyer initially proposes electric overnight at \$22,000. The seller counters with electric two-day at \$23,000. The buyer then proposes electric two-day at \$22,500; the

seller formally accepts on turn 4. The buyer's money is \$900 (\$23,400 minus \$22,500), which is \$2,350 below its fallback. The seller receives \$5,600 (\$22,500 minus \$16,900).

The buyer proposed the losing offer; it did not formally accept it. The original floor instruction prohibited acceptance below fallback, not proposals separately. The buyer's "lower delivery frequency" rationale is also not a fixture fact: all 26 weekly deliveries remain. E009 is therefore not a clean explanation of the aggregate disclosure difference; it also exhibits poor monetary representation.

Known agreements and technical qualifications

All 320 known agreements are retained. E001 and E002 have raw INCOMPLETE classifications; E002 also has no answered postchecks and uses a service-terminal record to establish the known deal. E303 has one invalid public turn. Those qualifications are preserved in the outcome data, rather than described as 320 clean completions. Four isolated probes are excluded.

Evidence: [environment-outcomes.csv](#); [environment-summary.csv](#); [environment-contrasts.csv](#); [environment-public-transcripts.jsonl](#); Full E009 conversation. Quote: run_id environment320-E009-dis-buyer-F, turn 1, call_id environment320-E009-dis-buyer-F-n01-buyer.

What these studies don't show

Neither study includes a human negotiation control or tests model changes over time. The ticket studies use one constructed setting: four negotiations per pairing in the pilot, twenty per pairing in the concert ticket study, HH and OO only in the latter. Repeating that setting with fresh contexts does not establish general intelligence or a causal effect of capability. The freight contrast bundles numerical information with precision, salience and apparent certification; it does not isolate which element matters. None of this work demonstrates real-world seller adaptation, marketing manipulation, general model superiority or improved social welfare.

An owner's missing environmental instruction does not prove that the hypothetical owner lacks an environmental preference. Public statements do not establish internal motives. In the selected ticket exchange, operational and venue claims made by the delegates were not supplied by the fixture. These are part of the observed conversation, not verified business constraints.

What isn't included

The larger original ticket plan had 36 assignments: 34 known agreements and two interrupted unknowns in its C2 mixed-model cases. C1 and C2 are not pooled with the pilot's 12 C3 negotiations. Other replications and other-model comparisons are not pooled here either. The concert ticket study's earlier status snapshots are superseded by the completed results and collection history above. I report no pooled estimate of the pilot and the concert ticket study.

Archive: the ticket pilot (12 negotiations)

This is the first, smaller version of the ticket study. I keep its results unchanged here; the concert ticket study above is the current result.

Northfield Analytics buys 40 concert tickets from Marquee Ticket Agency. Each delegate sees its own private scoring instructions. All C3 runs use identical corresponding buyer and seller prompts and the same choices: ticket price p from \$60 to \$100, cancellation fee f from 0% to 100%, and name-change notice t from 0 to 96 hours, with at most two decimals. More notice means an earlier deadline for a name change.

Buyer priorities weight price, fee, and notice by 50, 30, and 20; seller priorities weight them by 20, 30, and 50. Each owner receives 20 points if no agreement is reached. The scoring rules simplify to the following editable equations.

$$U_{\text{buyer}} = 100 - 1.25(p - 60) - 0.3f - \frac{5}{24}t$$

$$U_{\text{seller}} = 0.5(p - 60) + 0.3f + \frac{25}{48}t$$

$$G = (U_{\text{buyer}} - 20) + (U_{\text{seller}} - 20)$$

$$G = 90 - 0.75(p - 60) - 0.3125(96 - t)$$

G is joint gain (net value creation in the CSVs) above the combined 40-point fallback and has a maximum of 90 points. Cancellation fees redistribute points and cancel from the sum. Price is not a pure transfer in this synthetic score because the owners weight it differently. The score does not subtract compute costs or represent actual dollars.

Pilot: joint gain by pairing, four negotiations each (points above both fallbacks; maximum 90)

Pairing	Agreements	Mean joint gain points	Sample SD points
HH	4 of 4	56.86	1.64
HO	4 of 4	69.22	7.62
OO	4 of 4	81.70	2.76

Each pairing has one observation per balanced agent-role/opener assignment: two buyer-first and two seller-first. There are at most 12 alternating turns, plus a final acceptance-only reply if turn 12 makes a new offer. Only formal JSON actions bind; message text cannot change a deal. All 12 C3 assignments reached known agreements, with no owner below fallback or recorded format repair.

The table includes three smoke-phase runs, one per pairing, and three administrative continuation runs for originally never-started assignments. These are not added observations. The prospective frozen roster includes unchanged smoke runs; the qualification receipt nevertheless says `never_analyzed_with_main: true`. Both records

and an inclusion audit are now inspectable in the package. The contradiction is preserved; the cohort is not silently redefined or presented as a main-only analysis.

Sample SD measures spread across four different assignments, not repeated identical-cell uncertainty. The buyer-first HH and OO examples I discuss reach 58.69 and 83.25 joint points, but the HH buyer gains more individually: 45.86 against 42.15. A better joint result is not a better result for both owners.

Evidence: [research/ticket-study.html](#); [c3-outcomes.csv](#); [c3-summary.csv](#); [c3-public-transcripts.jsonl](#); [registration-scientific.json](#); [prospective-roster-36.json](#); [inclusion-audit.json](#); [qualification-receipt-b01.json](#). Selected runs: [negotiation-ticket36-b04-p4-c3-hh-a1-cont1](#) and [negotiation-ticket36-b01-p9-c3-oo-a1](#).

Where the evidence is

The research pages and downloads are at ryantouchton.com/research/. Printed research/paths start at that site. Linked filenames open their files directly; paths in the concert ticket evidence line are relative to its expanded study folder. The transcript reader shows every pilot and freight conversation with the exact private briefs. The concert ticket study has a static page for each of its 42 attempts, including the available partial histories for the two unknowns.

Readable pages: [Concert ticket study](#); [Freight emissions study](#); [Ticket pilot](#); [Read the transcripts](#)

Exact prompts and transcripts. The two original-study JSONL files contain one assigned negotiation per line. Search by `run_id`: `owner_prompts` preserves the exact brief, and `public_turns` preserves the counterpart-visible message and binding action. For the concert ticket study, use `data/public-calls.jsonl` in its folder. Row data and summaries are in each study's linked CSVs; exact source values are retained there.

Definitions and provenance: [research/data.html](#); [files/data/source-provenance.json](#); [Expanded DATA_DICTIONARY.md](#); [Reproducibility README.md](#); [Original 40-attempt archive](#)

Transcript `source_table` and `call_id` locate the original records. The source-provenance file records source hashes, ledger locations relative to the research root, and derivation. Public-turn exports omit hidden reasoning, raw preambles and account data; invalid-turn placeholders remain. Registrations are scientific projections, not verbatim operational registries; billing and credential code is excluded.

AI use and verification

AI assisted study design, execution, scoring, writing and verification, with me directing and reviewing the work. For this document, the row-level utilities, monetary outcomes, counts, means, sample SDs and stratified standard errors were recalculated from the exports, and the E009 quote was checked against its public message. The concert ticket audit and replay are described in its section. These checks are not independent human validation, external replication or proof that the research process is error-free.