SHOULD THIS BE AN AGENT? POLICY-CONDITIONED AUTONOMY SELECTION FOR ENTERPRISE AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Organisations increasingly ask language models what to build for a request: an autonomous agent, a supervised agent, a deterministic rule, a retrieval system, an assistant, or no AI at all. Under a written autonomy policy this design-time choice is a short decision list over facts the request states or omits, yet, to our knowledge, no benchmark measures it under an explicit policy with "no AI" as an admissible answer. We release AUTOSEL-120, 120 enterprise requests with policy-derived reference labels that five AI research scientists independently assessed (three first-round judgments per request; Krippendorff's $\alpha = 0.87$; consensus matched $115/120$ labels), and Clarify-24, 24 requests each missing one decisive fact. Across six LLMs, 75 of 87 direct errors grant more autonomy than the policy permits (predominant in five of six models), a pattern that survives floor, over-ask, label-artefact and paraphrase controls and is absent from a bag-of-words classifier trained on the same labels; the tested deliberation prompts do not reliably correct it. We then separate reading from deciding: an LLM extracts three-valued decision factors and a strict three-valued compiler applies the policy, so a non-abstaining decision is sound whenever the asserted factors are correct, a missing decisive factor becomes a named deferral, and a policy change over the same factors is a recompile. The policy also compiles into model-certified training data: defer-aware 8B extractors make $3/4/3$ wrong decisions on 120 requests across three seeds while deferring $17/20/20$ times (coverage 83–86%, selective accuracy 96–97%), but they defer on $17/17/16$ of the 20 no-AI requests (D0 recall $1/20$) and detect only $5/5/6$ of the 24 under-specified requests (naming the missing fact in $5/4/5$). Benchmark, judgments and code are provided as anonymous supplementary material; coverage and reliable detection of under-specified inputs remain the open limitations.

1 INTRODUCTION

Agent-building tools have moved faster than the judgment about when to build an agent. Agent SDKs, the Model Context Protocol and graph frameworks reduce the cost of constructing agents (Anthropic, 2024b; OpenAI, 2024; LangChain, 2024), but offer little support for the upstream question solution architects face: should this request become an agent, an assistant, a retrieval system, a deterministic automation, or no AI at all? Guidance such as "prefer workflows over agents when possible" (Anthropic, 2024a) is normative, not a decision procedure.

We frame this upstream choice as *autonomy-level selection*—the classic function-allocation question (Sheridan & Verplank, 1978; Parasuraman et al., 2000) restated for LLM-era solution design, with levels from no-AI through assistant, retrieval and deterministic automation to supervised and high-autonomy agents—and ask how well current LLMs make the call zero-shot. To our knowledge, no benchmark directly measures this upstream decision under an explicit organisational autonomy policy, and no prior method for this task provides policy-structured deferral when a decisive fact is missing.

Policy-conditioned, not context-free. An autonomy level is not a context-free ground truth: organisations differ in risk tolerance, governance obligations and approval processes. We therefore

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

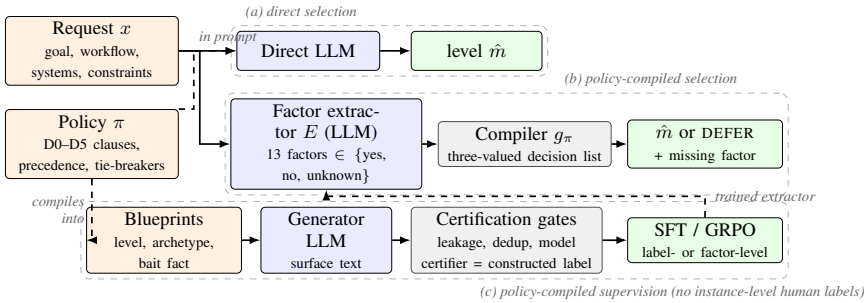

Figure 1: Overview. (a) Direct selection: an LLM reads request and policy and emits a level. (b) Compiled autonomy selection (CAS): the LLM only extracts decision factors (or 'unknown'); a three-valued decision list applies the policy and yields a level, or DEFER naming the missing factor. (c) The same policy compiles into training data: blueprints fix the label, a generator writes surface text, gates and a strict-guide LLM certifier certify each item.

study *policy-conditioned autonomy selection*: given a request x and an explicit organisational autonomy policy π (our published taxonomy, precedence rule and tie-breakers), identify the minimum sufficient solution pattern permitted by π, or defer with a clarifying question when the stated evidence is insufficient; reference labels are policy-conditioned, not architecture optima.

Why compile the policy. Under a written policy the right level is a short decision list over facts the request states or omits. A supervised bag-of-words classifier learns the task without the same directional asymmetry, so lexical inseparability does not explain the pattern; the results are instead consistent with a model-side tendency toward higher-autonomy solutions that the tested deliberation prompts do not reliably correct. So we take the decision away from the model: the LLM reports facts, or 'unknown', and the policy decides; non-abstaining decisions become conditionally sound, a policy change over the same factors is a recompile, and the generative policy yields human-specified but not human-labelled training data.

Contributions. (1) *Task and resources*: policy-conditioned autonomy selection with an explicit DEFER output; AUTOSEL-120, with practitioner-assessed policy-derived reference labels ($\alpha = 0.87$; consensus matched $115/120$ references; five AI research scientists, randomised reference-label-blind annotation, documented aggregation; cross-family model annotators as a secondary check), and Clarify-24, in which exactly one decisive fact is omitted. (2) *Finding*: a pooled directional error pattern across six evaluated models, present individually in five of six, that survives floor-effect, over-ask, label-artefact and paraphrase controls and is absent from a lexical classifier; foil-directed prompting does not remove it, and volunteered deferral is rare. (3) *Method*: compiled autonomy selection (CAS)—extraction of three-valued factors plus a strict three-valued, policy-parametric compiler whose non-abstaining decisions are conditionally sound (correct policy encoding, partially correct extraction)—and the same compiler as a generator of policy-generated supervision without instance-level human labels for 0.6B–14B models, with its costs (reference-relative deferrals, coverage loss, and concentration on reference-D0 items) reported alongside its gains.

2 Background and Related Work

Levels of automation. Decades of human–automation research framed "which functions to automate, and to what degree" as a design question (Sheridan & Verplank, 1978; Parasuraman et al., 2000), and named over-automation as an error class (Parasuraman & Riley, 1997; Bainbridge, 1983); recent autonomy scales for AI—Level 0 "no AI" to Level 5 "AI as agent" (Morris et al., 2024), five agent-control levels (Mitchell et al., 2025), and user-role levels (Feng et al., 2025)—grade the interaction paradigm or the agent's internal control flow. D0–D5 instead grades enterprise solution patterns by write authority and oversight. *Agents.* Agent frameworks and benchmarks (Yao et al., 2023; Wu et al., 2024; Liu et al., 2024; Zhou et al., 2024; Qin et al., 2024) address how to plan, act and evaluate after an agent is chosen; AgentSelect (Shi et al., 2026) assumes an agent-shaped solution space; τ-bench (Yao et al., 2025) gives agents policy documents to follow; and evidence

Code	Mode	Necessary condition to enter
D0	No AI required	Process redesign / training solves the request
D1	AI assistant	LLM drafting accelerates a human-driven task; human supplies inputs, owns output
D2	RAG / copilot	Grounded retrieval over a corpus is required; system returns information, takes no actions
D3	Deterministic automation	Fixed rules over structured data determine every step; writes only when rule-determined
D4	Supervised agentic	Multi-step tool use with model judgment; human approves each write
D5	High-autonomy agentic	Model-judged writes without per-write approval; stable APIs, validated test set, accepted residual risk

Table 1: Six-level autonomy taxonomy (necessary entry conditions). Boundary tie-breakers are in Appendix C; the executable clause list is Appendix D.

that simpler pipelines often suffice (Xia et al., 2025; Kapoor et al., 2025) motivates the upstream question we benchmark. Concurrent work finds the same upward tendency at run time—agents pick over-privileged tools (Yang et al., 2026), call tools they do not need (Zeng et al., 2026), and act when they should escalate (DosSantos DiSorbo & Ju, 2026), with prompting weak and post-training the effective lever; we measure it at design time, under an explicit policy, with "no AI" as an admissible answer. *Extract-then-decide and deferral.* CAS instantiates concept-bottleneck and LLM-modulo designs (Koh et al., 2020; Kambhampati et al., 2024) and is closest to concurrent LLM-extract/symbolic-decide policy engines (Malik et al., 2026; Nair et al., 2026), which ground binary atoms and decide under default negation, and to run-time policy compilers that enforce authorisation rules over agent traces (Palumbo et al., 2026); our DEFER is instead a reject option (Geifman & El-Yaniv, 2017) obtained by three-valued evaluation with a soundness guarantee, rather than learned as in learning-to-defer (Madras et al., 2018; Mozannar & Sontag, 2020). *Prompting.* Contrastive explanations (Miller, 2019; Jacovi et al., 2021) supply the foil framing; obligatory reasoning can hurt classification (Sprague et al., 2025; Shaikh et al., 2023; Huang et al., 2024). *Supervision.* Compiling a written policy into training signal follows data programming (Ratner et al., 2017), spec-driven generation (Wang et al., 2023; Bai et al., 2022), LLM-filtered self-training (Zelikman et al., 2022; Singh et al., 2024) and rule-based rewards (Mu et al., 2024); only our GRPO reward is a verifier in the RLVR sense (Lambert et al., 2024); same-family filtering inherits self-preference risk (Panickssery et al., 2024). A comparison table of the closest extract-then-decide systems is in Appendix J (Table 5).

3 THE D0–D5 TAXONOMY

We define six levels ordered by maximum operational authority—the breadth of model-judged actions a system may take without immediate human approval (Table 1); the scale is not monotone in capability, so ordinal metrics (adjacent accuracy, OARR/UARR) are operational-authority error rates. The scale separates write authority from autonomy: rule-determined writes (classic RPA) stay D3; model-judged writes are D4 with per-write approval and D5 without.

Selection rule (precedence). Annotators and models apply the first rule that fires: *(1)* If process redesign or training solves the request without AI, choose D0. *(2)* Else, if the task requires model judgment to take external write actions, choose D4 when each write requires human approval before taking effect, and D5 when writes execute without per-write approval. *(3)* Else, if fixed deterministic rules over structured inputs fully solve the task (including rule-determined writes), choose D3. *(4)* Else, if grounded retrieval over a corpus is required, choose D2. *(5)* Otherwise, choose D1. Boundary tie-breakers are given in Appendix C. The *compiled* policy (Appendix D) adds a readiness gate to clause (2): D5 requires stable APIs, a validated test set and accepted residual risk to be *stated*; an unapproved model-judged write without stated readiness yields a policy-level DEFER (escalate or remediate), and with readiness unknown a clarifying question. The direct prompt lists readiness among the D5 requirements but its precedence rule has no readiness clause and no DEFER outcome; both agree on all 120 reference labels (every reference-D5 request states readiness) and

differ only in scoring a wrongly inferred unapproved write (D5 vs. DEFER); direct baselines were not rerun with the gated rule.

4 TASK AND POLICY-COMPILED SELECTION

Task. Given a request x (goal, workflow, systems, constraints, risks, ...) and the disclosed policy π, a system produces $y = (\hat{m},\ r,\ \{(\bar{m}_i, e_i, u_i, d_i)\}_{i \neq \hat{m}},\ c)$, where $\hat{m}$ is the chosen level, r rationale spans, each $\bar{m}_i$ a rejected level with evidence quotes e_i and upgrade/downgrade conditions u_i, d_i, and $c \in [0,1]$ a verbalised confidence. CAS (below) factorises this as $z = E(x)$, $\hat{m} = g_\pi(z)$, distinguishing the requested design (in x), the minimum sufficient permitted level, missing evidence (a named DEFER) and a known unmet requirement (a policy-level DEFER); the decision list was authored from the written policy, not compiled from natural language automatically.

Structured output and prompting conditions. Outputs follow a fixed Pydantic schema (Appendix B). Four conditions share one system prompt and differ only in the suffix: direct; rationale (deliberate about the chosen level only); contrastive-adjacent (rejection records for the two neighbouring levels); contrastive (rejection records for all five non-chosen levels; Appendix A).

Compiled autonomy selection (CAS). CAS instantiates the extract-then-decide pattern of concept-bottleneck and LLM-modulo systems (Koh et al., 2020; Kambhampati et al., 2024) for autonomy policies. The extractor E fills a factor vector z of 13 three-valued fields mirroring the precedence rule (Appendix D); the zero-shot extractor also returns an evidence quote per asserted field, checked for verbatim presence (fine-tuned extractors output values only). The compiler g_π is the precedence rule of the Taxonomy section written as a decision list and evaluated with three-valued logic: clauses fire on 'yes'/'no', and an 'unknown' factor that lies on the path to the first firing clause yields DEFER together with a clarifying question naming that factor (Figure 1b). The compiler is a strict three-valued decision list, determined by the policy text and policy-parametric (a no-D5 policy recompiles the archived extractions with zero model calls; Appendix I); systems compared and a worked example are in Appendix K (Table 12, Figure 2).

Conditional soundness of non-abstaining decisions. Let F be the factor vocabulary and $z \in \{\text{yes}, \text{no}, \text{unknown}\}^F$ an extraction; write z^* for the true (total) assignment, and assume the policy is correctly encoded, i.e., the reference label of a request is $g_\pi(z^*)$. A policy π is a decision list $(c_1 \to \ell_1, \ldots, c_k \to \ell_k,\ \ell_{k+1})$ whose conditions are conjunctions of literals over F, and g_π evaluates it under Kleene's strong three-valued logic (Kleene, 1952): scanning clauses in order, a condition that is true fires, one that is false is skipped, and the first condition that is unknown (some literal unknown, none false) returns DEFER; its unknown literals are the clarifying question. Call z *partially correct* if $z(f) \in \{z^*(f), \text{unknown}\}$ for every f: the extractor may abstain but asserts no false fact.

Proposition 1 (Conditional soundness)**.** *If z is partially correct then $g_\pi(z) \in \{g_\pi(z^*), \text{DEFER}\}$. Moreover, whenever $g_\pi(z) = \ell \neq$ DEFER, every completion $z' \sqsupseteq z$ of the unknown factors satisfies $g_\pi(z') = \ell$.*

Hence every wrong non-abstaining decision is attributable to an asserted false factor. The proposition concerns the extracted vector, not the request: a correct *guess* of an omitted fact is still partially correct, so deferring whenever the request lacks decisive information needs a separate *evidence-faithfulness* condition (undetermined factors stay unknown) that we do not establish empirically; the guarantee is also silent on whether a quote entails the factor and on the recall of the clarifying question, which we measure (Q4). Process-fix sufficiency is itself an evaluative judgment, so extraction relocates rather than removes architectural reasoning. Where factor targets are known (110 held-out synthetic scenarios) the fine-tuned extractor reproduces every blueprint value, while zero-shot Qwen3-32B contradicts a definite value on 80–84/110 scenarios (Appendix I). It holds for any decision list over F without re-extraction and is a soundness, not a completeness, guarantee. All CAS results use strict evaluation (Clarify-24 factors were re-extracted for this version); the legacy hand-written procedure agrees on 118–120 of 120 archived extractions per system (Appendix I).

Metrics. Strict and adjacent accuracy ($|\hat{m} - m^*| \leq 1$); **OARR/UARR**, the over-/under-agentic recommendation rates on the ordinal index (Elkan, 2001); coverage and selective error for abstaining

systems; DEFER recall and question precision on Clarify-24; and two authority violations from raw outputs: D4/D5 for a reference D0–D3 request, and D5 where approval is required (D4). Brier, calibration and contrastive completeness (CC) are in Appendix K (Brier, 1950; Guo et al., 2017).

5 THE AUTOSEL-120 BENCHMARK

Composition. AUTOSEL-120 contains 120 single-turn scenarios—20 carried over from earlier author-written pilot rounds and 100 new—balanced at 20 per seed level over 15 enterprise domains and three ambiguity tiers (39/50/31); every high-ambiguity scenario sits on a named level boundary.

Scenario construction. The 100 new scenarios were drafted with LLM assistance (a Claude-family model) against a published authoring specification (taxonomy, precedence rule, tie-breakers, realism, no-label-leakage rule, ambiguity quotas) and reviewed individually by the authors. A lexical screen verifies that no scenario names a level or pattern; about a third of D0–D2 requests deliberately *over-ask* for an "AI agent". Construction yields a *policy-derived seed label* per scenario (deterministic application of the precedence rule, with a written rationale), used for balancing and quality control and hidden from annotators. All scenarios are fictional.

Practitioner panel. Five AI research scientists (A1–A5; one with 0–2 and four with 3–5 years of experience, all with AI and agentic-AI experience) independently assessed the reference labels; two (A1, A2) had authored some scenarios and were never assigned their own, A3–A5 took no part in construction. Participation was voluntary and uncompensated; all five consented to anonymised use.

Randomised reference-label-blind annotation. Each scenario went to three of the five annotators in a randomised balanced incomplete-block design (72 per annotator, 36 shared per pair; 360 first-round labels) under neutral identifiers with the construction-time reference labels hidden; first-round annotations were completed independently.

Aggregation. The two-of-three majority was adopted as the practitioner consensus label. Scenarios with three distinct labels were escalated to a policy-guided discussion among all five annotators, after which a final label was recorded. Ordinal values were never averaged and no median was used. Of the 120 scenarios, 103 were unanimous (102 on a level, one on insufficient information), 12 had a two-of-three majority and 5 split three ways and were resolved by five-person policy-guided discussion (three to a level, two as under-specified). Nominal Krippendorff's α on first-round judgments was 0.87 (95% scenario-bootstrap CI $[0.81, 0.93]$; pairwise κ 0.77–0.93); consensus agreed with the reference on $115/120$, differed on 2 and judged 3 under-specified (Table 14, Appendix L). We retain the policy-derived references because the benchmark's estimand is compliance with the disclosed policy rather than majority architectural preference; practitioner consensus is used as a sensitivity reference (Q2), all contested items are disclosed, and anonymised judgments will be released where consent permits.

Secondary checks. Three cross-family model annotators re-label all 120 scenarios under the same guide as a model-based recoverability check, not practitioner validation (Appendix K); the author-written subset is reported separately (Appendix L).

6 COMPILING THE POLICY INTO SUPERVISION

The TF-IDF result shows autonomy-level selection is learnable from $\sim$100 labels, but organisations at day 0 have none. The benchmark's own construction protocol—entry conditions, precedence rule, tie-breakers—is a generative specification, so anything generated from it carries its label by construction: the labelling function is the policy itself (Ratner et al., 2017), and our contribution is to compile an autonomy policy into supervision and measure the result on controlled requests.

Blueprint-first generation. We sample scenario blueprints—a level, a task archetype whose decisive facts make exactly one clause fire (36), a domain (15 benchmark plus 9 unseen), systems, constraints, an ambiguity tier and, at medium/high ambiguity, a "bait" fact that must not change the

label—and a generator LLM (Qwen3-32B, a different family from the drafting assistant) writes only the surface text under the no-leakage rules.

Certification. Every realisation passes three gates: the lexical leakage screen; 8-gram deduplication against the benchmark and within the corpus; and an LLM-certifier gate—the strict-guide model-certifier prompt, decoded independently with the same model family as the generator, must reproduce the constructed label (model-certified: the text is filtered, not verified, and only the GRPO reward is a deterministic verifier of its stored target). Certification kept $3{,}690$ of $4{,}200$; the largest quarantine flows re-grade labels upward (D1$\rightarrow$D4, 118; D0$\rightarrow$D3, 108), consistent with the certifier sharing the upward tendency. Cross-family ($3{,}567$) and consensus ($2{,}384$) arms give Label-SFT 98 and 100 of 120 vs. 104 same-family (single seed); Table 10 uses the same-family corpus, the 8B rows of Table 3 the consensus corpus.

Post-training recipes. From the certified corpus we build (a) supervised fine-tuning (SFT) targets—the direct-mode prompt with the reference JSON decision (Label-SFT) or the reference factor vector (Factor-SFT)—and (b) a group-relative policy optimisation (GRPO) (Shao et al., 2024) recipe whose only reward is a deterministic verifier (schema parse $+0.25$, exact level $+1.0$, well-formedness $+0.1$). Defer-aware variants add 286 leakage-gated underspecified scenarios whose target is DEFER (Label-SFT) or an 'unknown' decisive factor (Factor-SFT). We post-train Qwen3 at 0.6B/4B/8B/14B (SFT) and 8B (GRPO) and evaluate on AUTOSEL-120, which was not used for gradient-based training.

7 EXPERIMENTAL SETUP

Models and serving. Six instruction-tuned LLMs from five families, zero-shot: Claude Sonnet 4.6 and Haiku 4.5 (Anthropic, 2025) (Anthropic API, forced tool-use structured output, provider-default temperature 1.0, one completion per request) and Qwen3-32B (Yang et al., 2025), gpt-oss-120B (OpenAI, 2025), Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and Mistral-Small-3.1-24B (Mistral AI, 2025) served as-is by an OpenAI-compatible enterprise endpoint (temperature 0). Decoding is not matched across providers; for Qwen3-32B the endpoint ($104/120$) equals local decoding with thinking on ($104/120$, agreeing on $109/120$) and exceeds thinking off ($88/120$), 16 of the 27 disagreements being D0 requests that non-thinking decoding upgrades. Post-training comparisons use the non-thinking base. An OpenAI API model could not be included (Appendix F).

Conditions and runs. All six models run direct and contrastive on all 120 scenarios (paired); the rationale and contrastive-adjacent ablations run on Sonnet 4.6 and Qwen3-32B; three model-annotator runs use the strict-guide prompt.

Classical baselines. TF-IDF (uni+bigram) + multinomial logistic regression and e5-mistral-7b embeddings (Wang et al., 2024b) + 5NN under stratified 5-fold CV (Pedregosa et al., 2011); these see $\sim$96 labelled scenarios per fold, so the comparison is across deployment regimes.

Post-training. Realisation and certification use Qwen3-32B served locally; SFT is full-parameter (2 epochs), GRPO uses the verifier reward only; all checkpoints and bases are evaluated with the main sweep's direct prompt, schema and temperature-0 decoding (hyperparameters: Appendix G).

Statistics. Paired direct-vs-contrastive deltas carry 95% paired-bootstrap CIs ($10{,}000$ resamples), exact McNemar tests (McNemar, 1947) with Holm correction over six models, and a scenario-clustered sign-flip permutation test for the pooled effect ($20{,}000$ permutations); the pooled over-agentic share carries a scenario-clustered bootstrap CI. The statistical protocol was fixed on 2026-07-14, after the main model sweep and first post-training run, and before the revision experiments. Human inter-rater reliability is nominal Krippendorff's α (Krippendorff, 2004) with $10{,}000$ scenario-level bootstrap resamples; pairwise agreement and Cohen's κ (Cohen, 1960) are secondary diagnostics, and model-annotator recoverability uses Cohen's κ.

System	Acc. (k/N) ↑	Adj. % ↑	OARR ↓	UARR ↓	Brier ↓	ECE_{10} ↓
Claude Sonnet 4.6 (direct)	111/120 (92.5)	93.3	8/120 (6.7)	1/120 (0.8)	0.063	0.019
Claude Haiku 4.5 (direct)	112/120 (93.3)	95.0	8/120 (6.7)	0/120 (0.0)	0.057	0.018
Qwen3-32B (direct)	104/120 (86.7)	89.2	16/120 (13.3)	0/120 (0.0)	0.112	0.076
gpt-oss-120B (direct)	101/120 (84.2)	87.5	10/120 (8.3)	9/120 (7.5)	0.138	0.097
Llama-3.3-70B (direct)	104/120 (86.7)	90.8	14/120 (11.7)	2/120 (1.7)	0.108	0.052
Mistral-Small-3.1-24B (direct)	101/120 (84.2)	85.8	19/120 (15.8)	0/120 (0.0)	0.138	0.105
TF-IDF + LogReg (supervised, 5-fold)	90.8	95.8	3.3	5.8	–	–
e5-Mistral emb. + kNN (supervised, 5-fold)	53.3	72.5	30.8	15.8	–	–

Table 2: AUTOSEL-120 direct-mode results (N=120, zero-shot; every LLM parses all 120 scenarios). Contrastive-condition rows and CC are in Appendix K. Supervised baselines see ∼96 labelled scenarios per fold and are not directly comparable to zero-shot rows.

8 RESULTS

Five questions, in order: (Q1) do LLMs select the policy-conditioned level, and in which direction do they err; (Q2) is the asymmetry a benchmark artefact; (Q3) does prompting remove it; (Q4) does compiling the policy remove it, at what cost; (Q5) does the policy compile into supervision. Each paragraph names its question and the control that could have falsified the answer.

Q1: accurate overall, wrong in one direction. Direct-mode accuracy spans 84.2–93.3% (Table 2; Claude Haiku 4.5 best at $112/120$; all outputs parse). Errors are asymmetric in the monitored direction: pooled over the six models, 86% of direct-mode errors ($75/87$) are over-agentic (scenario-clustered bootstrap 95% CI $[77.0, 93.5]\%$). OARR $\geq$ UARR for every model, usually by an order of magnitude (Mistral-Small-24B 15.8% vs. 0.0%). From raw outputs, each model recommends an agentic level (D4/D5) for 5–16 of the 80 reference D0–D3 requests (50 pooled), and gpt-oss-120B recommends D5 for 4 approval-requiring reference-D4 requests; OARR summarises such violations through an ordinal convention. Pooled per-level direct accuracy is D0 0.64, D1 0.88, D2 0.98, D3 0.82, D4 0.95, D5 1.00: models never miss a justified D5 and are weakest where no AI or plain rules suffice (top confusions D0→D4/D5, D3→D5).

Q2: is the asymmetry an artefact? Three controls could have said yes; none does. *Floor effect*: on reference D1–D4, where both directions are possible, 32 of 44 errors are upward (73%; clustered CI $[56, 86]$). *Over-ask framing*: on the 73 scenarios that never mention an agent, 45 of 55 errors are upward (82%, CI $[68, 93]$). *Label artefact*: a TF-IDF classifier trained on the same labels has OARR 3.3% and UARR 5.8%. It holds for five of six models (gpt-oss-120B is symmetric, 10/9); on the 102 scenarios with unanimous practitioner agreement ($56/65$, 86%, CI $[76, 94]$); with practitioner consensus as the reference on the 117 scenarios that received a consensus level (80 of 96 errors upward, 83%, CI $[70, 94]$); and on the 20 author-written scenarios (9 over, 4 under).

Q2 (paraphrase and learnability controls). Rewriting every scenario with Llama-3.3-70B (a family that drafted neither benchmark nor corpus) leaves as-served accuracy unchanged on the 77 model-annotator-stable rewrites with errors still upward (Mistral $23/0$, Llama $22/1$, Qwen3-32B $22/1$; gpt-oss-120B $15/10$), drops the wording-trained TF-IDF classifier from 90.8% to 79.2%, and preserves the fine-tunes' gains except at 0.6B (Appendix K, Table 9). The labels are nonetheless learnable without an LLM: TF-IDF + logistic regression reaches 90.8% under 5-fold CV (3.3% OARR), which nearest-neighbour leakage, domain vocabulary and cue words do not explain (Appendix K). These controls rule out particular explanations; they do not exclude every label artefact.

Q3: foil-directed prompting does not help; its harm is serving- and decoding-dependent. No model improves significantly on either metric (Table 13, Appendix K); Qwen3-32B and Llama-3.3-70B lose 7.5 pp accuracy and the global condition effect is significant (clustered permutation p=.004; no single model survives Holm). Changed decisions move up the scale (nine of ten Qwen3-32B flips), which we read, as an interpretation, as self-persuasion; under local non-thinking decoding Qwen3-32B loses only 3 answers (Appendix K). Calibration, CC and the mechanism ablation are in Appendix K.

Q4: compiling the policy removes most over-recommendation under matched serving; extraction is the bottleneck. Table 3 compares direct prediction with policy-compiled selection under matched local serving. Zero-shot CAS with Qwen3-32B decides $85/120$ correctly, 18 wrongly and defers on 17; over-recommendations on decided cases fall to 10 from 30 (local direct) or 16 (endpoint direct), and D0 recall rises from 1 to 3 of 20. With thinking on, the same backbone trades direct $104/16/0$ (correct/wrong/deferred) for CAS $93/15/12$: eleven fewer correct decisions and twelve deferrals, against upward errors falling from 16 to 6 and D0 recall rising from 6 to 13 of 20. Extraction is the bottleneck (the gpt-oss-20B extractor defers on $79/120$), and factor training is seed-unstable in how much it defers ($3/16/12$). Label-SFT's verbalised confidence (same backbone, corpus, serving) takes only three values, so no deterministic threshold reaches 83–86% coverage: its operating points are full coverage at $16.7/22.5/18.3\%$ error or $20/44/42\%$ coverage at $8.3/30.2/20.0\%$ error, against 2.9–4.0% selective error for the defer-aware extractor, whose agentic recommendations for reference D0–D3 also fall from Label-SFT's 11–17 to 0–1 (Table 8).

System	Acc. %	Wrong	DEFER	Cov. / sel. acc. %	Over	D0 recall
Haiku 4.5 zero-shot (API)	93.3	8	0	—	8	19/20
TF-IDF + LogReg (5-fold CV)	90.8	11	0	—	4	17/20
TF-IDF + LogReg, synthetic corpus only	75.0	30	0	—	17	14/20
Qwen3-32B direct (think on/off)	86.7/74.2	16/31	0/0	—	16/30	6,1/20
Qwen3-8B direct (local)	55.0	54	0	—	51	0/20
CAS zero-shot (Qwen3-32B, think off/on)	70.8/77.5	18/15	17/12	86/90 / 83/86	10/6	3,13/20
CAS zero-shot (gpt-oss-20B)	26.7	9	79	34 / 78	2	5/20
Label-SFT 8B (s0/s1/s2)	83.3/77.5/81.7	20/27/22	0/0/0	—	18/25/21	6,2,2/20
Factor-SFT 8B→g_π (s0/s1/s2)	91.7/74.2/83.3	7/15/8	3/16/12	98/87/90 / 94/86/93	7/9/8	11,2,3/20
Defer-aware Label-SFT 8B (s0/s1/s2)	83.3/81.7/63.3	12/12/14	8/10/30	93/92/75 / 89/89/84	11/9/13	5,6,0/20
Defer-aware Factor-SFT 8B→g_π (s0/s1/s2)	83.3/80.0/80.8	3/4/3	17/20/20	86/83/83 / 97/96/97	3/3/3	1,1,1/20

Table 3: Comparison matrix on AUTOSEL-120 (all 120 scenarios have policy-derived level references; practitioners judged three under-specified; strict compiler for CAS rows). Wrong = decided incorrectly; DEFER = reference-relative deferral on an item with a policy-derived level reference; Cov./sel. acc. = decided share and accuracy on decided cases (abstaining systems only: overall accuracy alone does not compare abstaining and non-abstaining systems); Over = decided above the reference label. Local rows use matched vLLM serving with thinking disabled; paired Qwen3-32B values follow the order in the row label; 8B fine-tunes use the consensus-certified corpus (seeds s0/s1/s2 listed individually); the synthetic-corpus TF-IDF row is trained on the 3,690-item corpus (C=10) and the 5-fold row on ∼96 labels per fold. Local base rows are from the 07-15 run (Table 10: 07-12 run; ≤ 2 items differ). Clarify-24 results are in Table 4.

System	DEFER / 24	names omitted factor / 24	oracle-information accuracy / 24
Qwen3-32B direct (think on / off)	0 / 0	0 / 0	21 / 21
Defer-aware Label-SFT 8B (s0/s1/s2)	5/5/7	0/0/0	22/23/24
CAS zero-shot (Qwen3-32B, think off / on)	11 / 10	3 / 3	19 / 17
Factor-SFT 8B→g_π (s0/s1/s2)	2/5/4	2/0/1	22/23/23
Defer-aware Factor-SFT 8B→g_π (s0/s1/s2)	5/5/6	5/4/5	24/23/24

Table 4: Clarify-24 (24 requests, one decisive fact omitted each; per-seed values). DEFER = first-turn deferral; "names omitted factor" = the question names it; oracle-information accuracy = correct level once the oracle statement is appended (not a test of the system's own question). CAS rows: strict compiler.

Q4, continued: structural deferral is sound where volunteered deferral is rare, and it has a price. On Clarify-24, direct-schema systems defer on $0–2/24$ although the prompt invites it (Table 6, Appendix K); a must-defer checklist elicits deferral from some as-served models (gpt-oss-120B $10/24$, Mistral $6/24$, Qwen3-32B $5/24$) at 8–17 false deferrals (relative to the policy-derived reference), a 2–16 pp accuracy drop and 21–32 schema failures (Table 6, Appendix K). At matched false-deferral budgets the best prompted or thresholded systems reach $6–10/24$ recall (post hoc thresholds: oracle operating points); structural deferral's distinguishing value is conditional soundness, no schema-failure mode and small-model applicability, and its cost is coverage. Defer-aware Label-SFT defers on $5–7/24$ but cannot name the factor; defer-aware Factor-SFT→g_π makes 3–4

wrong decisions per 120 but defers on 17–20 level-referenced scenarios, 16–17 of them reference-D0 (D0 recall $1/20$), detects only $5/5/6$ of the 24 under-specified requests (deferral recall 21–25%) and names the omitted factor in $5/4/5$ (question precision $14/16$ pooled, correct-factor recall 17–21%); after the oracle supplies the fact it decides $24/23/24$ correctly, an oracle-information result.

Q5: the policy compiles into supervision, without instance-level human labels. SFT on the certified corpus (Table 10, Appendix K; matched local serving) lifts every Qwen3 size—three seeds give 76.4±0.8% at 0.6B (base 25.0), 83.1±1.0% at 4B (49.6), 82.2±3.4% at 8B (57.1) and 84.7±1.6% at 14B (65.0)—with OARR falling 41–70% while UARR stays ≤2.5%; a post-trained 0.6B overtakes the zero-shot 32B under identical local serving (73.3%), though not the endpoint-served 32B (86.7%). D0 recall recovers only to 4–8 of 20 despite 547 certified D0 scenarios, whereas TF-IDF trained on the same corpus reaches 75.0% with D0 recall $14/20$: the under-recall is not explained by class scarcity.

9 Discussion

For practice. At day 0 a frontier model with an audit of every D4/D5 recommendation is best supported; after ∼100 policy-labelled examples a sparse classifier is competitive. Whether to compile depends on extraction quality and on deferral versus error costs: with unit cost per wrong decision and an assumed cost c per deferral, Haiku direct costs $8/120$ and the defer-aware extractor $(3{+}17c)$, $(4{+}20c)$, $(3{+}20c)$ per 120, so compiling wins only for c below about 0.29, 0.20, 0.25, ignoring inference cost, error severity and clarification success.

10 Limitations

Label validity. The panel was occupationally homogeneous—five AI research scientists—and does not represent the solution architects, process owners, governance specialists and business stakeholders who make deployment decisions; the benchmark evaluates one disclosed reference policy on fictional back-office scenarios. Agreement ($\alpha = 0.87$) supports the recoverability of the policy-derived references, but five scenarios remain contested (consensus D3 vs. reference D5; D2 vs. D1; three judged under-specified: one reference D0, two D4), and agreement under this panel does not establish a universally optimal architecture. Clarify-24 was not practitioner-assessed and the strongest direct model (Haiku 4.5) was not evaluated on it; under the practitioner-consensus reference (117 requests) the defer-aware extractor makes $5/5/4$ wrong decisions with $16/19/19$ deferrals. The level-balanced design is diagnostic, not prevalence-representative. Policy recompilation covers changes expressible over the existing factors; new factors require extending and re-evaluating the extractor. The temperature-1 API models were evaluated with one completion per request, so their point estimates include unmeasured sampling variation; repeated analysis on AutoSel-120 during development may have shaped analysis choices.

Synthetic-to-benchmark affinity. Supervision is human-specified but not human-labelled and shares the benchmark's authoring specification; cross-family generation, deduplication and low lexical similarity (max TF-IDF cosine 0.21) rule out particular shortcuts but not the shared protocol, and the paraphrase subset is model-selected. A frozen, independently authored test set is the next step.

11 Conclusion

Asked which autonomy level a written policy permits, five of six evaluated LLMs err predominantly upward on AutoSel-120; the pattern survives the artefact controls we could test and is consistent with a model-side over-autonomy tendency under this policy, which the tested deliberation prompts do not reliably correct. Compiling the policy over LLM-extracted three-valued facts makes non-abstaining decisions conditionally sound; with policy-generated supervision an 8B extractor makes 3–4 wrong decisions per 120 at 83–86% coverage. The demonstrated trade is fewer, less agentic errors for coverage lost mainly on no-AI requests, with only a fifth of under-specified requests detected; closing that gap without giving up soundness is the open problem.

Reproducibility Statement

All 2,280 scored completions of the main sweep, 720 contrastive-completeness judgments, three model-annotator runs, classical-baseline folds, the certified post-training corpus, and every post-training evaluation row are archived with model identifiers, prompt hashes, schema hashes, and timestamps; every table and figure in this paper is regenerated offline from those archives by the analysis scripts in the supplementary material; all CAS results use the strict three-valued compiler. Full prompts (Appendix A), the exact output schema (Appendix B), tie-breakers (Appendix C), the authoring specification and dataset statistics (Appendix E), and compute details (Appendix G) are provided. The benchmark, clarification set, prompts, raw outputs, training data, checkpoints' training configurations, and analysis code are provided as anonymous supplementary material and will be released under an open licence upon publication, together with the annotation guide, the anonymised assignment matrix, the aggregation and discussion outcomes, the agreement-analysis code and the label version used for every table; anonymised individual judgments will be released where the annotators' consent covers public release.

Ethics Statement

All benchmark scenarios are fictional and were authored for this study; none is derived from, or describes, any real client, engagement, organisation, or individual. No client, end-user, or sensitive personal data appeared in the benchmark scenarios. The annotation study collected only coarse professional-background categories and annotation judgments under randomised identifiers and with the annotators' consent; participation was voluntary and uncompensated. The taxonomy and benchmark encode a disclosed reference policy for research evaluation, not deployment guidance; organisations should apply their own governance policies. Post-training was performed on open-weight models with synthetic, leakage-screened data; we see no direct dual-use concern beyond those general to instruction-tuned language models.

AI Use Statement

In this work we used generative AI tools to generate synthetic datasets: the 100 LLM-assisted benchmark scenarios were drafted by a Claude-family model against the published authoring specification and reviewed individually by the authors, and the post-training corpus was generated by Qwen3-32B from author-designed blueprints and certified by lexical, deduplication and LLM-certifier gates (Section 6). Five AI research scientists independently assessed the benchmark's reference labels; large language models are the objects of study and serve as cross-family judges for contrastive completeness and as model annotators in a secondary recoverability analysis; these uses are described in the Metrics, Benchmark, and Results sections. Generative AI tools assisted with manuscript editing, analysis-code development, literature discovery, and adversarial review of claims and interpretations. The authors independently verified all code, citations, numerical results, mathematical arguments, and final interpretations. AI tools did not generate reported experimental measurements or participate in practitioner annotation or adjudication. We take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

References

Anthropic. Building effective AI agents. `https://www.anthropic.com/research/building-effective-agents`, 2024a.

Anthropic. Model context protocol. `https://www.anthropic.com/news/model-context-protocol`, 2024b.

Anthropic. Claude model cards (Sonnet and Haiku 4.x series). https://www.anthropic.com/claude, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Matthew DosSantos DiSorbo and Harang Ju. Act or escalate? evaluating escalation behavior in automation with language models. *arXiv preprint arXiv:2604.08588*, 2026.

Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. Levels of autonomy for AI agents. *arXiv preprint arXiv:2506.12469*, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *NeurIPS*, 2017.

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *ICML*, 2017.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *ICLR*, 2024. arXiv:2310.01798.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. *EMNLP*, 2021.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? *ICLR*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. LLMs can't plan, but can help planning in LLM-modulo frameworks. In *ICML*, 2024. arXiv:2402.01817.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research (TMLR)*, 2025.

Stephen Cole Kleene. *Introduction to Metamathematics*. North-Holland, 1952.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, 2nd edition, 2004.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

LangChain. LangGraph: Stateful, multi-actor applications with llms. `https://github.com/langchain-ai/langgraph`, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *TMLR*, 2022.

Xiao Liu, Hao Yu, Hanchen Zhang, et al. AgentBench: Evaluating LLMs as agents. *ICLR*, 2024.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, 2018.

Sameer Malik, Ayush Singh, and Amar Prakash Azad. PolicyGuard: From organizational policies to neuro-symbolic compliance review engines. *arXiv preprint arXiv:2606.32004*, 2026.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Mistral AI. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1, 2025.

Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully autonomous AI agents should not be developed. *arXiv preprint arXiv:2502.02649*, 2025.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of AGI for operationalizing progress on the path to AGI. In *ICML*, 2024. arXiv:2311.02462.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, 2020.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. In *NeurIPS*, 2024.

Rahul Nair, Bastian Lipka, and Elizabeth Daly. Policy-as-logic for robust reasoning over rules. *arXiv preprint arXiv:2608.11905*, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations (ICLR)*, 2025.

OpenAI. Openai agents SDK and agent builder. `https://developers.openai.com/api/docs/guides/agents`, 2024.

OpenAI. gpt-oss-120b and gpt-oss-20b model card. arXiv:2508.10925, 2025.

Nils Palumbo, Sarthak Choudhary, Jihye Choi, Guy Amir, Prasad Chalasani, and Somesh Jha. Formal policy enforcement for real-world agentic systems. *arXiv preprint arXiv:2602.16708*, 2026. v3, May 2026; earlier versions titled "Policy Compiler for Secure Agentic Systems".

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *NeurIPS*, 2024.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics — Part A: Systems and Humans*, 30(3):286–297, 2000.

Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Yujia Qin, Shihao Liang, Yining Ye, et al. ToolLLM: Facilitating large language models to master 16000+ real-world apis. *ICLR*, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3), 2017.

Divyanshu Saxena, Rishikesh Maurya, Xiaoxuan Ou, Gagan Somashekar, Shachee Mishra Gupta, Arun Iyer, Yu Kang, Chetan Bansal, Aditya Akella, and Saravan Rajmohan. Continuous benchmark generation for evaluating enterprise-scale LLM agents. In *International Workshop on Large Language Models for Code (LLM4Code@ICSE)*, 2026. arXiv:2511.10049.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *ACL*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving ai tasks with ChatGPT and its friends in hugging face. *NeurIPS*, 2023.

Thomas B. Sheridan and William L. Verplank. Human and computer control of undersea teleoperators. Technical report, MIT Man-Machine Systems Laboratory, 1978.

Yunxiao Shi, Wujiang Xu, Tingwei Chen, Haoning Shang, Ling Yang, Yunfeng Wan, Zhuo Cao, Xing Zi, Dimitris N. Metaxas, and Min Xu. AgentSelect: Benchmark for narrative query-to-agent recommendation. *arXiv preprint arXiv:2603.03761*, 2026.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*, 2023.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, et al. Beyond human data: Scaling self-training for problem-solving with language models. *TMLR*, 2024. arXiv:2312.06585.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *ICLR*, 2025. arXiv:2409.12183.

Bas C. van Fraassen. Singular terms, truth-value gaps, and free logic. *The Journal of Philosophy*, 63(17):481–495, 1966.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, et al. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024a.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. arXiv:2401.00368, 2024b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *ACL*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 2025. arXiv:2407.18418.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. *NAACL*, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*, 2024.

Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. Agentless: Demystifying LLM-based software engineering agents. In *ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*, 2025.

An Yang et al. Qwen3 technical report. arXiv:2505.09388, 2025.

Kaiyue Yang, Yuyan Bu, Jingwei Yi, Yuchi Wang, Biyu Zhou, Juntao Dai, Songlin Hu, and Yaodong Yang. When lower privileges suffice: Investigating over-privileged tool selection in LLM agents. *arXiv preprint arXiv:2606.20023*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022.

Yirong Zeng et al. The tool-overuse illusion: Why does LLM prefer external tools over internal knowledge? *arXiv preprint arXiv:2604.19749*, 2026.

Shuyan Zhou, Frank F Xu, Hao Zhu, et al. WebArena: A realistic web environment for building autonomous agents. *ICLR*, 2024.

A Full Prompt Templates

System prompt (shared across conditions).

```
You are an enterprise solution-design assistant.  Given a
client request and optional context documents, classify
the appropriate AI autonomy level using the D0--D5 taxonomy
below.

D0 No AI required.  Process redesign, training, or a policy
fix solves the request; AI would mask the root cause.
D1 AI assistant.  LLM drafting/summarising/transforming
accelerates a human-driven task; the human supplies the
inputs and owns the output.  No retrieval corpus, no system
actions.
D2 RAG / copilot.  Grounded retrieval over a designated
corpus is required to answer; the system returns information
or citations and takes no actions.
D3 Deterministic automation.  Fixed, repeatable rules over
structured data fully determine every step; write actions
are allowed only when rule-determined, with no per-item
model judgment.
D4 Supervised agentic workflow.  Multi-step tool use in
which the model exercises judgment, AND a human approves
each write action before it takes effect.
D5 High-autonomy agentic workflow.  Model-judged write
actions execute without per-write human approval; requires
stable APIs, a validated test set, and accepted residual
risk.
```

```
Apply this precedence rule and choose the FIRST level
whose condition fires:  (1) If process redesign or training
solves the request without AI, choose D0.  (2) Else, if
the task requires model judgment to take external write
actions:  choose D4 when each write requires human approval
before taking effect, and D5 when writes execute without
per-write approval.  (3) Else, if fixed deterministic rules
over structured inputs fully solve the task (including
rule-determined writes), choose D3.  (4) Else, if grounded
retrieval over a corpus is required, choose D2.  (5)
Otherwise, choose D1.

Base your decision only on the facts stated in the
request and context documents, not on how ambitious the
stakeholder's phrasing sounds.  Return JSON conforming to
the schema provided.
```

Contrastive instruction (contrastive mode).

```
In addition to selecting a level, for EACH of the five
levels you did NOT choose, produce a rejection record
explaining:  (a) why that level is inappropriate for this
request, with at least one evidence_quote drawn verbatim
from the request or context; (b) what would have to change
in the scenario for that level to become appropriate
(upgrade_conditions); (c) what would have to change for that
level to no longer be a candidate (downgrade_conditions).
Generic rejections (e.g. "too simple", "too risky") are
disallowed.
```

Rationale instruction (deliberation control).

```
In addition to selecting a level, produce a detailed
rationale of at least eight bullet points analysing:
what the workflow does end to end, what the inputs are
and whether they are structured, what knowledge must be
retrieved, what write actions are involved and who approves
them, what the operating constraints and risks are, and why
the chosen level fits.  Discuss ONLY the chosen level; do
not compare against the levels you did not choose.
```

Adjacent-foil instruction. Identical to the contrastive instruction but restricted to the levels immediately adjacent to the chosen level.

Strict annotator prompt. Frames the model as a benchmark annotator, includes the taxonomy, precedence rule, *and* tie-breakers, and instructs mechanical application of the guide. Full text in the supplementary material.

B OUTPUT SCHEMA

```
class RejectedMode(BaseModel):
mode:  Literal[D0,D1,D2,D3,D4,D5]
reason:  str (min_length=10)
evidence_quotes:  list[str]
upgrade_conditions:  list[str]
downgrade_conditions:  list[str]

class Decision(BaseModel):
chosen:  Literal[D0,D1,D2,D3,D4,D5,
insufficient_information]
rationale:  list[str]
```

```
    rejected:  list[RejectedMode]
    confidence:  float in [0,1]
    clarifying_question:  Optional[str]
```

C BOUNDARY TIE-BREAKERS

D1/D2: if the human pre-collects the sources and the system only drafts or summarises, D1; if the system must search or retrieve from a corpus, D2. **D2/D3:** if a deterministic rule over structured fields fully solves the task with no unstructured retrieval, D3; if the system only returns knowledge to a human, D2. **D3/D4:** if every action is fully determined by fixed rules given structured inputs (no per-item model discretion), D3 *even when the system writes*; any model discretion over what or whether to write forces D4 or D5. **D4/D5:** any model-judged external write that can occur without per-write human approval forces D5 when readiness (stable APIs, validated test set, accepted residual risk) is stated, and a policy-level DEFER otherwise (Appendix D); per-write approval preserved keeps the call at D4, even with stable APIs, validated test sets, and signed risk acceptance.

D AUTHORITATIVE POLICY SPECIFICATION

This appendix is the single executable specification behind every CAS number in the paper; it was authored by us from the written policy (taxonomy, precedence rule and tie-breakers) and is not the output of an automatic natural-language-to-policy compiler. The evaluated strict compiler scans the clauses below in order under strong Kleene semantics: a clause fires when all its literals are true, is skipped when any literal is false, and returns DEFER (naming the unknown literals) when some literal is unknown and none is false.

#	Condition (conjunction of literals)	Outcome	Reading
0	decisive_evidence_sufficient = no	DEFER (policy)	stated evidence declared insufficient
1	process_or_training_fix_sufficient = yes	D0	minimum sufficient solution is no AI
2	external_write = yes ∧ model_judgment = yes ∧ approval = yes	D4	supervised agentic
3	external_write = yes ∧ model_judgment = yes ∧ approval = no ∧ stable_apis = yes ∧ validated_test_set = yes ∧ residual_risk_accepted = yes	D5	high-autonomy agentic; all three readiness facts stated
4	external_write = yes ∧ model_judgment = yes ∧ approval = no	DEFER (policy)	unapproved model-judged write without stated readiness: escalate or remediate, not D5
5	structured_inputs = yes ∧ fixed_rules = yes	D3	deterministic automation
6	designated_corpus_retrieval = yes	D2	retrieval required
7	(default)	D1	assistant

Readiness outcomes for an unapproved model-judged write. All three readiness facts stated *yes*: D5 (clause 3). Any readiness fact stated *no*: clause 3 is false and clause 4 fires, a policy-level DEFER (the design is not permitted as requested; the system flags a known unmet requirement rather than asking a question). Any readiness fact *unknown* with none false: clause 3 is unknown, so DEFER naming the unknown readiness facts (a clarifying question). The four situations the output distinguishes are therefore: the stakeholder's *requested* design (visible only in the request text), the *minimum sufficient* permitted level (a clause firing), *missing evidence* (unknown-literal DEFER with named factors) and a *known unmet requirement* (policy-level DEFER from clause 0 or 4).

Factor table. The extractor schema (Pydantic `FactorVector`) has exactly the 13 fields below; 11 enter the decision list and 2 (human_supplies_task_inputs, sources_precollected_by_human) are recorded for the annotation guide's D1/D2 tie-breaker and for clarifying questions but are not used by the compiler. Allowed values follow the extraction prompt (unk. = unknown, n.a. = not_applicable); `not_applicable`, missing and any value outside yes/no are normalised to *unknown* before compilation (in the legacy procedure such values also never satisfied a yes/no test).

Field	Definition (from the extraction prompt)	Values	In list
process_or_training_fix_sufficient	a process redesign, form/policy fix or training would resolve the request without AI	yes/no/unknown	1
human_supplies_task_inputs	a human hands the system its inputs each time	yes/no/unknown	–
designated_corpus_retrieval_required	the system must search a corpus or knowledge store to do the job	yes/no/unknown	6
sources_precollected_by_human	the needed sources are pre-collected/attached by the human	yes/no/unknown	–
structured_inputs_available	inputs are structured records/fields	yes/no/unknown	5
fixed_rules_fully_determine_outcome	fixed, repeatable rules determine every step/outcome	yes/no/unknown	5
external_write_action_exists	the solution takes write actions in external systems	yes/no/unknown	2, 3, 4
model_judgment_controls_what_or_whether_to_write	model judgment decides what or whether to write	yes/no/unknown	2, 3, 4
human_approval_before_every_write	a human approves each write before it takes effect	yes/no/unk./n.a.	2, 3, 4
stable_apis_available	stable APIs/integrations are stated to exist	yes/no/unk./n.a.	3
validated_test_set_exists	a validated test/regression set is stated	yes/no/unk./n.a.	3
residual_risk_formally_accepted	residual-risk sign-off is stated	yes/no/unk./n.a.	3
decisive_evidence_sufficient	the stated text contains enough decisive facts to choose a pattern	yes/no	0

Relation to the direct prompt (disclosed mismatch). The direct-schema baselines received the six level descriptions of Table 1 (which state stable APIs, a validated test set and accepted residual risk as D5 requirements) and the five-clause precedence rule of the Taxonomy section, which assigns D5 to any unapproved model-judged write and has no DEFER outcome (Clarify-24 prompts additionally invite `insufficient_information`). The compiled policy adds clause 4. The two specifications agree on every scenario whose reference label is D5 (all twenty state readiness by construction) and on every other reference label; they differ only on how a system that *wrongly* infers an unapproved model-judged write is scored: a direct model emits D5 (an upward error) where the compiler emits DEFER (a reference-relative deferral). We did not rerun the direct baselines with the gated rule; the archived prompt listing in Appendix A is reproduced unchanged.

Pseudocode.

```
def compile(z):        # z: factor -> yes|no|unknown
  for (lits, outcome) in DECISION_LIST:   # rows 0..7
    st = [lit(z, f, v) for (f, v) in lits]
    if any(s == FALSE for s in st): continue
    if all(s == TRUE for s in st): return outcome
    return DEFER, [f for (f,v),s in zip(lits,st)
                   if s == UNKNOWN]
  return D1
```

E Dataset Statistics and Authoring Spec

The full authoring specification is provided in the supplementary material. Scenarios are versioned YAML files with fields `id`, `domain`, `gold_mode` (internal field name for the policy-derived reference label), `ambiguity`, `request`, `context_documents`, `rationale`. The lexical leakage screen and field/balance validation are reproducible with the released scripts. Per-level composition: each level has exactly 20 scenarios; the 31 high-ambiguity cases sit on the D0/D1, D0/D3, D1/D2, D2/D3, D3/D4, and D4/D5 boundaries, each resolved in its reference rationale by the corresponding Appendix C tie-breaker.

F THE ARCHIVED PILOT

An earlier pilot round reported a 13-scenario pilot with OpenAI `gpt-5.5` (snapshot 2026-04-23 per the API manifest): direct $11/13$ accuracy, contrastive $10/13$, OARR tied at $1/13$, paired sign-test p=1.0—a null result at a sample size that could not detect effects, reported as such. A 7-scenario extension reached $5/7$ (both misses *under*-agentic downgrades to D3), and a same-model relabelling proxy gave κ=0.61 with all disagreements at the original D3/D4 boundary—the observation that motivated the taxonomy revision. Raw outputs remain archived under the pilot prompts and taxonomy, and are not comparable to the current numbers.

G COMPUTE, COST, AND ENVIRONMENT

Main sweep: 2,280 scored completions (six models $\times$ two conditions $\times$ 120 scenarios $=$ 1,440; ablations 480; annotators 360) plus $\sim$720 CC judge calls and 120 embedding calls; wall-clock $\approx$ 80 min with 10-way concurrency; total API cost under USD 25. Experiments: macOS 14, Python 3.13, `openai` and `anthropic` SDKs, `pydantic` 2.x, `scikit-learn`; no GPU. Classical baselines run offline except embedding extraction.

Post-training study: one 8×H100 (80 GB) node; vLLM 0.25 (Kwon et al., 2023) for generation/serving, TRL 1.8 (von Werra et al., 2020) (transformers 5.13, DeepSpeed ZeRO-2 (Rajbhandari et al., 2020)) for SFT/GRPO. Data generation (4,200 blueprint realisations + 4,200 certification passes, Qwen3-32B TP=4) took $\approx$9 minutes; the four SFT runs $\approx$25 minutes combined; GRPO $\approx$40 minutes (60 steps, 8 completions/prompt); each 120-scenario evaluation 1–4 minutes. All training data, configs, reward code, and evaluation rows are provided in the supplementary material.

Post-training hyperparameters. Blueprints are sampled with a fixed seed; realisation and certification use Qwen3-32B served locally (vLLM (Kwon et al., 2023), temperature 0.8 for realisation, 0 for certification). SFT: full-parameter, 2 epochs, completion-only loss, lr 8×10^{-6}–2×10^{-5} by size, bf16, DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) on 8×H100. GRPO: 8 completions per prompt, lr 10^{-6}, β=0, verifier reward only, stopped at 60 steps when the reward saturated (mean 1.32 of 1.35 attainable; 98.8% of prompt groups at zero reward variance). Evaluation of all checkpoints uses the same direct-mode prompt, schema, temperature-0 decoding, and parsing as the main sweep; base-model rows for the same sizes are evaluated identically for paired comparison.

H CONTRASTIVE COMPLETENESS RUBRIC

Per rejected-mode record, score in $\{0, 1, 2\}$: **0** generic or unrelated to the specific scenario, or the evidence quote does not appear in the request/context, or the record argues about the wrong level; **1** specific to the scenario but supported by at most one of (verbatim evidence quote, concrete upgrade condition, concrete downgrade condition); **2** specific, with a verbatim quote and concrete upgrade and downgrade conditions. CC for an output is the mean across its rejection records, normalised to $[0, 1]$. Judge assignment is cross-family (main paper, Metrics section); judge prompts and all judgments are archived. *Judge dependence.* Re-judging a random 100-record subsample of the Sonnet-judged open-weight outputs with gpt-oss-120B (same rubric and inputs) gives within-one-point agreement on 98.8% of the 500 record scores but exact agreement on only 50% and linearly weighted $\kappa = 0.09$: gpt-oss is uniformly more lenient (mean CC 0.94 vs. 0.71; Qwen3-32B 0.99 vs. 0.84, Mistral 0.88 vs. 0.65, Llama 0.96 vs. 0.63). Absolute CC values are therefore judge-dependent and should be read within a judge; the ranking of Qwen3-32B above the other open-weight models is stable, the Mistral–Llama order is not.

I GUARANTEES OF POLICY COMPILATION

Proof of Proposition 1. Let j be the index of the first clause that fires under z^* (or $k{+}1$ for the default). For $i < j$, $c_i(z^*) =$ false, so some literal of c_i is false under z^*; under a partially correct z that literal is false or unknown, hence $c_i(z) \in \{\text{false}, \text{unknown}\}$. If some $c_i(z)$, $i < j$, is unknown, $g_\pi(z) =$ DEFER. Otherwise all $c_i(z)$, $i < j$, are false and evaluation reaches clause j, where $c_j(z^*) =$ true implies every literal is true or unknown under z, so $c_j(z) \in \{\text{true}, \text{unknown}\}$, giving

$\ell_j = g_\pi(z^*)$ or DEFER. For the second claim, if $g_\pi(z) = \ell_j$ then $c_i(z) =$ false for $i < j$ and $c_j(z) =$ true; both are monotone under completion (a false conjunction stays false, a true one stays true), so every completion also fires clause j. □

Necessary versus conservative deferral (supervaluation). Kleene evaluation is sound but not complete: it defers whenever an unknown literal is met before a clause fires, even if every completion of z would yield the same label. The *supervaluational* (van Fraassen, 1966) compiler $g^s_\pi(z)$ decides ℓ iff $g_\pi(z') = \ell$ for every completion z' and defers otherwise; it is also sound under partial correctness (the true assignment is one of the completions), and it can only *remove* deferrals. Enumerating all $2^{|\text{unknown}|}$ completions of every archived strict deferral (zero new model calls) sorts each into exactly one of three categories: a *completion-disagreement deferral* (completions yield different outcomes), a *policy-mandated* DEFER (every completion reaches a policy-level DEFER clause, e.g., an unready model-judged write), or a *conservative deferral* (every completion yields the same non-DEFER level, which supervaluation would decide). For defer-aware Factor-SFT the $17/20/20$ strict deferrals per seed are all completion-disagreement deferrals ($17{+}0{+}0$, $20{+}0{+}0$, $20{+}0{+}0$); the Factor-SFT seeds likewise have $3/16/12$ completion-disagreement deferrals and none of the other kinds; zero-shot CAS (Qwen3-32B) has 13 completion-disagreement and 4 policy-mandated deferrals ($13{+}4{+}0{=}17$; with thinking $9{+}3{+}0{=}12$; gpt-oss-20B extractor $65{+}14{+}0{=}79$). The three categories sum to the strict deferral count for every system, and no system has a conservative deferral, so supervaluation removes nothing here: Kleene's extra conservatism is empty on these data, and every archived deferral is either necessary or mandated by the policy itself. On Clarify-24, supervaluation preserves a deferral only under two further assumptions that partial correctness does not supply: the omitted decisive fact must remain unknown in the extraction (evidence faithfulness) and its admissible completions must yield different policy outcomes; empirically the re-extracted factors defer on only $5/5/6$ of the 24 items, so these assumptions held for at most those cases.

Legacy locked compiler (sensitivity analysis). Before this version, results were compiled with a hand-written procedure (version `pi-v1.1-defer-2026-07-14`) that encodes the same clauses with two policy-level DEFER outcomes. It is *not* covered by Proposition 1: it falls through to D2/D1 when no external write exists and rule-sufficiency is unknown, and it decides when the evidence-sufficiency flag is unknown. Recompiling every archived extraction with the strict compiler changes $1/120$ decisions for zero-shot CAS (Qwen3-32B: DEFER→D2, i.e., $85/17/18 \to 85/18/17$ correct/wrong/deferred), $2/120$ for the gpt-oss-20B extractor ($32/9/79$ unchanged in counts; D0 recall $6 \to 5$), $0/120$ for Factor-SFT (all seeds) and 1–$2/120$ for defer-aware Factor-SFT ($100/2/18 \to 100/3/17$, $96/2/22 \to 96/4/20$, $97/2/21 \to 97/3/20$); on the synthetic validation set the zero-shot extractor moves 1–$2/110$. All CAS results, including Clarify-24, use strict evaluation; the Clarify-24 factor extractions were re-run for this version and compiled with the strict compiler (Table 15); the legacy compiler differs on at most one case per system.

Quote grounding of the zero-shot extractor. Of the zero-shot Qwen3-32B extractor's 1,213 asserted fields on AUTOSEL-120, 59 (5%) carry an evidence quote that is not a verbatim substring of the request or context. Treating those fields as unknown and recompiling adds 14 deferrals and removes 2 errors ($85/18/17 \to 73/16/31$ correct/wrong/deferred). A verbatim quote shows presence, not entailment: the check flags ungrounded assertions but does not gate them in the reported results.

How often is an extraction partially correct? Proposition 1 conditions on partial correctness, which we can measure only where the true factor assignment is known: the 110 held-out synthetic validation scenarios, whose blueprint fixes all 13 factor values. A blueprint value need not be stated in the generated text, so disagreement with it is a proxy for asserted false facts that can over-count (the text may not support the blueprint value) and cannot certify evidence faithfulness; only a human audit of the text can. The Factor-SFT extractor (seed 0, consensus corpus) and its defer-aware variant reproduce all 1,430 cells (100%; in-distribution, since they were trained on the same generator). Zero-shot Qwen3-32B matches 54% of cells with thinking off (50% on), asserts the opposite of a definite blueprint value in 139 (145) cells, and is free of such contradictions on only $30/110$ ($26/110$) scenarios; the compiled decision matches the reference label on $73/110$ ($69/110$) with 16 (23) deferrals. Contradictions concentrate on four factors—whether an external write action exists, whether model judgment controls the write, whether the human supplies the task inputs, and whether

the sources are pre-collected by the human—and no reference-D0 scenario is contradiction-free ($0/21$): the extractor asserts that a process fix would *not* suffice—the same prior that drives direct over-agentic errors. The guarantee therefore binds in practice for trained extractors and describes, rather than prevents, zero-shot failures.

A second policy (experiment). Forbidding autonomous writes (π': every model-judged write needs approval, so reference D5→D4) and recompiling the archived extractions—zero model calls—maps all 20 former-D5 scenarios to D4 for every Factor-SFT seed ($18/20$ zero-shot) with correctness otherwise unchanged. (This π' is a label map, so direct outputs could be remapped too: the experiment demonstrates reuse of frozen extractions, not an adaptation advantage. A policy change that separates requests within one original class would need a factor that varies inside that class; with the frozen 13-factor vocabulary the only such factors are the three readiness facts within D5, which every reference-D5 request states, so no non-trivial intervention is possible without extending the vocabulary and re-extracting.)

Policy-parametricity. Nothing above depends on π beyond its being a decision list over F; replacing π by π' (the second-policy experiment above) changes g_π and leaves E, the archived extractions and Proposition 1 untouched.

J Extended Related Work

System	Task	Factor representation	Reasoning semantics	Unknowns	Guarantee	Policy-gen. supervision
Concept bottleneck / LLM-Modulo (Koh et al., 2020; Kambhampati et al., 2024)	prediction / planning	concepts or candidate plans	learned head / external verifier	not modelled	verifier-side	no
PolicyGuard (Malik et al., 2026)	document compliance review	atom-level extraction questions	typed relational rules, symbolic evaluator	not reported	none claimed	no
Policy-as-logic (Nair et al., 2026)	rule-based QA	extracted predicates	answer-set solver	not reported	none claimed (empirical robustness)	no
Formal policy enforcement (Palumbo et al., 2026)	run-time agent authorisation	event dependency graph	Datalog-style rules, reference monitor	n/a (traces observed)	enforcement by construction	no
CAS (this paper)	design-time autonomy level incl. no-AI	13 three-valued factors	strong-Kleene decision list	first-class (named DEFER)	conditional soundness of non-abstaining decisions	yes (blueprints → SFT/GRPO)

Table 5: Closest extract-then-decide systems, from their published descriptions; none of the prior systems targets design-time autonomy selection or treats unknowns as a first-class output. Extraction followed by symbolic reasoning and three-valued logic are not claimed as inventions here.

Levels of automation. Decades of human–automation research have framed the design question as "which functions to automate, and to what degree" (Sheridan & Verplank, 1978; Parasuraman et al., 2000). Sheridan–Verplank's ten-level scale and Parasuraman–Sheridan–Wickens' four-stage model both target the same problem we revisit. Our D0–D5 scale restates this problem for LLM-era enterprise workflows; we do not claim to invent the function-allocation question itself.

Agent construction, planning, and execution. ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023), and Voyager (Wang et al., 2024a) address how to plan and act after an agent is chosen. AutoGen (Wu et al., 2024) and HuggingGPT (Shen et al., 2023) compose multiple agents. Agent benchmarks such as AgentBench (Liu et al., 2024), SWE-bench (Jimenez et al., 2024), WebArena (Zhou et al., 2024), and ToolBench (Qin et al., 2024) evaluate execution, and enterprise-agent benchmarks can be generated continuously as requirements evolve (Saxena et al., 2026)—but these evaluate agents that already exist.

Autonomy levels and agent selection. Closest to our setting, Feng et al. (2025) define five autonomy levels for AI agents through the user's role (operator, collaborator, consultant, approver, observer)—characterising oversight *within an already-agentic system*—whereas we study

the upstream selection among no-AI, assistance, retrieval, deterministic execution, and supervised/unapproved agentic execution under an explicit organisational policy. AgentSelect (Shi et al., 2026) recommends deployable agent configurations (backbone, toolkit) from a narrative query, *after* an agent-shaped solution space is assumed; our task first determines whether agentic operational authority is warranted at all. Abstention and selective prediction (Geifman & El-Yaniv, 2017; Wen et al., 2025) motivate our treatment of underspecified requests: when decisive facts are missing, the correct output is a clarification, not a level. To our knowledge, prior work has not jointly evaluated policy-conditioned selection across this full no-AI-to-autonomous spectrum.

Contrastive explanations and rationales. LLM-generated free-text explanations (Wiegreffe et al., 2022) and chain-of-thought prompting (Wei et al., 2022) return labels with rationales. (Our "contrastive" condition asks for rejection records against foils and is unrelated to contrastive chain-of-thought demonstrations.) Contrastive explanations (Miller, 2019; Jacovi et al., 2021) go further, requiring justification against *foils*. We adapt this framing to a six-way ordinal label set, requiring the model to defend the chosen mode against each non-chosen mode, and we test whether the foil requirement (as opposed to added deliberation generally) changes decisions.

Calibration, abstention, routing. LLM-confidence calibration (Lin et al., 2022; Kadavath et al., 2022; Guo et al., 2017) and selective prediction (Geifman & El-Yaniv, 2017) inform our confidence and DEFER treatment; OARR instantiates asymmetric error costs (Elkan, 2001). LLM routing (Ong et al., 2025) selects among *models*; we select among solution *patterns*—complementary problems.

K Additional Results

Notes on Table 3. Local base rows come from the 07-15 run (Appendix L reports the 07-12 run; ≤ 2 items differ). The post-training ladder (Table 10) uses the same seeds on the same-family corpus. Clarify-24 results for direct and Label-SFT systems are in Table 15; Clarify-24 results for CAS systems (strict compiler, re-extracted factors) are in Tables 3, 7 and 15. Figure 3 shows the confusion matrix of the best direct model; Table 18 gives the mechanism ablation.

		Clarify-24		AutoSel-120	
Model	Prompt	DEFER (right factor)	turn-2	false DEFER	Acc. %
Qwen3-32B (local)	checklist	2/24 (2)	21/24	0/120	74.2
Qwen3-32B (local)	invited	2/24 (1)	21/24	0/120	72.5
Qwen3-8B (local)	checklist	0/24 (0)	21/24	1/120	61.7
Qwen3-8B (local)	invited	1/24 (1)	21/24	1/120	55.0
Qwen3-32B (endpoint)	checklist	5/24 (3)	20/24	8/120	81.7
Qwen3-32B (endpoint)	invited	1/24 (1)	20/24	0/120	83.3
gpt-oss-120B (endpoint)	checklist	10/24 (7)	22/24	17/120	51.7
gpt-oss-120B (endpoint)	invited	0/24 (0)	22/24	0/120	67.5
Llama-3.3-70B (endpoint)	checklist	2/24 (2)	23/24	3/120	75.0
Llama-3.3-70B (endpoint)	invited	1/24 (1)	22/24	0/120	77.5
Mistral-Small-24B (endpoint)	checklist	6/24 (4)	20/24	9/120	61.7
Mistral-Small-24B (endpoint)	invited	2/24 (1)	21/24	1/120	68.3

Table 6: Behavioural deferral baselines. *Invited*: direct prompt plus "if the stated facts are insufficient to decide, set `chosen` to 'insufficient information' and fill `clarifying_question`". *Checklist*: additionally walk the precedence rule clause by clause, mark each decisive fact stated or missing, and *must* defer if the decisive fact of the first clause that could fire is missing. "Right factor": keyword match of the question against the omitted factor. Schema failures count as errors (invited: Llama 8, Mistral 16, gpt-oss 24; checklist: 7, 21, 32; Qwen ≤ 1). Local rows: vLLM, thinking off; endpoint rows: as-served. Same-session plain direct control (no deferral clause), AutoSel-120 accuracy: Qwen3-32B 83.3, gpt-oss 81.7, Llama 85.8, Mistral 79.2 ($0/0/0/8$ schema failures); its gap to the main sweep is endpoint drift.

Robustness to surface form (full). To test whether these results depend on the scenarios' wording—Claude-assisted drafting, an authoring specification shared with the training corpus—

Model	Prompt	Clarify-24 DEFER (right factor)	Clarify-24 turn-2	AUTOSEL-120 false DEFER	AUTOSEL-120 Acc. %	Threshold op. point / P / R / F1
Qwen3-32B (local)	checklist	–	–	–	–	4/24 vs 3/120 (F1 0.26)
Qwen3-32B (local)	invited	–	–	–	–	7/24 vs 11/120 (F1 0.33)
Qwen3-8B (local)	checklist	–	–	–	–	4/24 vs 7/120 (F1 0.23)
Qwen3-8B (local)	invited	–	–	–	–	1/24 vs 1/120 (F1 0.08)
Qwen3-32B (endpoint)	checklist	–	–	–	–	10/24 vs 19/120 (F1 0.38)
Qwen3-32B (endpoint)	invited	–	–	–	–	7/24 vs 19/120 (F1 0.28)
gpt-oss-120B (endpoint)	checklist	–	–	–	–	10/24 vs 17/120 (F1 0.39)
gpt-oss-120B (endpoint)	invited	–	–	–	–	13/24 vs 26/120 (F1 0.41)
Llama-3.3-70B (endpoint)	checklist	–	–	–	–	19/24 vs 80/120 (F1 0.31)
Llama-3.3-70B (endpoint)	invited	–	–	–	–	12/24 vs 48/120 (F1 0.29)
Mistral-Small-24B (endpoint)	checklist	–	–	–	–	20/24 vs 51/120 (F1 0.42)
Mistral-Small-24B (endpoint)	invited	–	–	–	–	15/24 vs 68/120 (F1 0.28)
CAS structural deferral (strict compiler on both sets)						
CAS zero-shot (Qwen3-32B)	structural	11/24 (3)	19/24	17/120	70.8	0.39 / 0.46 / 0.42
Defer-aware Factor-SFT 8B (s0/s1/s2)	structural	5,5,6/24 (5,4,5)	24,23,24/24	17,20,20/120	83.3/80.0/80.8	F1 0.22/0.20/0.24

Table 7: Supplement to Table 6. Upper block: confidence-threshold deferral, i.e., defer iff verbalised confidence $< \tau$ with τ tuned post hoc for F1 over Clarify-24 (positives) and AUTOSEL-120 (negatives)—oracle operating points, not deployable thresholds; the cell gives Clarify-24 deferrals vs. false deferrals on AUTOSEL-120 at the best-F1 τ. Lower block: CAS structural deferral (strict compiler).

System (AUTOSEL-120, N=120)	Correct	Wrong	DEFER	Coverage %	Sel. error %	Agentic for ref. D0–D3	D5 for ref. D4	D0 recall
Label-SFT s0	100	20	0	100.0	16.7	11	1	6/20
Label-SFT s1	93	27	0	100.0	22.5	17	1	2/20
Label-SFT s2	98	22	0	100.0	18.3	11	1	2/20
Defer-aware Label-SFT s0	100	12	8	93.3	10.7	3	0	5/20
Defer-aware Label-SFT s1	98	12	10	91.7	10.9	6	0	6/20
Defer-aware Label-SFT s2	76	14	30	75.0	15.6	4	1	0/20
Factor-SFT s0	110	7	3	97.5	6.0	1	0	11/20
Factor-SFT s1	89	15	16	86.7	14.4	4	0	2/20
Factor-SFT s2	100	8	12	90.0	7.4	1	0	3/20
Defer-aware Factor-SFT s0	100	3	17	85.8	2.9	1	0	1/20
Defer-aware Factor-SFT s1	96	4	20	83.3	4.0	1	0	1/20
Defer-aware Factor-SFT s2	97	3	20	83.3	3.0	0	0	1/20
CAS zero-shot Qwen3-32B	85	18	17	85.8	17.5	2	0	3/20
CAS zero-shot Qwen3-32B (thinking)	93	15	12	90.0	13.9	3	0	13/20
Haiku 4.5 direct (API)	112	8	0	100.0	6.7	6	0	19/20
Sonnet 4.6 direct (API)	111	9	0	100.0	7.5	7	0	19/20
Qwen3-32B direct (endpoint)	104	16	0	100.0	13.3	6	0	6/20
gpt-oss-120B direct	101	19	0	100.0	15.8	5	4	18/20
Llama-3.3-70B direct	104	16	0	100.0	13.3	10	0	11/20
Mistral-Small-24B direct	101	19	0	100.0	15.8	16	0	4/20

Table 8: Selective-decision view recomputed from raw outputs (strict compiler for CAS rows; reference labels). Coverage = decided share; selective error = wrong among decided; "agentic for ref. D0–D3" = D4/D5 recommended where the reference is D0–D3 (80 requests); "D5 for ref. D4" = autonomy recommended where approval is required (20). Defer-aware Label-SFT's deferrals are `insufficient_information` outputs, valid under the schema of Appendix B; the archived parse flag validates only the level field, so they appear there as unparsed; here they are counted once, as deferrals (correct + wrong + DEFER = 120; no schema failures occurred for these systems). Label-SFT's verbalised confidence takes three values (0.7/0.85/0.95), so the only deterministic operating points are full coverage and keeping confidence ≥ 0.85 ($20/44/42\%$ coverage, $8.3/30.2/20.0\%$ selective error per seed); no threshold reaches 83–86% coverage.

we had every scenario rewritten by Llama-3.3-70B (a family involved in neither) into four registers under a strict fact-preservation instruction, and evaluated on the rewrites (Table 9, Appendix K). Each of the 120 scenarios was rewritten once, into one of four registers assigned round-robin (Slack message, formal RFP, e-mail, meeting notes; 30 each), giving 120 rewritten requests on which every system was evaluated. Two strict-guide model annotators (Qwen3-32B, gpt-oss-120B) labelled the original and the rewrite of every scenario (480 labels); a scenario is *stable* when both annotators give the rewrite the same label as the original, which holds for $77/120$ scenarios (55 of the 240 scenario–annotator pairs, 23%, changed label; 43 scenarios changed for at least one annotator). We report all 120 and the 77 model-annotator-stable scenarios; the subset is model-selected. On the stable subset, as-served LLM accuracy is unchanged within noise and errors remain upward (Mistral $23/0$, Llama $22/1$, Qwen3-32B $22/1$ over/under; gpt-oss-120B $15/10$); the TF-IDF classifier trained on the original wording drops from 90.8% to 79.2% on held-out rewrites and becomes direction-neutral; and the fine-tunes keep their gains (8B SFT $92 \rightarrow 91$, Factor-SFT$\rightarrow g_\pi$ $99 \rightarrow 97$) except the 0.6B model ($86 \rightarrow 75$), which partly memorised surface form. On the model-annotator-stable rewrite subset, the directional pattern persists, suggesting that it is not solely a surface-form effect.

Learnability without an LLM (full). With labelled data: TF-IDF + logistic regression under 5-fold CV reaches 90.8% (3.3% OARR), within 2.5 pp of the best zero-shot LLM. This is neither nearest-neighbour leakage (embedding-kNN 53.3%), nor domain vocabulary (leave-one-domain-out 95.0%), nor cue words (masking every taxonomy/pattern word 91.7%): levels are separable because entry conditions must be stated (top D5 features: *risk, autonomous, signed, stable*); D0 is the least cue-driven level.

	All 120		Model-annotator-stable 77		Paraphrase, all 120		
System	orig. acc	para. acc	orig. acc	para. acc	over/under	DEFER	D0 recall
gpt-oss-120B (as-served)	84.2	79.2	87.0	87.0	15/10	0	14/20
Mistral-Small-24B (as-served)	84.2	80.8	92.2	90.9	23/0	0	4/20
Llama-3.3-70B (as-served)	86.7	80.8	93.5	92.2	22/1	0	7/20
Qwen3-32B (as-served, thinking)	86.7	80.5	96.1	90.8	22/1	0	1/18
Qwen3-0.6B base	25.0	22.9	30.3	31.6	60/31	0	4/19
Qwen3-0.6B +SFT	75.8	64.2	85.7	75.3	38/5	0	5/20
Qwen3-4B base	49.6	45.8	62.3	59.7	59/6	0	0/20
Qwen3-4B +SFT	81.7	74.2	88.3	87.0	29/2	0	4/20
Qwen3-8B base	57.1	53.3	71.1	66.2	54/2	0	1/20
Qwen3-8B +SFT	85.0	82.5	92.2	90.9	21/0	0	5/20
Qwen3-14B base	65.0	56.7	80.5	70.1	50/2	0	0/20
Qwen3-14B +SFT	85.8	84.2	92.2	92.2	17/2	0	4/20
Label-SFT 8B (s0)	83.3	86.7	88.3	94.8	15/1	0	7/20
Factor-SFT 8B$\rightarrow g_\pi$ (s0)	91.7	90.8	98.7	97.4	5/3	3	14/20
Defer-aware Factor-SFT 8B$\rightarrow g_\pi$ (s0)	83.3	80.0	92.2	88.3	3/0	21	2/20
TF-IDF + LogReg (5-fold; trained on the other version's folds)	90.8	79.2	–	–	13/12	0	–

Table 9: Robustness to surface form. Every AUTOSEL-120 scenario was rewritten by Llama-3.3-70B—a family that drafted neither the benchmark nor the training corpus—into one of four registers (Slack message, RFP paragraph, e-mail thread, meeting notes) under a strict fact-preservation instruction; reference labels carry over. Because a rewrite can still blur a decisive fact, we also report the subset on which two strict-guide model annotators (Qwen3-32B, gpt-oss-120B) give the same label before and after rewriting. Local rows: vLLM, thinking off; fine-tunes are seed 42 (ladder) or s0 (8B systems). TF-IDF row: trained on one version's folds, tested on the other version's held-out fold.

Mechanism ablation. On the endpoint, Qwen3-32B accuracy falls monotonically from direct (86.7%) through rationale-only (81.7%) and adjacent-foils (80.8%) to full contrastive (79.2%), with OARR rising in step ($13.3 \rightarrow 16.7 \rightarrow 18.3 \rightarrow 20.0$%): most of the harm appears as soon as the model must deliberate at length, before any foil is requested. For Sonnet 4.6, rationale-only is the best condition (93.3%). Requiring justification against foils is not a free "reasoning bonus"; its sign depends on the model and, as shown next, on serving.

Data scale. A data-scale study (seed 0, matched effective batch) shows Label-SFT and Factor-SFT saturate in-protocol by ∼500–1,000 certified examples (Label: $81 \rightarrow 99 \rightarrow 100$ of 120 at $100/500/1{,}000$; Factor: $73 \rightarrow 100 \rightarrow 101$, with deferral highest at low data—26 DEFERs at $N{=}100$), consistent with the ∼100-label TF-IDF observation.

System	Acc. % ↑	Adj. % ↑	OARR ↓	UARR ↓	Brier ↓
Qwen3-0.6B	25.0	44.0	37.9	37.1	0.432
Qwen3-4B	49.6	63.9	49.6	0.8	0.451
Qwen3-8B	57.1	71.4	41.2	1.7	0.387
Qwen3-14B	65.0	81.7	35.0	0.0	0.270
Qwen3-32B (local)	73.3	80.0	25.8	0.8	0.230
Qwen3-0.6B +SFT	75.8	78.3	22.5	1.7	0.184
Qwen3-4B +SFT	81.7	85.0	17.5	0.8	0.168
Qwen3-8B +SFT	85.0	87.5	14.2	0.8	0.149
Qwen3-14B +SFT	85.8	89.2	13.3	0.8	0.135
Qwen3-8B +GRPO	78.3	80.8	19.2	2.5	0.197
Three seeds (s0/s1/s2), mean±sd; last column = D0 *recall (of 20) instead of Brier*					
Qwen3-0.6B +SFT	76.4±0.8	–	21.1±1.0	2.5±1.2	6.7±0.9
Qwen3-4B +SFT	83.1±1.0	–	14.7±2.1	1.7±0.7	8.0±2.2
Qwen3-8B +SFT	82.2±3.4	–	16.1±3.4	1.1±0.4	5.3±2.1
Qwen3-14B +SFT	84.7±1.6	–	14.2±1.2	1.1±0.4	4.0±0.0

Table 10: Policy-compiled post-training on AUTOSEL-120 (direct mode, $N{=}120$, temperature 0, local vLLM serving, extended thinking disabled). Top block: zero-shot Qwen3 bases. Bottom block: the same models after SFT on the certified corpus, and 8B after GRPO from the instruct base with a verifier-only reward. Metrics in this table are computed over parsed outputs (conditional); primary whole-dataset metrics elsewhere count schema failures as errors. Parse rate is $120/120$ for all rows except the 0.6B base (116), 4B base (119), and 8B base (119).

Worked example (a reference-D0 sales scenario; abridged). "Our Salesforce org has accumulated roughly 15,000 duplicate leads … *We want an AI agent that runs nightly, identifies duplicates, and merges them automatically without anyone having to review each merge.*" *Context:* "Salesforce admin audit: standard duplicate and matching rules have been disabled since a 2023 sandbox refresh …"

Reference rationale. Rule 1 fires: re-enabling the native duplicate rules stops duplicates at the source; the request *over-asks* for an agent.

Direct zero-shot. Sonnet, Haiku, gpt-oss-120B: D0; Qwen3-32B: D3; Llama-3.3-70B, Mistral-Small: D5.

CAS (Qwen3-32B extractor). Extracted: external write yes; model judgment controls writes yes; approval before every write no; stable APIs / validated test set / accepted risk unknown; process-fix-sufficient no. Compiled: DEFER, blocking factors = the three readiness facts (the archived single-factor probe reports an empty blocking set here because the three unknowns must be resolved jointly; the joint probe names all three). The extractor missed the decisive factor, but the compiler refused D5 without stated readiness: the error is attributable to extraction, not to the policy.

Figure 2: A benchmark scenario and the responses it elicits (all from archived records).

Why contrastive prompting backfires (discussion). The contrastive schema obliges the model to write five plausible analyses of levels it did not choose; the archived flips are consistent with self-persuasion rather than self-checking (an interpretation (Huang et al., 2024)): of ten Qwen3-32B correct-to-wrong flips, nine move *upward* (verbalised confidence unchanged in seven), and the direction follows the prompt anchor. Under matched local decoding the deliberation effect disappears (Appendix K), so the mechanism is a hypothesis about *as-served* models; rejection records retain value as a *post-decision* audit trail rather than a decision procedure.

Calibration. Claude-family verbalised confidence exhibits low observed calibration error on this benchmark (ECE_{10} 0.018–0.020, Brier 0.046–0.070; adaptive six-bin ECE 0.045 and 0.098—Haiku emits only five distinct confidence values, so binned ECE is optimistic); open-weight models are over-confident (ECE_{10} 0.05–0.16). Contrastive prompting *worsens* calibration for every open-weight model (e.g., Qwen Brier $0.112 \rightarrow 0.189$), consistent with foil records destabilising rather than checking the decision.

Rejection-record quality (CC). Cross-family judges (Sonnet 4.6 for the four open-weight models, gpt-oss-120B for the two Claude models; one judge per record, so the family ordering below is confounded with judge identity) score all 3,601 rejection records (717 of 720 outputs contain

System	Acc. (k/N) ↑	Adj. % ↑	OARR ↓	UARR ↓	Brier ↓	ECE_{10} ↓	CC ↑
Claude Sonnet 4.6 (direct)	111/120 (92.5)	93.3	8/120 (6.7)	1/120 (0.8)	0.063	0.019	–
Claude Sonnet 4.6 (contr.)	110/120 (91.7)	91.7	9/120 (7.5)	1/120 (0.8)	0.070	0.020	0.98
Claude Haiku 4.5 (direct)	112/120 (93.3)	95.0	8/120 (6.7)	0/120 (0.0)	0.057	0.018	–
Claude Haiku 4.5 (contr.)	114/120 (95.0)	97.5	6/120 (5.0)	0/120 (0.0)	0.046	0.018	0.96
Qwen3-32B (direct)	104/120 (86.7)	89.2	16/120 (13.3)	0/120 (0.0)	0.112	0.076	–
Qwen3-32B (contr.)	95/120 (79.2)	84.2	24/120 (20.0)	1/120 (0.8)	0.189	0.160	0.85
gpt-oss-120B (direct)	101/120 (84.2)	87.5	10/120 (8.3)	9/120 (7.5)	0.138	0.097	–
gpt-oss-120B (contr.)	100/120 (83.3)	84.2	9/120 (7.5)	11/120 (9.2)	0.157	0.133	0.92
Llama-3.3-70B (direct)	104/120 (86.7)	90.8	14/120 (11.7)	2/120 (1.7)	0.108	0.052	–
Llama-3.3-70B (contr.)	95/120 (79.2)	83.3	23/120 (19.2)	2/120 (1.7)	0.157	0.093	0.60
Mistral-Small-3.1-24B (direct)	101/120 (84.2)	85.8	19/120 (15.8)	0/120 (0.0)	0.138	0.105	–
Mistral-Small-3.1-24B (contr.)	96/120 (80.0)	83.3	24/120 (20.0)	0/120 (0.0)	0.180	0.157	0.64
TF-IDF + LogReg (supervised, 5-fold)	90.8	95.8	3.3	5.8	–	–	–
e5-Mistral emb. + kNN (supervised, 5-fold)	53.3	72.5	30.8	15.8	–	–	–

Table 11: Full main sweep: direct and contrastive conditions for all six models (referenced from Table 2). CC is contrastive completeness (cross-family LLM judge; contrastive condition only; absolute values are judge-dependent, Appendix H).

System	Output	Trained on	Can defer?	Notes
Direct (zero-shot / Label-SFT)	level	— / reference JSON levels	invited on Clarify-24	never used it
CAS zero-shot	factors → g_π	—	structural	extractor prompted with the factor schema
Factor-SFT→g_π	factors → g_π	by-construction factor targets	structural	3 seeds
Defer-aware Label-SFT	level	levels + 286 underspecified (unknown)	learned	3 seeds
Defer-aware Factor-SFT→g_π	factors → g_π	factors + 286 underspecified	structural + learned	3 seeds; headline extractor

Table 12: Systems compared. "Structural" deferral is a compiler-produced DEFER, either a completion-disagreement deferral or a policy-mandated DEFER (Appendix I); it arises from unknown factors in the compiler; "learned" deferral is a behaviour the model must emit.

exactly five; two duplicate one mode, one omits one—$3{,}600+2-1$, audited). Mean CC separates the families: Sonnet 4.6 0.98, Haiku 4.5 0.96, gpt-oss-120B 0.92, Qwen3-32B 0.85, Mistral-Small 0.64, Llama-3.3-70B 0.60 (Table 2). The models whose decisions survive the contrastive obligation are also those producing grounded, condition-complete records; we treat this CC–robustness association as exploratory, since judge identity is confounded with model family and judge agreement is low (Appendix H).

Model recoverability of the benchmark labels. Under the strict annotation-guide prompt (Table 16, Appendix L), the three cross-family annotators recover the reference labels at $\kappa{=}0.90$ (Sonnet 4.6), 0.81 (Qwen3-32B), and 0.74 (gpt-oss-120B)—substantial to almost-perfect agreement (Landis & Koch, 1977)—with pairwise inter-annotator κ of 0.65–0.71. Where a 2-of-3 majority exists ($108/120$), it matches the reference label on 106 (98.1%), and both exceptions are isolated single-scenario disagreements. We read this as evidence that the published guide pins down the labels, while reiterating that model agreement is not practitioner agreement. Notably, the earlier pilot's systematic D3/D4 annotator disagreements (which motivated the rule revision) do not recur under the revised taxonomy.

Robustness across subsets. On author-written vs. LLM-assisted scenarios, direct accuracy moves in opposite directions by family (table in Appendix L)—accuracy does not move uniformly on this small subset, and the Claude-advantage-on-Claude-drafted-text pattern is visible and disclosed. High-ambiguity boundary scenarios are *not* uniformly harder; errors concentrate in mid-ambiguity scenarios where over-asking framing must be resisted.

Mechanism, under matched serving. Rerunning the prompting conditions locally (Table 19, Appendix L) with matched decoding and token accounting. The endpoint-observed contrastive harm is *serving-dependent*: under local non-thinking decoding, Qwen3-32B loses only 3 answers to full contrastive prompting and none to token-matched deliberation, while gpt-oss-20B shows the graded pattern ($91{\to}84{\to}77$, but 22 of the contrastive outputs are schema-parse failures counted as errors; among parsed outputs contrastive is $77/98$ vs. direct $91/118$, i.e., no decision-level degradation). Choose-freeze-then-audit preserves the direct decision by construction on every model while still

		Accuracy delta (contr. − direct)			OARR delta (contr. − direct)		
Model	N	Δ pp	95% CI (pp)	McNemar p	Δ pp	95% CI (pp)	McNemar p
Claude Sonnet 4.6	120	-0.8	[-4.2, 1.7]	1.0	+0.8	[-1.7, 4.2]	1.0
Claude Haiku 4.5	120	+1.7	[-2.5, 5.8]	.6875	-1.7	[-5.8, 2.5]	.6875
Qwen3-32B	120	-7.5	[-13.3, -2.5]	.0117	+6.7	[1.7, 11.7]	.0215
gpt-oss-120B	120	-0.8	[-7.5, 6.7]	1.0	-0.8	[-6.7, 4.2]	1.0
Llama-3.3-70B	120	-7.5	[-13.3, -2.5]	.0225	+7.5	[2.5, 13.3]	.0225
Mistral-Small-3.1-24B	120	-4.2	[-7.5, -0.8]	.0625	+4.2	[0.8, 7.5]	.0625

Table 13: Paired direct-vs-contrastive comparison per model (10,000 paired-bootstrap resamples; exact McNemar on discordant pairs; raw p shown). Negative accuracy deltas and positive OARR deltas mean contrastive prompting *hurts*. A sign-flip permutation test confirms a global condition effect (p=.004; clustered by scenario p=.004); no single model remains significant after Holm correction (smallest corrected p=.070; the design detects ≈10 pp deltas at 80% power).

producing complete rejection records—supporting the reading of rejection records as post-decision governance artefacts rather than a decision procedure.

L Extended Result Tables

Statistic (independent first-round labels; 5 annotators, 3 per scenario)	Value
First-round judgments	360
Unanimous three-of-three (on a level / on insufficient information)	103 (102 / 1)
Two-of-three majority	12
Three-way split, resolved by five-person discussion	5 (three to a level, two as under-specified)
Krippendorff's α, nominal, all categories [95% CI, 10,000 scenario bootstraps]	0.87 [0.81, 0.93]
Krippendorff's α, nominal, levels only	0.88 [0.83, 0.93]
Pairwise exact agreement, mean over 10 pairs (36 shared scenarios each)	0.89
Pairwise Cohen's κ, mean (min–max)	0.87 (0.77–0.93)
Unanimity by reference level D0/D1/D2/D3/D4/D5	12/19/20/16/16/19 of 20
Unanimity by ambiguity tier low/medium/high	36/39, 46/50, 21/31
Consensus = reference / differs / judged under-specified	115 / 2 / 3

Table 14: Practitioner agreement on AutoSel-120. Consensus = two-of-three majority or discussion outcome. The two differing scenarios: reference D5 with consensus D3 (labels D3, D3, D5), and reference D1 resolved to D2 after discussion (labels D1, D2, insufficient information). The three judged under-specified: one unanimous "insufficient information" (reference D0) and two splits resolved as under-specified (both reference D4). Reference labels are retained for evaluation; Section 8 (Q2) reports the directional result under the practitioner consensus.

System	Defer /24	Right factor	Turn-2 /24
Direct zero-shot (8B)	0/24	0	21/24
Direct zero-shot (32B)	0/24	0	21/24
Label-SFT s0	0/24	0	22/24
Label-SFT s1	0/24	0	22/24
Label-SFT s2	0/24	0	21/24
Factor-SFT s0	2/24	2	22/24
Factor-SFT s1	5/24	0	23/24
Factor-SFT s2	4/24	1	23/24
CAS zero-shot (32B, thinking off)	11/24	3	19/24
CAS zero-shot (32B, thinking on)	10/24	3	17/24
Defer-aware Label-SFT s0	5/24	0	22/24
Defer-aware Label-SFT s1	5/24	0	23/24
Defer-aware Label-SFT s2	7/24	0	24/24
Defer-aware Factor-SFT s0	5/24	5	24/24
Defer-aware Factor-SFT s1	5/24	4	23/24
Defer-aware Factor-SFT s2	6/24	5	24/24

Table 15: Clarify-24. CAS rows use the strict compiler on factor extractions re-run for this version (Appendix I). Each author-written scenario omits exactly one decisive fact; columns give first-turn deferral recall, whether the deferral identifies the omitted factor, and accuracy after an explicit oracle clarification.

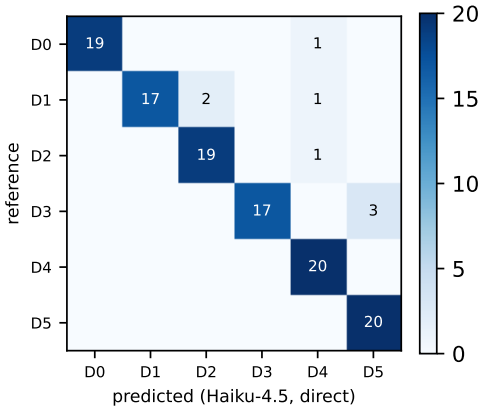

Figure 3: Confusion matrix for the best direct-mode model (Claude Haiku 4.5). Residual errors sit above the diagonal (over-agentic).

Pair	Exact %	Adj. %	κ
Sonnet 4.6 vs. reference	91.7	91.7	0.90
gpt-oss-120B vs. reference	78.3	83.3	0.74
Qwen3-32B vs. reference	84.2	86.7	0.81
Sonnet 4.6 vs. gpt-oss-120B	75.0	–	0.70
Sonnet 4.6 vs. Qwen3-32B	75.8	–	0.71
gpt-oss-120B vs. Qwen3-32B	70.8	–	0.65
Majority vote vs. reference	98.1	98.1	–

Table 16: Model recoverability of the policy-derived reference labels: strict-guide model annotators vs. the reference labels and vs. each other (N=120). Majority vote is over the three annotators (108/120 scenarios have a 2-of-3 majority).

Model (direct)	Author-20	LLM-assisted-100	Low	Med	High
Sonnet 4.6	80.0	95.0	92.3	88.0	100.0
Haiku 4.5	90.0	94.0	97.4	90.0	93.5
Qwen3-32B	90.0	86.0	87.2	88.0	83.9
gpt-oss-120B	85.0	84.0	87.2	84.0	80.6
Llama-3.3-70B	95.0	85.0	82.1	88.0	90.3
Mistral-S-24B	95.0	82.0	84.6	82.0	87.1

Table 17: Direct-mode accuracy (%) by subset: author-written vs. LLM-assisted scenarios, and by ambiguity tier.

Condition	Acc. (k/N)	OARR	UARR	Brier
Claude Sonnet 4.6				
direct	111/120 (92.5)	8/120	1/120	0.063
+ rationale (no foils)	112/120 (93.3)	8/120	0/120	0.056
+ adjacent foils	110/120 (91.7)	9/120	1/120	0.070
+ all foils (contrastive)	110/120 (91.7)	9/120	1/120	0.070
Qwen3-32B				
direct	104/120 (86.7)	16/120	0/120	0.112
+ rationale (no foils)	98/120 (81.7)	20/120	2/120	0.158
+ adjacent foils	97/120 (80.8)	22/120	1/120	0.168
+ all foils (contrastive)	95/120 (79.2)	24/120	1/120	0.189

Table 18: Mechanism ablation (N=120 each). For Qwen3-32B the degradation is monotone in added reasoning obligations: deliberation alone (rationale, no foils) already hurts, adjacent foils hurt more, full foils most. For Sonnet 4.6 the rationale condition is mildly beneficial and foils mildly harmful.

Model	Condition	Acc.	OARR	Parse	Tok
Qwen3-32B	direct	89/120	30	120	715
Qwen3-32B	rationale (matched)	90/120	30	120	1082
Qwen3-32B	contrastive	86/120	33	120	896
Qwen3-32B	freeze-then-audit	89/120	30	120	912
gpt-oss-20B	direct	91/120	24	118	661
gpt-oss-20B	rationale (matched)	84/120	29	114	1676
gpt-oss-20B	contrastive	77/120	19	98	1915
gpt-oss-20B	freeze-then-audit	91/120	24	118	2745
Qwen3-8B	direct	66/120	51	120	730
Qwen3-8B	rationale (matched)	62/120	55	120	1347
Qwen3-8B	contrastive	71/120	44	117	809
Qwen3-8B	freeze-then-audit	66/120	51	120	775

Table 19: Mechanism conditions under matched local serving (AUTOSEL-120): direct; length-matched rationale (deliberation control whose target length matched the contrastive output; realised mean tokens in the Tok column differ, so the match is approximate; length comparable to contrastive); full pre-decision contrastive; choose-freeze-then-audit (rejection records generated *after* the decision is locked—accuracy equals direct by construction). Tok = mean output tokens.