GOVERNED AUTONOMY

Human–LLM Capability Elicitation: Studying Validated Joint Performance

A paper in the Governed Autonomy collection

Andrew Durden
Independent research
Publication edition 2026.10.8

Research on authority, evidence, and continuity in autonomous systems.

Content identifier: c321294d8c141120daaaf471

All rights reserved · Research exposition

Contents

Human–LLM Capability Elicitation: Studying Validated Joint Performance 3

Abstract . . . 3

1. The research object is a work episode . . . 3

2. Four different ways performance can improve . . . 4

3. Intent fidelity and human authority . . . 6

4. Representation can reveal structure and introduce error . . . 6

5. Search, independence, and useful depth . . . 7

6. External validation is claim-specific . . . 8

7. Worked conceptual examples . . . 8

8. What would establish an interaction effect? . . . 9

9. Relationship to governed autonomy and limits on automation . . . 10

Source guide . . . 11

All rights reserved · Research exposition

Human–LLM Capability Elicitation: Studying Validated Joint Performance

Governed Autonomy · Publication 2026.10.8
Research study: conceptual exposition of the Human–LLM Capability Elicitation research corpus
Source: Public exposition, successor 2026.10.8; historical research and evidence separately qualified

Abstract

A language model may produce different results when a task is expressed differently, when relevant information is supplied, when tools become available, or when a qualified person helps evaluate the answer. These improvements have different causes. Calling all of them “elicitation” obscures what changed and makes it difficult to reproduce a result or determine where it applies.

The Human–LLM Capability Elicitation corpus proposes the bounded work episode as a research unit: people, models, task, evidence, representations, interaction, resources, authority, and validation considered together. It distinguishes changes that make a fixed model's competence more accessible from improvements obtained by adding information, computation, tools, or human expertise. It treats intent as something that may need clarification or formation, and useful depth as a quality profile constrained by correctness, cost, and risk. Five scenarios show how these distinctions affect the attribution and validation of improved performance. The contribution is a provisional research framework with testable alternatives. Its proposed protocols and experiments have yet to establish a general capability improvement.

1. The research object is a work episode

A prompt and an answer are useful units for some questions. They are often too narrow for work in which the objective is clarified, evidence changes, several candidates are considered, and someone decides whether the result can be used. A correct-looking final answer can conceal that the task changed halfway through. An unsuccessful first response can conceal that a simple representation change would have made the problem tractable.

The corpus therefore studies a **human–LLM epistemic work episode**: a bounded attempt to formulate, investigate, compare, and validate claims about a task. “Epistemic” here concerns how the participants distinguish what is known, assumed, inferred, or still uncertain. The episode includes the people and model, the information available to them, their interaction, the resources used, and the standard by which a result can be accepted.

This wider unit makes causal claims more demanding. If an expert supplies a missing fact and the model then succeeds, the episode improved. That does not demonstrate an effect of wording alone. If a program checks thousands of cases, the resulting capability belongs to a system with that computational support. If a specialist catches an error, their contribution belongs in the account of performance.

All rights reserved · Research exposition

The conceptual contribution is an integrated research vocabulary for these conditions. It brings established questions about intent, representation, search, evidence, human judgment, and validation into the same analysis. The usefulness of considering them together is itself an empirical question. Public review scope.

2. Four different ways performance can improve

The corpus distinguishes **intrinsic elicitation** from several forms of augmentation. In the narrow elicitation case, the model, relevant information, tools, and resource budget are held fixed while the interaction changes. A clearer representation or a better-suited operation may then improve validated performance. The causal claim concerns that controlled contrast.

Contextual augmentation adds information: a source passage, a missing observation, a demonstration, or expert knowledge. **Computational scaffolding** adds computational or action capacity, such as tools or additional search. **Joint-system augmentation** adds human expertise, independent adjudication, experiments, or institutional support. These are all legitimate ways to improve an outcome. Naming them accurately makes their benefits and costs assessable.

The categories can interact. A useful tool may contribute little if the model cannot represent the problem correctly. A clearer representation may make missing evidence visible. An expert may improve both the specification and the evaluation. Real episodes therefore need not fit a single category; they need an account that does not attribute the whole gain to whichever component is most noticeable.

This also changes how failure is interpreted. A weak response can reflect inadequate model competence, inaccessible competence under the tested interaction, missing evidence, a mistaken objective, a poor representation, insufficient search, or an inadequate evaluator. One unsuccessful prompt cannot distinguish these possibilities. Conversely, one successful interaction does not show that every absent capability is waiting to be unlocked.

The corpus uses **accessible capability** for a fixed model's attainable expected validated performance over a declared feasible class of interaction procedures and resource budgets. Because performance has several dimensions, its default object is a frontier of quality profiles, rather than one winning answer or an unqualified scalar maximum. An observed profile describes recorded results under tested conditions. An estimated profile concerns expected performance across a specified population and repeated trials, under stated assumptions. The two must remain separately reported.

An assessment declares the task population; model and version; information, tool, and human contributions; admissible interaction procedures; token, call, time, financial, and human-attention constraints; risk limits; validator; and acceptance criteria. A change to those boundaries changes the quantity being assessed. For a closed task, correctness, fidelity to the authorized task, and validation integrity are gates. For an open research task, the evaluated object may be a candidate's precision and testability, while the candidate's truth remains unestablished.

All rights reserved · Research exposition

Best then has a specified meaning. In the default frontier interpretation, one eligible quality profile dominates another only when it is at least as good on every declared primary dimension and better on at least one, within the compared resource constraints. Profiles that trade quality dimensions remain incomparable by that relation. A study can use a single decision value instead, but its objective, weights, constraints, and sensitivity analysis must be declared before observing the results. A higher novelty score cannot compensate for failing a required correctness gate.

Quantity	What is being reported	Required interpretation
Observed validated result	A particular trial or recorded collection of trials met its stated validator	Establishes performance in those cases; one favorable trial is not the expected performance of the procedure.
Observed tested frontier	Nondominated profiles among the conditions actually evaluated	Describes the best profiles found in the tested set, with measurement uncertainty retained. It need not include the best feasible untested procedure.
Estimated expected profile	A declared average, probability of satisfying all required gates, or other predeclared outcome across a task-and-trial population	Requires the estimator, sampling and dependence assumptions, uncertainty, and any inference beyond observed conditions.
Estimated accessible frontier	An estimate of attainable expected-quality profiles over the declared feasible procedure class	Requires evidence about search coverage and uncertainty in the expected profiles; broader procedure classes or budgets change the frontier.
Represented capability	A latent claim about task behavior supported by the model's learned structure	Cannot be read directly from an answer or identified as the absolute ceiling from a finite tested frontier.

For example, suppose one admissible procedure offers higher validated reliability at greater review cost, while another offers lower cost and more explanatory detail. If both satisfy the mandatory gates, neither is automatically the best. They can occupy different positions on the quality–cost frontier. A separately declared cost limit may exclude the first, or a named decision objective may prefer one. Reporting the highest reliability from the first together with the lowest cost from the second as one attainable profile would invent a result that neither procedure achieved.

Finite trials also create selection uncertainty. Choosing the apparent winner after evaluating many noisy candidates can overstate its expected performance. Repetition, fresh tasks, and separation of selection from evaluation help test that inference. In a procedure class without suitable attainment conditions, even a well-defined upper bound need not be achieved by any available procedure. The paper's claims therefore concern observed or estimated bounded performance and explicitly declared optimization, rather than the model's absolute ceiling. Public review scope.

All rights reserved · Research exposition

3. Intent fidelity and human authority

People do not always begin with a complete specification. They may have a clear goal but lack the vocabulary to express it. They may instead have examples, exclusions, competing aims, and an intuition that the usual framing misses something. Interaction can recover an existing intention, clarify alternatives, or help form a more definite objective.

Intent fidelity concerns whether the task pursued and result delivered satisfy the authorized objective, constraints, and prohibited consequences. It should not be inferred merely from fluent agreement or from a user's assent after seeing a persuasive answer. A technically excellent solution to a neighboring problem can still fail the intended task.

The source distinguishes clarification from substitution. A person may deliberately revise an objective after understanding its consequences. That is different from a model silently replacing an unusual request with a more familiar one. The original expression and the reasons for material changes help make the difference inspectable.

Human authority also has a defined scope. People and accountable institutions authorize objectives, values, resource commitments, risk tolerances, and consequential actions. Their factual or mathematical assertions remain open to evidence and valid reasoning. Respecting a person's objective does not require accepting an incorrect premise; challenging a premise does not authorize changing the objective.

The corpus preserves **unresolved residue**: an observation, intuition, or concern that has not yet been satisfactorily formalized. Such material can guide inquiry without becoming a premise. A sense that “the boundary is wrong” may later prove useful, or it may prove misleading. Keeping it available does not assign it the status of evidence.

Clarification itself has costs. Excessive questioning can transfer work to the person, interrupt progress, or bias their preferences toward options suggested by the model. The source treats the value of clarification as a research question dependent on ambiguity, task importance, and human burden. It does not prescribe an elaborate negotiation for every simple request. Public review scope.

4. Representation can reveal structure and introduce error

A problem representation determines which relationships are easy to inspect. A narrative can preserve context and exceptions. A timeline can make ordering visible. A dependency view can expose prerequisites. A constraint account can reveal that a proposed solution is infeasible. None is universally superior.

Changing representation can redistribute difficulty without adding facts. A relationship that is buried across several paragraphs may become obvious when expressed in a different form. This is one reason representation is a plausible causal variable in capability research. It is also why a representation change needs scrutiny: the transformation may add an unstated assumption, omit a constraint, or imply causality that the evidence does not support.

Several representations may reveal complementary defects. They may also repeat the same omission or contradict one another. An extra diagram is not evidence of deeper

All rights reserved · Research exposition

understanding. The relevant question is whether the representation exposes task-relevant structure, preserves meaning, and improves a result that survives validation.

The same distinction applies to reasoning operations. Asking for a counterexample, an explanation, a proof, or a comparison involves different work. An operation is meaningful when its output can be evaluated against its purpose. Merely naming an impressive reasoning style does not establish that a distinct internal faculty was activated.

Visible rationales are artifacts of the interaction. They can help a person inspect assumptions or check steps, but they are not privileged access to the model's internal computation. A convincing explanation of how an answer was reached may not faithfully report the process that produced it. The source therefore bases its research claims on observable changes and checkable artifacts rather than inferred internal mental modules. Public review scope.

5. Search, independence, and useful depth

Difficult tasks may benefit from considering several plausible explanations or solution families before selecting one. Early commitment can shape later reasoning, making alternatives appear only as variations on the first answer. Yet producing more candidates also increases the number of persuasive errors that must be assessed.

The corpus treats breadth, depth, and diversity as distinct. Many differently worded answers can share the same unsupported assumption. A deeply developed branch can be wrong at its first premise. A diverse collection can include no useful candidate. Search quality therefore depends on the relationship between generation, comparison, validation, and the available budget.

Generation proposes candidates. **Adjudication** compares them against criteria and evidence. **Validation** applies a method capable of changing their claim status. These functions may be separated, but labels alone do not make their errors independent. Calling the same model a critic after it has produced an answer may detect some defects without creating an independent test.

Meaningful independence concerns the sources of error. Different evidence or a different method may catch a shared misconception. A formal checker can reject an invalid formal artifact. A new observation can contradict a plausible explanation. Another model can add a useful perspective, but it may share training influences, familiar assumptions, or evaluation weaknesses. Independence is a matter of degree and kind.

The corpus also resists a single scalar measure of “maximum useful depth.” Correctness, fidelity to intent, explanation, novelty, robustness, calibration, reproducibility, and cost can trade off. A concise checked proof may serve one objective better than an extended exploration. A research inquiry may benefit from explicit alternatives and tests even when no final explanation has been established.

Depth is useful when it exposes supported mechanisms, dependencies, constraints, or invariants that improve understanding or action. Length, technical vocabulary, and confidence are unreliable proxies. An intricate false account remains false, and an untested hypothesis remains a hypothesis however much structure surrounds it. Public review scope.

All rights reserved · Research exposition

6. External validation is claim-specific

The source distinguishes novelty within an episode from a discovery supported by an appropriate external standard. A model can generate a new combination, conjecture, design, or mechanism candidate. Whether that contribution is true, new to the relevant field, and useful requires separate evidence.

Different claims require different validators. A formal proof can establish a proposition under its stated assumptions, while a separate check is needed to ensure that the formal proposition matches the intended question. Software execution can establish behavior in the tested environment, while broader claims about reliability or security require broader evidence. Empirical mechanisms need observations and tests that distinguish them from alternatives.

Legal analysis depends on current primary authority, jurisdiction, facts, and procedural context. Biomedical mechanism proposals require layered scientific evidence and qualified review; plausibility in a laboratory model does not establish patient benefit or safety. Institutional design requires attention to incentives, enforcement, distributional effects, and behavior over time. A coherent rule is not evidence that an institution will enact it successfully.

Validation is consequently neither a universal ladder nor a stamp attached to an entire answer. An episode may contain a checked factual claim, an untested causal hypothesis, and a normative preference that cannot be settled by the same method. Preserving these differences prevents a narrow successful check from lending unwarranted authority to the rest of the output.

The evaluator is part of this system. People can be influenced by fluency, confidence, familiar explanations, and source prestige. Experts can disagree or share blind spots. A model judge can reward stylistic cues or its own preferred formulations. Better presentation may increase trust without increasing correctness. Evidence of improved judgment must therefore extend beyond preference for the answer's appearance. Public review scope.

7. Worked conceptual examples

Five hypothetical scenarios examine the attribution, validation, and limits of improved performance.

Illustration 1 — a clearer representation succeeds. A planning task describes several dependencies in prose. The model proposes an order that violates one of them. The same information is then presented in a compact dependency view, and the resulting plan satisfies the stated constraints when checked. This episode demonstrates a successful result under the second condition. To claim a general elicitation effect, a study would still need repeated comparable tasks, credible alternatives, and resource accounting. One useful correction does not establish a universal advantage for dependency views.

Illustration 2 — missing evidence limits a diagnosis. A service behaves unexpectedly, but the supplied record lacks the relevant time window. Several explanations remain compatible with what is available. Asking the model to reason more deeply does not create the absent observations. An authorized evidence source later supplies the missing interval,

All rights reserved · Research exposition

allowing a hypothesis to be tested. The improvement should be attributed at least partly to new information. Presenting it as a pure prompt improvement would obscure the actual cause of success.

Illustration 3 — authority remains with the objective owner. A user asks for a comparison of possible policy changes and prohibits any live change. The model produces a strong candidate and estimates that it would improve performance. Those conclusions do not authorize implementation. The comparison can be completed, its evidence assessed, and its limitations reported within the task. Carrying out the candidate would require authority beyond the analytical objective. This illustrates an action boundary, not a claim about the quality of the candidate itself.

Illustration 4 — recovery after a contradicted premise. An investigation initially assumes that two records refer to the same object. Later evidence shows that they refer to different objects. Correcting the final sentence would leave the dependent analysis unreliable. The episode must revisit the claims that depended on the mistaken identity, preserve unaffected evidence, and test the revised account. Recovery succeeds when the resulting claims fit the authorized task and available evidence, not when the earlier explanation has merely been rewritten more persuasively.

Illustration 5 — a useful analogy remains a hypothesis. A researcher notices that a resource problem resembles a process from another field. The resemblance suggests a candidate mechanism and an observation that would distinguish it from a simpler local explanation. Until the target is tested, the analogy has helped generate a question; it has not established that the mechanism transfers. A negative result can still be useful by eliminating an attractive but misleading account.

These scenarios emphasize accurate attribution. A better outcome may reflect representation, evidence, correction, a useful tool, or human judgment. Identifying the contribution is part of understanding whether the improvement is reproducible.

8. What would establish an interaction effect?

A causal study needs a declared population of tasks and users, a specified change, a credible comparator, and an outcome tied to an appropriate validator. It should distinguish the effect of a complete interaction arrangement from the effects of its individual components. Success after a bundle of changes does not show that every component helped.

Resource accounting matters because apparently better interaction can purchase more computation or human attention. A method that improves correctness while greatly increasing review cost may still be useful for some tasks, but its result should be stated that way. Quality at equal resources and resources required for equal quality answer different questions.

Task sampling also constrains inference. Repeating one puzzle many times gives information about variation on that puzzle. It does not establish broad competence. Synthetic tasks provide known answers but can favor the structure chosen by their designer. Realistic open tasks can be valuable while lacking decisive outcomes. A research program needs to preserve the strengths and limitations of each evidence source.

All rights reserved · Research exposition

Reproducibility is not usually identical wording. The relevant target may be the same validated claim, a materially equivalent artifact, a comparable effect across repeated trials, or successful transport to new tasks and participants. These are different achievements. Model updates and opaque runtime changes can limit the precision with which earlier conditions can be recreated.

The source's proposals must also be capable of losing. Representation-first interaction may add no value over a concise expert specification. Extra branches may create noise. An explicit claim record may consume more attention than it saves. A simpler method may achieve the same validated result at lower cost. Such outcomes would narrow or weaken the framework rather than count as insufficiently elaborate applications of it. Public review scope.

9. Relationship to governed autonomy and limits on automation

This corpus has a methodological relationship to the rest of the collection. It helps distinguish proposed behavior from tested behavior, a plausible explanation from a validated mechanism, and a model-generated recommendation from authority to act. These are conceptual correspondences with the thesis, architecture, and evidence guide. The source does not establish an implementation contract with the other systems discussed here.

The baseline is a research framework. Its manual protocols organize experiments rather than prescribe a universal process for every task. Adopting a protocol still requires an account of its suitability, costs, and authority for the intended use.

Automation can preserve a useful procedure, but it can also freeze a provisional theory, hide judgment calls, and make procedural compliance look like capability. The source therefore asks for evidence that a proposed automated analytical function improves validated outcomes, retains meaningful correction, and remains appropriate after conditions change. It distinguishes those claims from clerical transformation of already approved material, which does not itself claim a capability improvement.

The source's references provide adjacent foundations and empirical precedents, rather than independent validation of the complete framework. Its worked examples are conceptual or synthetic, and its study plans do not report completed experiments. Claims about the framework's effectiveness therefore remain hypotheses.

Open questions include how to measure meaningful branch independence, how to recognize a defective representation before ground truth is known, how to preserve tacit expertise without distorting it, and which laboratory effects survive consequential work. The practical promise is a more precise account of why an interaction helps and when it stops helping. Realizing that promise requires validated outcomes, preserved negative evidence, and conclusions bounded by what was actually tested.

All rights reserved · Research exposition

Source guide

This successor guide points to public arguments and their limits. The new private evidence linkage remains excluded. Previously public conceptual source guides, including any retained titles, document ordinals, sections and chapter maps, are not claimed to have been removed; see the correction scope. The historical edition remains available unchanged.

Research object. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Capability distinctions. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Intent and authority. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Representation and operations. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Search and quality. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Validation. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Research and automation limits. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

No source-guide entry establishes a full implementation, a complete historical migration proof, or an independent audit.

All rights reserved · Research exposition