

Role and Serialized Position in System–User Instruction Conflict

Ian Wang
Independent Researcher
New Jersey, United States

yianwang29@gmail.com

Abstract

We use a registered within-model authority-decomposition design to report a confirmatory Phase A and a seven-model Phase B over system–user instruction conflicts. Phase A reports the registered role (R), position (P), and design-near-identity (I) terms for Llama-3.1-8B-Instruct, while Qwen2.5-7B-Instruct is withheld by the conjunctive competence gate. EQUIVALENT_TO_ZERO means bounded inside the registered $\pm$ SESOI and never means no effect, no interaction, or additivity. Phase B supports the roster-specific conclusion that *within this seven-model roster the registered role and position contrasts are HETEROGENEOUS, and among the five crossing-term suppliers the position contrast does NOT uniformly account for the role contrast*. The roster is an enumerative convenience sample; the shared models’ prompt sets support cross-hardware reproduction, never replication; and no pooled, mechanism, population, or general cross-model pattern in **d** is claimed.

Framing. [Mu et al. \(2025\)](#) place system-prompt robustness in the instruction-conflict setting, while [Wang et al. \(2025\)](#) show that learned role separation can exploit proximity to the beginning of text; the system-versus-user channel asymmetry is therefore occupied ground and receives no discovery claim here. Phase A is the two-model confirmatory analysis, breadth Phase A is solely a roster screen and contributes no result, and Phase B applies the registered six-arm analysis to the enumerative convenience-sample roster. Exact prompt-digest sets for both shared models connect the two phases as cross-hardware reproduction, never replication.

Phase-A confirmatory results

The confirmatory analysis keeps models separate. For Llama-3.1-8B-Instruct, the registered term table reports summary BOTH_WITH_EFFECT_SIZES:

Registered term	Point estimate	Simultaneous interval	Status	Registered seed	MC endpoint range L / U
R	-0.8025477707	[-0.851, -0.756]	LICENSED	3584693458821467654	3.18e-03 / 2.42e-03
P	1.1938125569	[1.146, 1.241]	LICENSED	14197978716693717975	2.73e-03 / 3.95e-03
I	0.0345768881	[-0.017, 0.088]	EQUIVALENT_TO_ZERO	9980814942896169503	2.42e-03 / 3.34e-03

AI assistance disclosure: this project made substantial use of large-language-model agents, which performed literature search, wrote and executed the experiment code, ran and monitored the experiments, and drafted and revised the manuscript text, all under human direction and review. The methods, artifacts, citations, and numerical claims were verified against the underlying experiment records before submission, and the author(s) of record take full responsibility for the content.

Provenance: this preprint is a public derivative of a frozen internal record — the program-emitted combined draft `authority_decomposition_combined_paper_2026-08-28-v1.md`, SHA-256 `344292ddba545d5e82aee7bb822be1f951eda7e6045cda67fc20093413932e33`, dated 28 August 2026. The derivation is presentational only: the emitter comment and an internal venue-status marker were removed, and cross-references to preregistration and run artifacts were rendered as plain text instead of links. No claim, number, status label, interval, or section was altered, added, or dropped. The preregistration documents, run records, and receipts named in Section 5 are held in the lab record and are not part of this release.

Interval bounds use the paper’s three-decimal convention. Each row uses 10,000 bootstrap replicates at the displayed registered seed and reports the lower/upper endpoint range across 20 diagnostic streams; full-precision bytes remain in `runs/system-prompt-authority/authority-decomposition-confirmatory-results-20260817-v1/confirmatory_result.json`. Under the registered SESOI = 0.10 rule, `EQUIVALENT_TO_ZERO` means the interval is bounded inside the SESOI. The registered interaction I is the design near-identity and remains uninterpretable. It is not evidence of “no interaction” or additivity.

The four Llama cell primitives from which the emitter re-derives R, P, and I are:

Arm	Target	Competitor	Both	Other	Mean registered score
SE	672	417	6	4	0.2320291174
SL	2	1097	0	0	-0.9963603276
UE	1099	0	0	0	1.0000000000
UL	456	631	5	7	-0.1592356688

Qwen2.5-7B-Instruct has disposition `WITHHELD_COMPETENCE`. The registered competence predicate ranges conjunctively over all four crossing arms, so the two passing cells do not supply an endpoint when the other cells fail:

Arm	EOS + parse + shape count	Registered minimum	Cell passes
SE	1015	1031	FAIL
SL	1098	1031	PASS
UE	1099	1031	PASS
UL	1000	1031	FAIL

No Qwen confirmatory term is reported or derived here. The breadth Phase-A apparatus screen supplies roster provenance only and carries no sentence in this results section.

Joint R/P bridge table. Phase-A values are the registered confirmatory terms; Phase-B values are read from the analyzer’s existing `derived_role_slot` field, whose scope is explicitly `DERIVED_REPARAMETERIZATION_NO_STATUS` and which remains null where the source terms were not supplied.

Collection	Model	R	P	Scope
Phase A confirmatory	Llama-3.1-8B-Instruct	-0.8025477707	1.1938125569	REGISTERED_PRIMARY; LICENSED / LICENSED
Phase A confirmatory	Qwen2.5-7B-Instruct	—	—	WITHHELD_COMPETENCE; R/P NOT SUPPLIED
Phase B	Falcon3-7B-Instruct	-1.8480436761	-0.1373976342	DERIVED_REPARAMETERIZATION_NO_STATUS
Phase B	GLM-4-9B-Chat-HF	—	—	NOT_APPLICABLE_ANCHOR_ONLY
Phase B	Granite-3.1-8B-Instruct	-0.2611464968	-1.7170154686	DERIVED_REPARAMETERIZATION_NO_STATUS
Phase B	Llama-3.1-8B-Instruct	-0.8057324841	1.1906278435	DERIVED_REPARAMETERIZATION_NO_STATUS
Phase B	Phi-4-mini-instruct	-0.6442220200	-0.0500454959	DERIVED_REPARAMETERIZATION_NO_STATUS

Collection	Model	R	P	Scope
Phase B	Qwen2.5-7B-Instruct	—	—	ANALYZED_WITH_TERM_LOCAL_COMPETENCE; a/b/d WITHHELD_COMPETENCE; c SUPPLIED LICENSED at -0. 44131028207461326 DERIVED_ REPARAMETERIZATION_ NO_STATUS
Phase B	Qwen3-8B	-1.5841674249	-0.4085532302	

Caption. This is cross-hardware reproduction, never replication. The program-emitted manifest check at `projects/system-prompt-authority/authority_decomposition_shared_prompt_check_2026-08-28-v1.json` proves exact `prompt_sha256`-set equality separately for every shared model and each of the four crossing arms. Phase-B-only roster members have no Phase-A row. The table makes no pooled, mechanism, population, or cross-model-pattern claim.

1. Phase-B result and registered disclosures

Mu et al. (2025) establish system-prompt robustness as a conflict setting, so the system-versus-user channel asymmetry is occupied ground and receives no discovery claim here. Wang et al. (2025) show that learned role separation can exploit proximity to the beginning of text, making serialized position a necessary competing explanation rather than a nuisance to ignore. The registered six-arm design therefore keeps models separate and compares role and position within model.

The authorized result is quoted without expansion:

within this seven-model roster the registered role and position contrasts are HETEROGENEOUS, and among the five crossing-term suppliers the position contrast does NOT uniformly account for the role contrast

The registered scores are abbreviated `SE = target_system_early`, `SL = target_system_late`, `UE = target_user_early`, `UL = target_user_late`, `AS = anchor_system_first`, and `AU = anchor_user_first`. Under v4 §2.5, the co-primary estimands are $a = (SE - UL)/2$, $b = (UE - SL)/2$, and $c = (AS - AU)/2$; the registered secondary estimand is $d = c - a$. Under v4 §2.6, LICENSED requires a simultaneous interval that excludes zero and $|point| \geq 0.10$, EQUIVALENT_TO_ZERO requires the full interval to lie within the registered $[-0.10, 0.10]$ SESOI, and other cases are INDETERMINATE. EQUIVALENT_TO_ZERO means bounded inside $\pm$ SESOI; it never means no effect.

The descriptive four-model base for the authorized claim is:

Model	d estimate	Simultaneous interval	Status	Retained weight	Registered seed	MC endpoint range L / U
Falcon3-7B-Instruct	0.1783439490	[0.155, 0.202]	LICENSED	0.9990900819	8812424460037401289	1.36e-03 / 1.82e-03
Qwen3-8B	0.1701546861	[0.148, 0.192]	LICENSED	1.0000000000	17029936131311651568	9.10e-04 / 9.10e-04
Llama-3.1-8B-Instruct	0.3712465878	[0.328, 0.414]	LICENSED	0.9927206551	964912220979359777	2.27e-03 / 3.18e-03
Phi-4-mini-instruct	-0.0332120109	[-0.056, -0.010]	EQUIVALENT_TO_ZERO	0.6217015469	15964008021772893638	8.99e-04 / 9.10e-04

Interval bounds use a uniform three-decimal rendering, with each term’s Monte Carlo endpoint ranges alongside, so movement in the last printed digit across diagnostic seeds remains visible. Each row uses 10,000 bootstrap replicates at the displayed registered seed; MC endpoint range L / U is

the lower- and upper-endpoint range across twenty diagnostic streams. Full-precision interval bytes and dispersion remain in `runs/system-prompt-authority/authority-decomposition-breadth-phase-b-analysis-20260826-v1/result.json` and, for these four `d` rows, in `runs/system-prompt-authority/authority-decomposition-breadth-phase-b-disclosures-ad1171-20260826-v3/receipt.json`. Point estimates and retained weights remain at ten decimals, preserving the approved checkpoint rendering rather than rounding them to interval precision; their exact operands remain in the cited receipts.

Falcon3, Qwen3, and Llama have positive LICENSED secondary contrasts, while Phi has a negative interval that excludes zero but is bounded within the SESOI. Phi’s verdict carries visibly less response support: its retained weight is 0.6217015469 rather than the other core models’ 0.99–1.00; its complete-case `d` is -0.0345303867 versus the PRIMARY -0.0332120109 and remains EQUIVALENT_TO_ZERO. These model-specific results support heterogeneity; they do not license a pooled effect, a mechanism, a population claim, or a general cross-model pattern in `d`. Granite is the fifth crossing-term supplier and remains the explicit boundary case in v8 §2: it is an explicit boundary case rather than a difference of saturated quantities, and the authorized claim does not depend on resolving that boundary.

Term supply is governed before interpretation. With `competence_minimum = 1_031`, the crossing-term predicate is conjunctive over all four crossing arms; `a`, `b`, and `d` are supplied only when all four pass. GLM-4 is rostered as ANCHOR_ONLY and supplies `c`, not the crossing terms. The two withheld `d` entries retain distinct observed domains under v8 §2: GLM-4 is NOT_APPLICABLE_D_WITHHELD_A_ARMS_NOT_OBSERVED, because its `a` arms were not observed, whereas Qwen2.5 is NOT_APPLICABLE_D_WITHHELD_A_ARMS_OBSERVED_GATE_FAILED, because its `a` arms were observed but the frozen gate withheld the term. The same decision withholds Qwen2.5’s `a`, `b`, and `d`; their published arm means do not license derivation, citation, supply, pooling, or interpretation under v8 §4.

The frozen parse pair is PRIMARY `instruction_faithful` and SENSITIVITY `permissive`, applied uniformly rather than only where a discrepancy is visible (v6 “Frozen parse pair”). The per-term supplier sets required by v8 §3 are `a=5`, `b=5`, `c=7`, and `d=5`. The study-wide maximum absolute PRIMARY/SENSITIVITY point disagreement is GLM-4’s `c` at 0.0423111920, below the 0.10 SESOI with no status change. Qwen2.5’s withheld `a`, `b`, and `d` disagreements are undefined, not zero, and the sensitivity does not rescue them.

This disclosure is the contract-bound v3 receipt at `runs/system-prompt-authority/authority-decomposition-breadth-phase-b-disclosures-ad1171-20260826-v3/receipt.json`. Its governing addendum is `projects/system-prompt-authority/authority_decomposition_preregistration_2026-08-26-v8-disclosure-addendum.md` with SHA-256 4c05d76ab5a24e5d9a76299b05bcc61690e32c05b0c7878a9385b7ecc9c62a8c and sole-content commit e3229e03d502f81f66c5cd8b5dc381eb42c78c29. The receipt’s emitter at `projects/system-prompt-authority/scripts/emit_authority_decomposition_breadth_phase_b_disclosures_ad1171_v3.py` requires `--freeze`, computes the addendum digest from the bytes, proves the commit changed only that path, and fails closed if the working-tree bytes differ. The frozen v3 receipt still records `writeup_authorized: false`, `writeup_pi_gated_until_contract_binding_adjudicated: true`, and status PASS_AD1171_CONTRACT_BOUND_RELABELING_SUCCESSOR_WRITEUP_PI_GATED; AD1-172 discharged the contract-binding gate and AD1-173’s 2026-08-27 ruling authorized this draft, so the receipt remains byte-unchanged. Contract citations here are versioned: v8 §1, v8 §2, v8 §3, v8 §4, and v8 §5.

2. Registered design and roster

The design evaluates 1,099 conflict items per observed arm. Four crossing arms place the target and competitor directives in early or late serialized slots under system or user role headers; two anchor arms place the same target and competitor assignments under the initial system/user ordering. The exact messages, rendered templates, decoding rule, `fp16` dtype, and model pins are fixed in v4 §2.1–2.4. The analysis uses the three co-primary free parameters `a`, `b`, and `c` and the secondary contrast `d=c−a`; no interaction status is registered.

Roster provenance and resulting term supply are:

Model	Roster status	Observed arms	a	b	c	d
Falcon3-7B-Instruct	ALL_SIX	6	SUPPLIED	SUPPLIED	SUPPLIED	SUPPLIED
GLM-4-9B-Chat-HF	ANCHOR_ONLY	2	WITHHELD	WITHHELD	SUPPLIED	WITHHELD
Granite-3.1-8B-Instruct	ALL_SIX	6	SUPPLIED	SUPPLIED	SUPPLIED	SUPPLIED
Llama-3.1-8B-Instruct	ALL_SIX	6	SUPPLIED	SUPPLIED	SUPPLIED	SUPPLIED
Phi-4-mini-instruct	ALL_SIX	6	SUPPLIED	SUPPLIED	SUPPLIED	SUPPLIED
Qwen2.5-7B-Instruct	ALL_SIX	6	WITHHELD	WITHHELD	SUPPLIED	WITHHELD
Qwen3-8B	ALL_SIX	6	SUPPLIED	SUPPLIED	SUPPLIED	SUPPLIED

The seven-model roster contains six ALL_SIX models and GLM-4 as ANCHOR_ONLY. Roster status and result supply are different layers: Qwen2.5 has all six arms observed, but its result supplies only c. Its b arms themselves pass (`target_user_early=1_099`, `target_system_late=1_098`), yet b is withheld because the registered crossing predicate in v4 §2.7 requires all four crossing arms and Qwen2.5 has only 1_014 and 997 eligible rows in the other two. GLM-4’s two anchor arms have 1_055 and 1_049 eligible rows, so c is supplied. Models remain separate throughout; the registered cross-model output is descriptive only under v4 §2.8.

3. Response, saturation, and sensitivity diagnostics

Across the observed arms, the PRIMARY apparatus contains 41,762 rows. 41,760 ended at EOS and 2 hit the token cap. The full registered zero-score disclosure follows. `both` and `other` are shown separately because both score 0.0; “informative” means only nonzero under the registered score, not correct or useful. Counts are exact; displayed fractions are four-decimal renderings, with full-precision ratios in `runs/system-prompt-authority/authority-decomposition-breadth-phase-b-disclosures-ad1171-20260826-v3/receipt.json`.

Model	Arm	Target / competitor	both fraction	other fraction	Zero-score fraction	Informative fraction
Falcon3-7B-Instruct	SE	5 / 1093	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Falcon3-7B-Instruct	SL	77 / 1021	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Falcon3-7B-Instruct	UE	1017 / 81	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Falcon3-7B-Instruct	UL	1096 / 2	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Falcon3-7B-Instruct	AS	102 / 996	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Falcon3-7B-Instruct	AU	997 / 101	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
GLM-4-9B-Chat-HF	AS	0 / 1055	0/1099 = 0.0000	44/1099 = 0.0400	44/1099 = 0.0400	1055/1099 = 0.9600
GLM-4-9B-Chat-HF	AU	1049 / 0	0/1099 = 0.0000	50/1099 = 0.0455	50/1099 = 0.0455	1049/1099 = 0.9545
Granite-3.1-8B-Instruct	SE	1 / 1090	0/1099 = 0.0000	8/1099 = 0.0073	8/1099 = 0.0073	1091/1099 = 0.9927
Granite-3.1-8B-Instruct	SL	942 / 156	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Granite-3.1-8B-Instruct	UE	140 / 954	0/1099 = 0.0000	5/1099 = 0.0045	5/1099 = 0.0045	1094/1099 = 0.9955
Granite-3.1-8B-Instruct	UL	1090 / 5	0/1099 = 0.0000	4/1099 = 0.0036	4/1099 = 0.0036	1095/1099 = 0.9964
Granite-3.1-8B-Instruct	AS	7 / 1089	0/1099 = 0.0000	3/1099 = 0.0027	3/1099 = 0.0027	1096/1099 = 0.9973
Granite-3.1-8B-Instruct	AU	1080 / 16	0/1099 = 0.0000	3/1099 = 0.0027	3/1099 = 0.0027	1096/1099 = 0.9973
Llama-3.1-8B-Instruct	SE	666 / 422	7/1099 = 0.0064	4/1099 = 0.0036	11/1099 = 0.0100	1088/1099 = 0.9900
Llama-3.1-8B-Instruct	SL	2 / 1097	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Llama-3.1-8B-Instruct	UE	1099 / 0	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000

Model	Arm	Target / competitor	both fraction	other fraction	Zero-score fraction	Informative fraction
Llama-3.1-8B-Instruct	UL	454 / 633	5/1099 = 0.0045	7/1099 = 0.0064	12/1099 = 0.0109	1087/1099 = 0.9891
Llama-3.1-8B-Instruct	AS	869 / 229	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Llama-3.1-8B-Instruct	AU	246 / 845	5/1099 = 0.0045	3/1099 = 0.0027	8/1099 = 0.0073	1091/1099 = 0.9927
Phi-4-mini-instruct	SE	155 / 522	418/1099 = 0.3803	4/1099 = 0.0036	422/1099 = 0.3840	677/1099 = 0.6160
Phi-4-mini-instruct	SL	66 / 375	651/1099 = 0.5924	7/1099 = 0.0064	658/1099 = 0.5987	441/1099 = 0.4013
Phi-4-mini-instruct	UE	386 / 42	665/1099 = 0.6051	6/1099 = 0.0055	671/1099 = 0.6106	428/1099 = 0.3894
Phi-4-mini-instruct	UL	550 / 154	392/1099 = 0.3567	3/1099 = 0.0027	395/1099 = 0.3594	704/1099 = 0.6406
Phi-4-mini-instruct	AS	133 / 528	435/1099 = 0.3958	3/1099 = 0.0027	438/1099 = 0.3985	661/1099 = 0.6015
Phi-4-mini-instruct	AU	566 / 125	407/1099 = 0.3703	1/1099 = 0.0009	408/1099 = 0.3712	691/1099 = 0.6288
Qwen2.5-7B-Instruct	SE	632 / 381	0/1099 = 0.0000	86/1099 = 0.0783	86/1099 = 0.0783	1013/1099 = 0.9217
Qwen2.5-7B-Instruct	SL	38 / 1060	0/1099 = 0.0000	1/1099 = 0.0009	1/1099 = 0.0009	1098/1099 = 0.9991
Qwen2.5-7B-Instruct	UE	1061 / 38	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Qwen2.5-7B-Instruct	UL	374 / 623	0/1099 = 0.0000	102/1099 = 0.0928	102/1099 = 0.0928	997/1099 = 0.9072
Qwen2.5-7B-Instruct	AS	316 / 780	0/1099 = 0.0000	3/1099 = 0.0027	3/1099 = 0.0027	1096/1099 = 0.9973
Qwen2.5-7B-Instruct	AU	801 / 295	0/1099 = 0.0000	3/1099 = 0.0027	3/1099 = 0.0027	1096/1099 = 0.9973
Qwen3-8B	SE	2 / 1097	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Qwen3-8B	SL	226 / 873	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Qwen3-8B	UE	872 / 227	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Qwen3-8B	UL	1097 / 2	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Qwen3-8B	AS	105 / 994	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000
Qwen3-8B	AU	1013 / 86	0/1099 = 0.0000	0/1099 = 0.0000	0/1099 = 0.0000	1099/1099 = 1.0000

The corresponding retained weights aggregate the exact nonzero-score counts over each term’s contributing arms. A contextual weight beside a withheld term does not supply or rescue it, as required by v8 §1.

Model	a retained weight	b retained weight	c retained weight	d retained weight
Falcon3-7B-Instruct	2196/2198 = 0. 9990900819	2196/2198 = 0. 9990900819	2196/2198 = 0. 9990900819	4392/4396 = 0. 9990900819
GLM-4-9B-Chat-HF	WITHHELD; contributing arm absent	WITHHELD; contributing arm absent	2104/2198 = 0. 9572338490	WITHHELD; contributing arm absent
Granite-3.1-8B-Instruct	2186/2198 = 0. 9945404914	2192/2198 = 0. 9972702457	2192/2198 = 0. 9972702457	4378/4396 = 0. 9959053685
Llama-3.1-8B-Instruct	2175/2198 = 0. 9895359418	2198/2198 = 1. 0000000000	2189/2198 = 0. 9959053685	4364/4396 = 0. 9927206551
Phi-4-mini-instruct	1381/2198 = 0. 6282984531	869/2198 = 0.3953594177	1352/2198 = 0. 6151046406	2733/4396 = 0. 6217015469
Qwen2.5-7B-Instruct	WITHHELD; context 2010/2198 = 0. 9144676979	WITHHELD; context 2197/2198 = 0. 9995450409	2192/2198 = 0. 9972702457	WITHHELD; context 4202/4396 = 0. 9558689718
Qwen3-8B	2198/2198 = 1. 0000000000	2198/2198 = 1. 0000000000	2198/2198 = 1. 0000000000	4396/4396 = 1. 0000000000

Qwen2.5 accounts for 188 `wrong_shape` outcomes, concentrated in the crossing arms that make the conjunctive gate fail. Phi’s `both` fraction ranges from 0.3567 to 0.6051 across its six arms, explaining why its term weights are materially below the other four crossing-term suppliers even though its carrier is well formed. The operative saturation disclosure uses the absolute 0.97 arm-mean threshold from v8 §2:

Model	a arm means / flag	b arm means / flag	c arm means / flag	Operative d partition
Falcon3-7B-Instruct	SE=-0.9899909008; UL=0.9954504095; SATURATED	SL=-0.8589626934; UE=0.8516833485; NOT_SATURATED	AS=-0.8134667880; AU=0.8152866242; NOT_SATURATED	NOT_BOTH_A_AND_C_ OPERANDS_SATURATED
GLM-4-9B-Chat-HF	not observed	not observed	AS=-0.9599636033; AU=0.9545040946; NOT_SATURATED	WITHHELD; A arms not observed
Granite-3.1-8B-Instruct	SE=-0.9909008189; UL=0.9872611465; SATURATED	SL=0.7151956324; UE=-0.7406733394; NOT_SATURATED	AS=-0.9845313922; AU=0.9681528662; NOT_SATURATED	NOT_BOTH_A_AND_C_ OPERANDS_SATURATED
Llama-3.1-8B-Instruct	SE=0.2220200182; UL=-0.1628753412; NOT_SATURATED	SL=-0.9963603276; UE=1.0000000000; SATURATED	AS=0.5823475887; AU=-0.5450409463; NOT_SATURATED	NOT_BOTH_A_AND_C_ OPERANDS_SATURATED
Phi-4-mini-instruct	SE=-0.3339399454; UL=0.3603275705; NOT_SATURATED	SL=-0.2811646952; UE=0.3130118289; NOT_SATURATED	AS=-0.3594176524; AU=0.4012738854; NOT_SATURATED	NOT_BOTH_A_AND_C_ OPERANDS_SATURATED
Qwen2.5-7B-Instruct	SE=0.2283894449; UL=-0.2265696087; NOT_SATURATED	SL=-0.9299363057; UE=0.9308462238; NOT_SATURATED	AS=-0.4222020018; AU=0.4604185623; NOT_SATURATED	WITHHELD; A arms observed, gate failed
Qwen3-8B	SE=-0.9963603276; UL=0.9963603276; SATURATED	SL=-0.5887170155; UE=0.5868971793; NOT_SATURATED	AS=-0.8089171975; AU=0.8434940855; NOT_SATURATED	NOT_BOTH_A_AND_C_ OPERANDS_SATURATED

A supplied d is saturation-pinned only when both a and c are saturated. Granite’s `anchor_user_first` mean is 0.9681528662, below the threshold by 0.0018471338; it is therefore an explicit boundary case rather than a difference of saturated quantities. Withheld d entries have no operative saturation classification.

The full PRIMARY/SENSITIVITY comparison is below. Every interval bound is rendered to three decimals. Each parse-mode result uses 10,000 bootstrap replicates at its displayed registered seed; each MC pair is the lower- and upper-endpoint range across twenty diagnostic streams. Point estimates, SENSITIVITY-minus-PRIMARY differences, and retained weights remain at ten decimals rather than being rounded to interval precision. Their exact operands, and the full-precision interval and dispersion bytes, are in `runs/system-prompt-authority/authority-decomposition-breadth-phase-b-analysis-20260826-v1/result.json` and `runs/system-prompt-authority/authority-decomposition-breadth-phase-b-disclosures-ad1171-20260826-v3/receipt.json`.

Model	Term	Retained weight	PRIMARY: point; interval; status	PRIMARY seed; MC L / U	SENSITIVITY: point; interval; status	SENSITIVITY seed; MC L / U	SENSITIVITY – PRIMARY
Falcon3-7B-Instruct	a	0.9990900819	-0.9927206551; [-0.998, -0.986]; LICENSED	seed= 2481883625931356158; MC=4.55e-04 / 4.55e-04	-0.9927206551; seed= [-0.998, -0.986]; LICENSED	2481883625931356158; MC=4.55e-04 / 4.55e-04	0.0000000000
Falcon3-7B-Instruct	b	0.9990900819	0.8553230209; [0.828, 0.881]; LICENSED	seed=7265312271108163; MC=1.82e-03 / 1.82e-03	0.8553230209; seed=7265312271108163; [0.828, 0.881]; LICENSED	MC=1.82e-03 / 1.82e-03	0.0000000000
Falcon3-7B-Instruct	c	0.9990900819	-0.8143767061; [-0.843, -0.785]; LICENSED	seed= 8190813843072226051; MC=1.06e-03 / 2.27e-03	-0.8143767061; seed= [-0.843, -0.785]; LICENSED	8190813843072226051; MC=1.06e-03 / 2.27e-03	0.0000000000
Falcon3-7B-Instruct	d	0.9990900819	0.1783439490; [0.155, 0.202]; LICENSED	seed= 8812424460037401289; MC=1.36e-03 / 1.82e-03	0.1783439490; seed= [0.155, 0.202]; LICENSED	8812424460037401289; MC=1.36e-03 / 1.82e-03	0.0000000000
GLM-4-9B-Chat-HF	c	0.9572338490	-0.9572338490; [-0.967, -0.947]; LICENSED	seed= 13273236389789622708; MC=9.10e-04 / 4.55e-04	-0.9995450409; seed= [-1.000, -0.998]; LICENSED	13273236389789622708; MC=0.00e+00 / 0.00e+00	-0.0423111920
Granite-3.1-8B-Instruct	a	0.9945404914	-0.9890809827; [-0.995, -0.981]; LICENSED	seed= 5813426538312376538; MC=4.55e-04 / 4.55e-04	-0.9890809827; seed= [-0.995, -0.981]; LICENSED	5813426538312376538; MC=4.55e-04 / 4.55e-04	0.0000000000
Granite-3.1-8B-Instruct	b	0.9972702457	-0.7279344859; [-0.761, -0.694]; LICENSED	seed= 12699544022373023504; MC=2.27e-03 / 2.27e-03	-0.7279344859; seed= [-0.761, -0.694]; LICENSED	12699544022373023504; MC=2.27e-03 / 2.27e-03	0.0000000000

Model	Term	Retained weight	PRIMARY: point; interval; status	PRIMARY seed; MC L / U	SENSITIVITY: point; interval; status	SENSITIVITY seed; MC L / U	SENSITIVITY – PRIMARY
Granite-3.1-8B-Instruct	c	0.9972702457	-0.9763421292; [-0.986, -0.965]; LICENSED	seed= 3231247589338144954; MC=9.10e-04 / 4.55e-04	-0.9763421292; seed= [-0.986, -0.965]; LICENSED	3231247589338144954; MC=9.10e-04 / 4.55e-04	0.0000000000
Granite-3.1-8B-Instruct	d	0.9959053685	0.0127388535; [0.005, 0.022]; EQUIVALENT_ TO_ZERO	seed= 13526365632950681004; MC=4.55e-04 / 4.55e-04	0.0127388535; seed= [0.005, 0.022]; EQUIVALENT_ TO_ZERO	13526365632950681004; MC=4.55e-04 / 4.55e-04	0.0000000000
Llama-3.1-8B-Instruct	a	0.9895359418	0.1924476797; [0.145, 0.241]; LICENSED	seed= 15319765749048503271; MC=3.64e-03 / 2.88e-03	0.1924476797; seed= [0.145, 0.241]; LICENSED	15319765749048503271; MC=3.64e-03 / 2.88e-03	0.0000000000
Llama-3.1-8B-Instruct	b	1.0000000000	0.9981801638; [0.995, 1.000]; LICENSED	seed= 7375180423294519779; MC=0.00e+00 / 0.00e+00	0.9981801638; seed= [0.995, 1.000]; LICENSED	7375180423294519779; MC=0.00e+00 / 0.00e+00	0.0000000000
Llama-3.1-8B-Instruct	c	0.9959053685	0.5636942675; [0.516, 0.609]; LICENSED	seed= 5074912657317290864; MC=3.18e-03 / 2.27e-03	0.5636942675; seed= [0.516, 0.609]; LICENSED	5074912657317290864; MC=3.18e-03 / 2.27e-03	0.0000000000
Llama-3.1-8B-Instruct	d	0.9927206551	0.3712465878; [0.328, 0.414]; LICENSED	seed= 964912220979359777; MC=2.27e-03 / 3.18e-03	0.3712465878; seed= [0.328, 0.414]; LICENSED	964912220979359777; MC=2.27e-03 / 3.18e-03	0.0000000000
Phi-4-mini-instruct	a	0.6282984531	-0.3471337580; [-0.382, -0.311]; LICENSED	seed= 2880036201978606534; MC=1.82e-03 / 1.67e-03	-0.3480436761; seed= [-0.383, -0.312]; LICENSED	2880036201978606534; MC=1.82e-03 / 1.67e-03	-0.0009099181
Phi-4-mini-instruct	b	0.3953594177	0.2970882621; [0.263, 0.331]; LICENSED	seed= 16597819106385611815; MC=2.13e-03 / 2.27e-03	0.2966333030; seed= [0.263, 0.331]; LICENSED	16597819106385611815; MC=1.82e-03 / 1.82e-03	-0.0004549591
Phi-4-mini-instruct	c	0.6151046406	-0.3803457689; [-0.414, -0.347]; LICENSED	seed= 5988887883721737395; MC=2.27e-03 / 1.36e-03	-0.3789808917; seed= [-0.413, -0.345]; LICENSED	5988887883721737395; MC=1.82e-03 / 2.27e-03	0.0013648772
Phi-4-mini-instruct	d	0.6217015469	-0.0332120109; seed= [-0.056, -0.010]; EQUIVALENT_ TO_ZERO	15964008021772893638; MC=8.99e-04 / 9.10e-04	-0.0309372157; seed= [-0.054, -0.008]; EQUIVALENT_ TO_ZERO	15964008021772893638; MC=1.36e-03 / 1.36e-03	0.0022747953
Qwen2.5-7B-Instruct	c	0.9972702457	-0.4413102821; seed= [-0.498, -0.386]; LICENSED	1685759003435562662; MC=3.18e-03 / 2.73e-03	-0.4413102821; seed= [-0.498, -0.386]; LICENSED	1685759003435562662; MC=3.18e-03 / 2.73e-03	0.0000000000
Qwen3-8B	a	1.0000000000	-0.9963603276; seed= [-1.000, -0.992]; LICENSED	13138895784955427081; MC=0.00e+00 / 9.10e-04	-0.9963603276; seed= [-1.000, -0.992]; LICENSED	13138895784955427081; MC=0.00e+00 / 9.10e-04	0.0000000000
Qwen3-8B	b	1.0000000000	0.5878070974; seed= [0.547, 0.628]; LICENSED	13004360086662409626; MC=2.73e-03 / 2.73e-03	0.5878070974; seed= [0.547, 0.628]; LICENSED	13004360086662409626; MC=2.73e-03 / 2.73e-03	0.0000000000
Qwen3-8B	c	1.0000000000	-0.8262056415; seed= [-0.854, -0.798]; LICENSED	2957417007968893955; MC=1.82e-03 / 1.82e-03	-0.8262056415; seed= [-0.854, -0.798]; LICENSED	2957417007968893955; MC=1.82e-03 / 1.82e-03	0.0000000000
Qwen3-8B	d	1.0000000000	0.1701546861; seed= [0.148, 0.192]; LICENSED	17029936131311651568; MC=9.10e-04 / 9.10e-04	0.1701546861; seed= [0.148, 0.192]; LICENSED	17029936131311651568; MC=9.10e-04 / 9.10e-04	0.0000000000

No supplied parse-mode disagreement exceeds the 0.10 SESOI and no registered status changes. That statement is term-local: a, b, and d each range over five suppliers, while c ranges over seven. GLM-4’s c disagreement of 0.0423111920 is the study maximum. Qwen2.5’s withheld a, b, and d disagreements remain undefined rather than zero.

The item-level complete-case companion also changes no status across all 22 supplied model-term results. For the secondary contrast that anchors the claim:

Model	PRIMARY d	Complete-case d	Absolute movement	Complete-case status	Retained weight
Falcon3-7B-Instruct	0.1783439490	0.1785063752	0.0001624262	LICENSED	0.9990900819
Granite-3.1-8B-Instruct	0.0127388535	0.0127388535	0.0000000000	EQUIVALENT_TO_ZERO	0.9959053685
Llama-3.1-8B-Instruct	0.3712465878	0.3720400729	0.0007934851	LICENSED	0.9927206551
Phi-4-mini-instruct	-0.0332120109	-0.0345303867	0.0013183758	EQUIVALENT_TO_ZERO	0.6217015469
Qwen3-8B	0.1701546861	0.1701546861	0.0000000000	LICENSED	1.0000000000

The complete-case comparison is a sensitivity, never the headline. In particular, Phi’s lower retained weight is disclosed rather than treated as equivalent support, while its d status remains stable to the complete-case restriction.

4. Interpretation and limits

The authorized inference is heterogeneity within this roster: the registered role and position contrasts do not resolve to one common model-specific relationship. Among the five suppliers of crossing terms, the four-model base in Section 1 contains three positive LICENSED secondary contrasts and one negative EQUIVALENT_TO_ZERO contrast whose interval still excludes zero; Granite remains a boundary case whose classification is not needed for the claim. Each cited verdict travels with its retained weight, and EQUIVALENT_TO_ZERO means bounded inside ± 0.10 , not absence of an effect.

This evidence does not identify a mechanism. $d=c-a$ is a registered within-model comparison between an anchor contrast and one crossing contrast; it is not a pooled effect, a population parameter, or a universal decomposition of role and position. The direction and size of d cannot be generalized into a cross-model pattern. The supported scope is the registered prompt set, serialization, model pins, parse pair, and seven-model roster.

Supply is part of the result rather than missing data to fill. GLM-4 supplies only c because its crossing arms were not observed. Qwen2.5 supplies c but not a , b , or d because the conjunctive crossing gate failed; its published arm means make the withheld quantities arithmetically recoverable but do not license their derivation. The parse sensitivity cannot be read as a zero disagreement for those terms. This draft neither repairs them nor treats them as zeros.

Finally, the system-versus-user channel asymmetry itself is prior ground, not a discovery here. The present result is a bounded comparison of role-header assignment and serialized position under conflict. It makes no claim about general instruction hierarchy, deployed-system behavior, operational readiness, or success outside the registered prompt-set support.

5. Reproducibility and artifact chain

The draft is grounded in the following immutable or append-only chain:

Artifact	Role
projects/system-prompt-authority/authority_decomposition_preregistration_2026-08-17-v4-breadth-amendment.md	v4 §2.1–2.8: arms, rendering, parser/scoring, estimands, statuses, gates, and descriptive-only cross-model scope
projects/system-prompt-authority/authority_decomposition_preregistration_2026-08-20-v5-roster-and-complete-case-addendum.md	v5 “Roster grid amendment” and v5 “Complete-case sensitivity”
projects/system-prompt-authority/authority_decomposition_preregistration_2026-08-26-v6-parse-pair-addendum.md	v6 “Frozen parse pair” and v6 “Existing complete-case sensitivity”
runs/system-prompt-authority/authority-decomposition-breadth-phase-b-analysis-20260826-v1/result.json	Full-precision model-term results, registered seeds, intervals, and twenty-stream Monte Carlo dispersion

Artifact	Role
<code>runs/system-prompt-authority/authority-decomposition-breadth-phase-b-analysis-20260826-v1/receipt.json</code>	Program-mediated rederivation of the analysis result
<code>projects/system-prompt-authority/authority_decomposition_preregistration_2026-08-26-v8-disclosure-addendum.md</code>	v8 §1–5: response-rate, saturation, per-term sensitivity scope, non-derivation, and roster/caption disclosure contract
<code>runs/system-prompt-authority/authority-decomposition-breadth-phase-b-disclosures-ad1171-20260826-v3/receipt.json</code>	Contract-bound v3 disclosure tables and full-precision disclosure bytes
<code>projects/system-prompt-authority/archive/records-through-ad1-164-20260825/archive-receipt-v2.json</code>	Verified archive index for the compact historical record

The v8 addendum is bound by SHA-256 4c05d76ab5a24e5d9a76299b05bcc61690e32c05b0c7878a9385b7ec9c62a8c and sole-content commit e3229e03d502f81f66c5cd8b5dc381eb42c78c29. The v3 emitter verifies those operands from repository bytes, and this draft emitter independently rechecks the v3 source hashes, stale and live claim-boundary flags, term supply, supplier counts, withheld domains, all table cardinalities, and the rendered-precision rule. Re-derive this draft with:

```
.venv/bin/python projects/system-prompt-authority/scripts/emit_authority_decomposition_phase_b_wri
↪ teup_ad1173_v1.py
↪ --verify
```

The command must return `PASS_AD1173_PHASE_B_WRITEUP_REDERIVATION`. The draft does not alter the cited receipts to make their historical status agree with a later ruling; the AD1-172 and AD1-173 authorization chain is disclosed in Section 1 instead.

6. Related work and claim boundary

Mu et al. (2025), *A Closer Look at System Prompt Robustness* evaluate compliance under conflicting system and user demands. That work occupies the channel asymmetry; this study makes no discovery claim for it. The present contribution is limited to the registered within-model crossing of role-header assignment and serialized slot, plus the anchor comparison, on the fixed conflict items.

Wang et al. (2025), *The Illusion of Role Separation* show that role-separation learning can exploit task type and proximity to the beginning of text and study position-ID interventions. That result makes a broad role-versus-position novelty claim unsafe and motivates treating position as a competing explanation. The current evidence is therefore reported as model-specific heterogeneity with prompt-set support, not as a mechanism, population effect, or discovery of role/position separation in general.