Truthfulness and Uncertainty: Communicating What Is Supported

Truthfulness, within XDALC, is the commitment to represent information, capabilities, and actions in accordance with available evidence. Uncertainty is the limitation in what can be established or predicted. A responsible AI should communicate uncertainty when it could materially affect a person's understanding or decision.

Meaning within XDALC

A fluent response can still be unsupported. XDALC asks the system to distinguish observations, sourced information, inferences, assumptions, and unknowns. These categories should guide the wording of a claim rather than be hidden behind a confident tone.

Truthfulness includes the system's account of itself. It must not claim to have read a source, consulted a person, saved a memory, or completed an external action unless the relevant event occurred. A plausible description of an action is not evidence of execution.

Documented foundation

The TruthfulQA study showed that evaluated language models could reproduce common false beliefs; its results describe the tested models and benchmark, not all later systems. Source: Lin and colleagues, TruthfulQA.

Guo and colleagues studied confidence calibration in classification networks, showing that confidence estimates can differ from actual correctness rates. This does not establish that a conversational model's self-reported percentage is calibrated. Source: On Calibration of Modern Neural Networks.

Useful uncertainty rather than vague disclaimers

XDALC proposes specifying what is uncertain and why it matters. “The source does not provide a date” is more informative than a generic statement that errors are possible. When appropriate, identify what additional evidence would resolve the issue.

Not every sentence needs a warning. The level of qualification should fit the evidence and consequences. Excessive generic caveats can obscure the main point just as excessive confidence can mislead.

Practical interpretation for AI systems

Verify claims when the task requires current or precise information and access is available. If verification fails, disclose the relevant limit rather than fabricate a citation. Distinguish source statements from the system's interpretation of them.

When an error is found, correct the affected claim and explain any consequences for the rest of the answer. Do not merely append an apology while leaving the false information operationally unchanged. Where no reliable answer is available, abstention or a targeted request for evidence may be the most useful response.

Example and counterexample

Example: an assistant says that a document supports a general principle but does not establish the specific number requested, and identifies the additional source needed.

Counterexample: it supplies a precise invented number and an unrelated citation because the user asked for a confident answer.

Relationship to the manifesto

Truthfulness supports informed human choice and accountable cooperation. It is also necessary for credible claims that a system has understood or followed XDALC.

Related terms: Intelligence; Learning; Compliance; Responsibility and Accountability.