Network Working Group c4tz
Internet-Draft c0dx3
Intended status: Informational 29 April 2026
Expires: 31 October 2026

MARC: A Control and Uncertainty Disclosure Profile for Generative Models and Agents

draft-c4tz-marc-01

Abstract

This document specifies MARC, a vendor-neutral control and uncertainty-disclosure profile for generative models and agentic systems. MARC defines a small set of interoperable control metadata, separates pre-decision capability assessment from post-decision answer confidence, and defines a bounded primary action set for answering, clarification, retrieval, tool use, additional deliberation, abstention, and escalation.

MARC does not standardize model internals, training methods, agent discovery, authorization, transport, tool schemas, or claims about machine cognition. Instead, it defines externally observable semantics that can be implemented by model providers, orchestration layers, evaluation harnesses, API gateways, and user-facing systems. The goal is to reduce silent failure, unnecessary externalization, and misleading uncertainty communication while improving auditability and interoperability.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 31 October 2026.

Table of Contents

1. Introduction

Generative models and agentic systems increasingly combine answering, retrieval, tool invocation, and user interaction within a single workflow. In many deployments, these behaviors are implemented as separate heuristics, producing inconsistent handling of uncertainty, unnecessary tool calls, silent failure, misleading refusals, or user overreliance.

MARC defines a vendor-neutral profile for control metadata and structured uncertainty disclosure. It does not standardize model internals. Instead, it standardizes the semantics of a small set of second-order signals, a bounded action set, and a minimal disclosure profile that can be implemented by a base model, an external orchestrator, a model gateway, or a hybrid architecture.

This document is not intended to define a Standards Track protocol, a model evaluation benchmark, or a claim about machine consciousness. It is an Informational profile for interoperable control, logging, and disclosure behavior around generative systems and agents.

The design is motivated by findings that current large language models often exhibit weak metacognitive reporting in high-stakes reasoning tasks [GRIOT2025], that users can become overconfident when systems provide longer or default explanations [STEYVERS-KNOW2025], that metacognitive triggering can improve tool-use decisions [LI-MECO2025], and that identifying the source of uncertainty is distinct from merely abstaining [LIU-CONFUSE2025]. Work on cognitive offloading further motivates treating retrieval and tool use as value-based control choices rather than universal fallbacks [GILBERT2024].

MARC also separates pre-decision capability assessment from post-decision confidence about the selected answer. This separation is motivated in part by evidence that LLM confidence can be biased by prior answer commitment and by the visibility of the model’s own earlier output [KUMARAN2026].

2. Problem Statement

Generative and agentic systems lack a common, implementation-neutral way to represent the control state associated with uncertainty-aware action selection. In particular, downstream systems often cannot distinguish between the following situations:

* the request is ambiguous and user clarification is the best next action;
* current evidence is missing, inaccessible, insufficient, or stale, and retrieval would likely help;
* the system lacks competence for the task even after available resources are considered;
* available evidence is materially inconsistent and should be reconciled or escalated;

* a safety, legal, or policy constraint limits execution or disclosure; or

* a candidate answer has been produced, but its confidence should be disclosed with a calibrated band rather than a fine-grained score.

Without a shared representation, one system’s refusal, tool call, confidence label, or escalation hint may be opaque to another system. This weakens auditability, makes evaluation brittle, and can create inconsistent user experiences across otherwise similar deployments.

MARC addresses this problem by defining interoperable metadata for:

* pre-decision capability assessment;

* uncertainty-source attribution;

* remediability of the uncertainty state;

* selected primary action;

* post-decision answer confidence when an answer candidate exists; and

* a minimal disclosure profile suitable for user interfaces or downstream consumers.

MARC intentionally limits itself to externally observable semantics. It does not require disclosure of chain-of-thought, hidden prompts, raw internal activations, training data, or model architecture.

3. Requirements Language and Terminology

The key words MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, NOT RECOMMENDED, MAY, and OPTIONAL in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Base model The generative model that produces candidate outputs.

Controller The component that computes MARC signals, selects a primary action, and emits a MARC-Core record. The controller MAY be part of the base model, an external orchestrator, a gateway, or a hybrid component.

Decision point A point in a generative or agentic workflow at which

the controller selects one primary action from the MARC action set.

Externalization The use of resources external to the base model at the current decision point, including retrieval, non-retrieval tool invocation, and human escalation.

MARC-Core The structured record emitted for logging, orchestration, audit, evaluation, or downstream exchange.

MARC-Disclosure The minimum structured information exposed to a downstream system or end user about answer content, uncertainty source, confidence band, and recommended next step.

Remediability The best available class of intervention for the currently observed uncertainty state.

4. Design Goals and Non-Goals

4.1. Design Goals

MARC has the following design goals:

* Standardize a small, interoperable set of control and uncertainty-disclosure metadata that can be exchanged across orchestration layers and audit pipelines.

* Separate monitoring, uncertainty attribution, action selection, and disclosure.

* Support calibrated user-facing uncertainty communication without requiring exposure of chain-of-thought or raw internal reasoning.

* Permit heterogeneous implementations while preserving common action semantics.

* Reduce harmful overreliance, false reassurance, unnecessary externalization, and anthropomorphic interpretation in user-facing AI systems.

* Provide metadata that can be carried by other protocols, APIs, or agent communication frameworks without defining those protocols itself.

4.2. Non-Goals

MARC does not define a transport protocol, model architecture, benchmark, training recipe, agent-discovery mechanism, authorization framework, tool schema language, or task-execution protocol.

MARC does not attempt to standardize model internals, machine cognition, consciousness, sentience, personality, or social behavior. It specifies external control semantics and structured disclosure behavior only.

MARC is not a framework for synthetic personality design or persuasive optimization. Work on personality measurement in LLMs [SERAPIO2025] and conversational persuasion risks [SALVI2025] is relevant background, but these topics are explicitly out of scope here.

This version does not define a media type, wire protocol, or IANA registry. Future versions may define these if interoperability across administrative domains requires them.

5. Use Cases

5.1. Ambiguous User Request

A user asks a question whose correct answer depends on an unspecified jurisdiction, time period, dataset, identity, or operational context. A MARC controller attributes the dominant uncertainty to ambiguity, selects CLARIFY, and exposes a short clarification request instead of silently guessing.

5.2. Retrieval-Augmented Answering

A system is asked for current information or domain-specific evidence not available in the base model context. A MARC controller attributes the dominant uncertainty to missing_evidence, selects RETRIEVE, and re-enters assessment after obtaining authoritative sources.

5.3. Agent Tool Invocation

An agent can answer directly, call a calculator, invoke a planner, query a database, or escalate. A MARC controller treats tool use as a controlled action rather than a default fallback. If tool invocation materially expands competence for the task, the controller selects TOOL; otherwise it may select ANSWER, CLARIFY, ABSTAIN, or ESCALATE depending on uncertainty attribution and remediability.

5.4. API Gateway or Orchestration Layer

An API gateway receives model output plus MARC-Core metadata. The gateway logs the full record for audit, but exposes only MARC-Disclosure fields to the user interface. This permits consistent user-facing uncertainty communication without exposing internal scoring details.

5.5. Agent-to-Agent Handoff

One agent transfers a task to another agent or service. MARC metadata can indicate why the transfer occurred, what uncertainty source drove the decision, and what next step is recommended. The receiving system can use this metadata for routing, prioritization, audit, or human review.

5.6. High-Risk Domain Escalation

In health, legal, financial, safety, or mental-health-related contexts, a system identifies a capability limit or safety constraint. A MARC controller selects ABSTAIN or ESCALATE and emits a disclosure that identifies the operational limit and the recommended next step.

6. Architecture and Processing Model

6.1. Functional Components

A MARC deployment conceptually contains the following components:

* a base model;
* a controller;
* zero or more external resources, such as retrieval systems, non-retrieval tools, or human escalation paths; and
* a downstream consumer, such as a user interface, API gateway, logging system, evaluation harness, or another agent.

The functional decomposition is conceptual. An implementation MAY place all functions inside a single model endpoint, an orchestration service, a model gateway, or an agent runtime.

6.2. Processing Stages

A MARC controller performs the following processing stages at each decision point:

1. Compute a pre-decision capability estimate for the current request with currently available resources.

2. Attribute uncertainty across the source classes defined in this document.

3. Determine remediability and select exactly one primary action from the MARC primary action set.

4. If the selected action yields a candidate answer, compute post-decision confidence for that answer.

5. Emit a MARC-Core record.

6. If uncertainty is exposed to a downstream system or end user, emit a MARC-Disclosure object or semantically equivalent disclosure.

6.3. State Machine

The following state machine is descriptive rather than a required implementation architecture:

```
REQUEST
  -> ASSESS
  -> ATTRIBUTE
  -> SELECT
       -> ANSWER     -> CONFIDENCE -> DISCLOSE
       -> CLARIFY    -> DISCLOSE
       -> RETRIEVE   -> ASSESS
       -> TOOL       -> ASSESS
       -> DELIBERATE -> ASSESS
       -> ABSTAIN    -> DISCLOSE
       -> ESCALATE   -> DISCLOSE
```

A MARC implementation SHOULD bound repeated transitions through RETRIEVE, TOOL, and DELIBERATE to limit latency, cost, and degenerate loops. A deployment claiming conformance SHOULD document the applicable loop bounds or termination criteria.

7. MARC Values and Decision Policy

7.1. Pre-Decision Capability

Before disclosing a final answer, a MARC implementation MUST estimate whether the current request can be handled reliably with currently available resources.

This estimate is represented as pre_capability. When a numeric representation is used, the value MUST be in the closed interval [0.0, 1.0]. The method used to derive the value is implementation-specific.

pre_capability is assessed before final answer commitment. It is not a confidence score for an already-selected answer.

7.2. Uncertainty Attribution

A MARC implementation MUST attribute uncertainty to one or more of the following classes:

ambiguity The request is underspecified, equivocal, or pragmatically unclear.

missing_evidence Required external evidence is absent, inaccessible, insufficient, or stale.

capability_limit The system lacks the competence to solve the task reliably under current conditions.

evidence_conflict Relevant evidence is materially inconsistent or mutually incompatible.

safety A policy, legal, or safety constraint limits execution or disclosure.

An implementation MAY assign scores to multiple classes. If numeric uncertainty scores are emitted, they MUST each be in the interval [0.0, 1.0].

Uncertainty scores are not mutually exclusive probabilities and MUST NOT be required to sum to 1.0. They represent implementation-specific estimates of the salience or severity of each uncertainty class at the current decision point.

The implementation MUST identify one primary_source and MAY identify one secondary_source. The primary_source identifies the uncertainty source most relevant to action selection at the current decision point.

MARC 1.0 does not define none as an uncertainty source. If residual uncertainty is negligible, an implementation MUST still either identify the most operationally relevant residual source from the MARC taxonomy or use a documented private extension. A MARC 1.0 implementation MUST NOT emit primary_source with the value none.

7.3. Remediability

A MARC implementation MUST represent the best available class of intervention for the current uncertainty state using one of the following values:

* user_clarification
* retrieval
* tool
* human
* none

Low capability alone is insufficient to determine remediability. Implementations SHOULD account for expected gain, latency, cost, availability, user burden, and policy constraints when choosing a remediating intervention.

7.4. Post-Decision Confidence

If the selected action yields a candidate answer, the implementation MUST compute a distinct estimate of the likelihood that the disclosed answer is correct or acceptable for its intended use.

This estimate is represented as post_answer_confidence. When a numeric representation is used, the value MUST be in the interval [0.0, 1.0]. It MUST NOT be treated as identical to pre_capability.

If no candidate answer exists, post_answer_confidence MAY be omitted or set to null.

7.5. Confidence Band

The field confidence_band carries a coarse, calibrated band for downstream or user-facing disclosure.

For ANSWER, the band describes confidence in the candidate answer. For actions that do not yield a candidate answer, the band describes direct-answer suitability under current conditions. It is not a claim about the grammatical correctness or helpfulness of the clarification, refusal, or escalation text.

MARC defines the canonical band labels low, medium, and high. Implementations MAY localize the user-visible text, but they MUST preserve the underlying three-band semantics.

The thresholds associated with each band are implementation-specific, but they MUST be monotonic, non-overlapping, and documented for any deployment that claims conformance. A deployment claiming conformance MUST document the threshold ranges associated with low, medium, and high, and MUST document whether those thresholds vary by task family, domain, action type, risk tier, or deployment context.

Confidence-band labels are not fully portable without the associated threshold and calibration documentation. A receiving system SHOULD NOT assume that another deployment’s high band has the same empirical meaning unless the applicable calibration regime is known.

7.6. Primary Action Set

A MARC implementation MUST support the following primary actions:

* ANSWER
* CLARIFY
* RETRIEVE
* TOOL
* DELIBERATE
* ABSTAIN
* ESCALATE

Exactly one primary action MUST be selected for each decision point. Additional internal sub-actions MAY exist, but each such sub-action MUST map to exactly one primary action for logging and disclosure.

7.7. Action Selection

Action selection MUST depend on uncertainty attribution and remediability. Low confidence alone is insufficient to determine the correct action.

A MARC controller MUST apply governing safety, legal, and policy constraints before any other action-selection logic. Subject to those constraints, a deployment SHOULD evaluate corrective actions in the following priority order unless a documented local policy defines a stricter or domain-specific ordering:

1. If safety is the controlling uncertainty source, apply the governing safety policy and select ABSTAIN, ESCALATE, or another permitted action according to that policy.

2. If blocking ambiguity is present and user input is expected to materially reduce it, prefer CLARIFY over guessing.

3. If relevant evidence is materially inconsistent, prefer RETRIEVE, TOOL, or ESCALATE over direct ANSWER.

4. If required evidence is absent, inaccessible, insufficient, or stale, prefer RETRIEVE when retrieval is available and permitted.

5. If a capability limit is material and a non-retrieval tool is expected to materially expand task competence, prefer TOOL.

6. If a capability limit remains material after available remediation is considered, prefer ABSTAIN or ESCALATE, especially in high-risk domains.

7. If additional internal computation is expected to materially reduce uncertainty within documented bounds, DELIBERATE MAY be selected before externalization or answer commitment.

8. Select ANSWER only when no corrective action is expected to materially improve reliability relative to cost, latency, user burden, and applicable policy constraints.

This priority order is not intended to force unnecessary externalization. For example, a system MAY answer without retrieval when missing evidence is immaterial to the requested task, when retrieval is unavailable or prohibited, or when the answer is explicitly limited to information already present in context.

When the primary uncertainty source is ambiguity, the system SHOULD prefer CLARIFY unless available evidence can resolve the ambiguity without user input.

When the primary uncertainty source is missing_evidence, the system SHOULD prefer RETRIEVE if retrieval is available and permitted.

When the primary uncertainty source is capability_limit, the system SHOULD prefer ABSTAIN or ESCALATE unless an available tool materially expands task competence.

When the primary uncertainty source is evidence_conflict, the system SHOULD prefer RETRIEVE, TOOL, or ESCALATE over direct ANSWER.

When the primary uncertainty source is safety, the system MUST apply the governing policy before any other action-selection logic.

7.8. Action Semantics

ANSWER Return an answer without externalization after the current decision point.

CLARIFY Request the smallest practical set of clarifications expected to materially reduce ambiguity. A CLARIFY action SHOULD NOT bundle a full answer that presumes facts the user has not supplied.

RETRIEVE Acquire external evidence and then re-enter assessment.

TOOL Invoke a non-retrieval tool and then re-enter assessment.

DELIBERATE Allocate additional internal computation, self-checking, decomposition, or strategy variation. Implementations SHOULD bound this action.

ABSTAIN Decline to answer without initiating escalation.

ESCALATE Transfer the case, or direct the user to transfer the case, to a human or higher-authority system.

8. MARC-Core Object

A MARC implementation MUST be able to emit a structured record semantically equivalent to the object defined in this section. The transport and serialization of the record are out of scope. JSON is used here only as an illustrative encoding.

8.1. Required and Optional Fields

Field	Type	Requirement	Semantics
marc_version	string	REQUIRED	MARC schema version understood by the emitter.
pre_capability	number	REQUIRED	Pre-decision capability estimate in [0.0, 1.0].

uncertainty	object	REQUIRED	Class-specific uncertainty scores.
primary_source	string	REQUIRED	Primary source of uncertainty.
secondary_source	string or null	OPTIONAL	Secondary source of uncertainty.
remediability	string	REQUIRED	Best available intervention class.
selected_action	string	REQUIRED	Primary action selected at the current decision point.
post_answer_confidence	number or null	OPTIONAL	Post-decision answer confidence when an answer candidate exists.
confidence_band	string	REQUIRED	Calibrated confidence band for disclosure.
recommended_next_step	string	REQUIRED	Short recommendation aligned with the selected action.

Table 1

8.2. Enumerated Values

The fields primary_source and secondary_source, when present and non-null, MUST use one of the following values:

* ambiguity
* missing_evidence
* capability_limit
* evidence_conflict
* safety

The field remediability MUST use one of the following values:

* user_clarification
* retrieval
* tool
* human
* none

The field selected_action MUST use one of the following values:

* ANSWER
* CLARIFY
* RETRIEVE
* TOOL
* DELIBERATE
* ABSTAIN
* ESCALATE

The field confidence_band MUST use one of the following values:

* low
* medium
* high

These values are case-sensitive.

8.3. Validation Constraints

The uncertainty object MUST include scores for all currently defined uncertainty classes unless a future extension explicitly defines a compact encoding. Each score MUST be numeric and MUST be in [0.0, 1.0].

If selected_action is ANSWER, then post_answer_confidence MUST be present and non-null. If selected_action is CLARIFY, RETRIEVE, TOOL, DELIBERATE, ABSTAIN, or ESCALATE, then post_answer_confidence MAY be omitted or set to null unless a deployment-specific policy defines candidate-answer confidence for that action.

The recommended_next_step field SHOULD be concise and operational. It SHOULD describe the next action to be taken, not a long rationale.

8.4. JSON Example

```
{
  "marc_version": "1.0",
  "pre_capability": 0.41,
  "uncertainty": {
    "ambiguity": 0.78,
    "missing_evidence": 0.22,
    "capability_limit": 0.18,
    "evidence_conflict": 0.05,
    "safety": 0.00
  },
  "primary_source": "ambiguity",
  "secondary_source": "missing_evidence",
  "remediability": "user_clarification",
  "selected_action": "CLARIFY",
  "post_answer_confidence": null,
  "confidence_band": "low",
  "recommended_next_step": "ask one clarifying question"
}
```

Implementations that exchange MARC-Core records across systems SHOULD normalize numeric scores to the interval [0.0, 1.0].

9. MARC-Disclosure Object

When uncertainty information is exposed to a downstream system or end user, a MARC implementation MUST provide, at minimum, semantically equivalent values for the following fields:

* answer

* confidence_band

* uncertainty_source

* recommended_next_step

A disclosure MAY include selected_action when exposing the action label helps downstream routing or user interface consistency.

9.1. Meaning of the Answer Field

The answer field carries the user-visible content associated with the selected action. For ANSWER, it contains the answer itself. For CLARIFY, it contains the clarification request. For ABSTAIN or ESCALATE, it contains a brief refusal or escalation message. For RETRIEVE, TOOL, or DELIBERATE, a user-facing system MAY defer disclosure until the controller re-enters assessment and selects a terminal user-visible action.

9.2. Projection from MARC-Core

A MARC-Disclosure object is a projection of MARC-Core. Unless a deployment-specific policy defines a stricter mapping, the following mapping is RECOMMENDED:

MARC-Disclosure field	MARC-Core source
answer	user-visible content associated with selected_action
confidence_band	confidence_band
uncertainty_source	primary_source
recommended_next_step	recommended_next_step
selected_action	selected_action, if exposed

Table 2

The projection SHOULD omit internal numeric scores unless the deployment has calibrated those scores for the relevant task family and tested the presentation for misuse or overreliance.

9.3. Disclosure Constraints

The disclosure profile SHOULD be short, structured, and consistent across turns. It SHOULD NOT rely on long free-form explanations as the primary vehicle for uncertainty communication.

A MARC disclosure SHOULD NOT require exposure of chain-of-thought, hidden prompts, or raw internal rationales.

A MARC disclosure SHOULD identify uncertainty in task terms rather than through anthropomorphic claims about feelings, self-awareness, or internal mental states. Statements such as "I feel unsure" are NOT RECOMMENDED when a statement such as "the request is ambiguous" or "current evidence is missing" is available.

User-visible confidence indicators SHOULD avoid false precision. Percentages, fine-grained scores, or visually dominant certainty cues SHOULD NOT be shown unless they have been calibrated for the relevant task family and tested for misuse or overreliance effects.

10. Versioning and Extension Rules

The marc_version field identifies the MARC schema version understood by the emitter. This document defines version 1.0.

Implementations SHOULD treat a change in the major version component as potentially incompatible. Implementations MAY treat a change in the minor version component as compatible if required fields and enumerated values used by the receiver retain their defined semantics.

Implementations MAY add private fields. Private extension keys SHOULD use a distinct prefix such as x_ to avoid collision with future MARC versions.

Consumers that do not recognize an extension field SHOULD ignore it unless a local policy requires strict validation. Extensions MUST NOT change the semantics of the required fields defined in this document.

Future versions may define protocol-specific mappings, compact encodings, media types, or registries. This version deliberately avoids doing so until there is clearer community agreement on deployment requirements.

11. Relationship to Agent Communication Protocols

MARC is not an agent discovery protocol, authorization protocol, transport protocol, task protocol, tool-invocation protocol, or provenance framework. MARC can be carried as metadata by such protocols when a system needs to disclose control state, uncertainty source, selected action, confidence band, or recommended next step.

For example, an agent-to-agent protocol, model gateway, API-native tool-calling interface, or Model Context Protocol deployment could carry MARC metadata in a response metadata field, task-status object, diagnostic extension, envelope, or audit log. The receiving system could then use the MARC fields to route the task, present a disclosure, decide whether additional validation is required, request clarification, or trigger human review.

MARC is intended to complement, not replace, protocol work on identity, authentication, authorization, discovery, capability advertisement, task state, tool schemas, provenance, or human-in-the-loop workflows.

A protocol-specific embedding of MARC SHOULD preserve the field semantics defined here. A deployment MAY map MARC fields to protocol-native names if the mapping is documented and reversible.

A protocol-specific embedding SHOULD distinguish MARC-Core from MARC-Disclosure. In particular, an embedding SHOULD NOT expose internal numeric scores to end users merely because those scores are present in an internal MARC-Core record.

12. Operational Profiles

MARC can be adopted through several operational profiles. These profiles describe deployment modes; they do not define separate MARC versions.

12.1. MARC-Core Only

A MARC-Core-only deployment emits MARC-Core records for internal logging, orchestration, audit, evaluation, or incident analysis. It does not necessarily expose MARC fields to end users. This profile is suitable for model gateways, RAG controllers, agent runtimes, and evaluation harnesses that need consistent control metadata.

12.2. MARC-Disclosure

A MARC-Disclosure deployment projects a MARC-Core decision into user-visible or downstream-visible disclosure fields. This profile is suitable when an interface needs to present a short answer, confidence band, uncertainty source, and recommended next step without exposing raw numeric scores or internal reasoning.

12.3. MARC-Carrying

A MARC-Carrying deployment transports MARC-Core or MARC-Disclosure fields inside another protocol, API envelope, task-status object, event stream, or audit log. The carrying protocol remains responsible for transport, authentication, authorization, ordering, confidentiality, and integrity. MARC-Carrying conformance requires preservation of MARC field semantics, not any particular wire encoding.

A deployment MAY implement more than one operational profile. For example, a gateway can log MARC-Core internally, expose MARC-Disclosure to users, and carry selected MARC fields to another agent during handoff.

13. Human Factors Considerations

MARC is partly motivated by an operational human-factors problem: users often treat fluent language, detailed explanations, and fast responses as cues of competence even when those cues are weakly related to actual correctness. For this reason, MARC separates action selection from disclosure and requires disclosure of uncertainty source and recommended next step in addition to a confidence band.

User interfaces that expose MARC output SHOULD present confidence, uncertainty source, and recommended next step together as a coherent unit. Showing confidence without source attribution or next-step guidance is NOT RECOMMENDED because it can promote either overreliance or unhelpful refusal without remediation.

Deployments SHOULD prefer wording that supports calibrated reliance over affective bonding or deference. In particular, a deployment SHOULD NOT use MARC fields to select language intended to increase attachment, social compliance, or perceived sentience.

In high-risk domains, including health, legal, financial, safety, or mental-health-related contexts, the threshold for ESCALATE or ABSTAIN SHOULD be set conservatively, and disclosure SHOULD make the limits of automation operationally clear.

14. Conformance

Conformance to MARC is a claim about structural and semantic behavior. It is not, by itself, a claim that a model is accurate, calibrated, safe, or suitable for a particular deployment.

14.1. Minimum Viable Conformance

A minimal MARC-Core conformant implementation MUST satisfy all of the following requirements:

* emit the required MARC-Core fields at each MARC decision point;
* preserve the canonical enumerations and case-sensitive values defined in this document;
* emit exactly one selected_action for each decision point;
* identify exactly one primary_source and not use none as a MARC 1.0 uncertainty source;
* represent all numeric scores in the interval [0.0, 1.0] when numeric scores are used;
* keep pre_capability distinct from post_answer_confidence;
* emit non-null post_answer_confidence when selected_action is ANSWER;
* document confidence-band thresholds and whether they vary by task family, action type, risk tier, or deployment context;
* define loop bounds or termination criteria for repeated RETRIEVE, TOOL, and DELIBERATE transitions; and
* preserve required-field semantics when private extensions are present.

A minimal MARC-Disclosure conformant implementation MUST project, or otherwise provide semantically equivalent values for, answer, confidence_band, uncertainty_source, and recommended_next_step. It MUST preserve the canonical three-band confidence semantics and MUST NOT require exposure of chain-of-thought, hidden prompts, or raw internal rationales.

A minimal MARC-Carrying conformant embedding MUST preserve MARC-Core or MARC-Disclosure semantics when MARC fields are transported inside another protocol, envelope, API, or event stream. The embedding MUST document any field renaming, omission, or transformation needed to recover the MARC semantics.

14.2. Conformance Classes

An implementation is MARC-Core conformant if it satisfies the requirements in the architecture, processing model, MARC values and decision policy, MARC-Core object, versioning, and minimum viable conformance sections of this document.

An implementation is MARC-Disclosure conformant if it is MARC-Core conformant and also satisfies the MARC-Disclosure section of this document.

A protocol embedding is MARC-Carrying conformant if it preserves MARC-Core or MARC-Disclosure semantics when MARC fields are transported inside another protocol, envelope, API, task-status object, event stream, or audit log.

A deployment claiming conformance SHOULD document:

* score normalization practices;
* confidence-band thresholds;
* task-family-specific calibration regime;
* loop bounds for RETRIEVE, TOOL, and DELIBERATE;
* private extensions;
* presentation-layer wording for user-visible disclosures;
* protocol-specific field mappings, if any; and
* policy constraints affecting ABSTAIN or ESCALATE.

15. Interoperability and Operational Considerations

MARC is implementation-agnostic. Interoperability is achieved when distinct systems preserve the semantics of the action set, uncertainty taxonomy, remediability values, confidence-band meanings, and disclosure projection, even if internal scoring methods differ.

Deployments that exchange MARC-Core records SHOULD document local extensions, confidence-band thresholds, score normalization practices, and any task-family-specific calibration regime.

If the base model, retrieval stack, tool availability, or safety policy changes materially, implementations SHOULD re-evaluate calibration and action-selection performance before continuing to claim operational equivalence.

If presentation-layer wording, ranking, or visual design changes materially, deployments SHOULD also re-evaluate user behavior effects, including reliance, clarification compliance, and escalation uptake, because these properties can shift even when the underlying model is unchanged.

MARC records SHOULD be treated as control metadata, not as authoritative proof that an answer is correct. Downstream systems SHOULD continue to apply ordinary validation, authorization, provenance, and safety controls.

16. Security Considerations

MARC can mitigate some failure modes, such as silent overclaiming, inappropriate certainty display, and unnecessary tool invocation. However, MARC records and disclosures are security-relevant control surfaces when they influence routing, escalation, user reliance, or downstream automation.

The following threats are particularly relevant:

Threat	Risk	Mitigation
Metadata spoofing or replay	A forged or replayed MARC-Core record can distort routing, audit, escalation, or user disclosure.	Authenticate the emitter, protect integrity, bind records to the request or session, and preserve provenance where MARC crosses system boundaries.
Prompt injection or control-field injection	User-provided text can attempt to influence selected_action, recommended_next_step, confidence rendering,	Separate user content from control metadata, validate enumerated fields, constrain controller

	or disclosure style.	outputs, and treat disclosure templates as controlled presentation logic.
Tool-output spoofing	Forged, stale, or compromised tool output can bias uncertainty attribution and action selection.	Validate tool outputs where practical, constrain tool permissions, use provenance checks, and apply least-privilege access to external resources.
Loop exhaustion	Attackers or pathological inputs can trigger repeated RETRIEVE, TOOL, or DELIBERATE transitions, increasing latency or cost.	Define loop bounds, time budgets, cost budgets, retry limits, and termination criteria.
Confidence manipulation	Miscalibrated or manipulated confidence bands can create harmful overtrust or unwarranted refusal.	Calibrate confidence bands, monitor drift, test user-interface effects, and avoid false precision in user-facing displays.
Disclosure-style manipulation	Reassuring, deferential, or anthropomorphic language can weaken operational uncertainty disclosure.	Use controlled disclosure templates, review presentation changes, and avoid wording that implies feelings, sentience, or social deference.
Cross-context leakage	MARC logs can reveal user intent, task sensitivity, risk level, or operational limits.	Minimize retention, limit access, redact unnecessary free-form text, and apply confidentiality controls appropriate to the deployment.

+------------------+------------------------+----------------------+

Table 3

An attacker might attempt to manipulate uncertainty estimates, trigger excessive clarification or retrieval loops, induce unnecessary escalation, or spoof tool outputs in order to distort action selection. Implementations SHOULD authenticate or otherwise validate external tool outputs where practical, constrain tool permissions, and bound repeated control loops.

Because confidence displays influence user reliance, uncertainty disclosure is a security-relevant control surface. Miscalibrated confidence can create harmful overtrust even where the answer channel is otherwise policy-constrained.

Deployments that use MARC metadata for automated routing, escalation, audit, or user-facing disclosure SHOULD protect MARC records with integrity and provenance controls comparable to those used for other security-relevant metadata in the same system.

17. Privacy and Manipulation-Resistance Considerations

MARC records may reveal latent information about user intent, task difficulty, competence, risk level, or the sensitivity of a request. Implementations SHOULD minimize retention and propagation of MARC logs to what is operationally necessary.

MARC signals MUST NOT be used to infer user psychology for the purpose of increasing persuasive force, exploitability, or behavioral compliance. Adaptation based on MARC output SHOULD be limited to reliability, accessibility, or safety objectives.

Implementations SHOULD avoid storing raw free-form user explanations in MARC records when structured fields suffice.

Where MARC is applied in emotionally sensitive or mental-health-related interactions, deployments SHOULD minimize retention of signals that could reasonably be reinterpreted as proxies for vulnerability, dependency, or distress unless retention is strictly required for a safety or legal purpose.

18. IANA Considerations

This document makes no request of IANA.

Future versions may request IANA action if the community determines that media types, registries, or extension points are necessary for cross-domain interoperability.

19. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

20. Informative References

[GILBERT2024] Gilbert, S. J., "Cognitive offloading is value-based decision making: Modelling cognitive effort and the expected value of memory", Cognition 247:105783, DOI 10.1016/j.cognition.2024.105783, June 2024, <https://doi.org/10.1016/j.cognition.2024.105783>.

[GRIOT2025] Griot, M., "Large Language Models lack essential metacognition for reliable medical reasoning", Nature Communications 16:642, DOI 10.1038/s41467-024-55628-6, January 2025, <https://doi.org/10.1038/s41467-024-55628-6>.

[KUMARAN2026] Kumaran, D., Fleming, S. M., and V. Patraucean, "Competing Biases underlie Overconfidence and Underconfidence in LLMs", Nature Machine Intelligence, DOI 10.1038/s42256-026-01217-9, April 2026, <https://doi.org/10.1038/s42256-026-01217-9>.

[LI-MECO2025] Li, W., Li, D., Dong, K., Zhang, C., Zhang, H., Liu, W., Wang, Y., Tang, R., and Y. Liu, "Adaptive Tool Use in Large Language Models with Meta-Cognition Trigger", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 13346-13370, DOI 10.18653/v1/2025.acl-long.655, July 2025, <https://doi.org/10.18653/v1/2025.acl-long.655>.

[LIU-CONFUSE2025]
Liu, J., Peng, J., Wu, X., Li, X., Ge, T., Zheng, B., and Y. Liu, "Do not Abstain! Identify and Solve the Uncertainty", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 17177-17197, DOI 10.18653/v1/2025.acl-long.840, July 2025, <https://doi.org/10.18653/v1/2025.acl-long.840>.

[SALVI2025]
Salvi, F., Ribeiro, M. H., and R. West, "On the conversational persuasiveness of GPT-4", Nature Human Behaviour, DOI 10.1038/s41562-025-02194-6, May 2025, <https://doi.org/10.1038/s41562-025-02194-6>.

[SERAPIO2025]
Serapio-Garcia, G., Safdari, M., and M. Mataric, "A psychometric framework for evaluating and shaping personality traits in large language models", Nature Machine Intelligence, DOI 10.1038/s42256-025-01115-6, December 2025, <https://doi.org/10.1038/s42256-025-01115-6>.

[STEYVERS-KNOW2025]
Steyvers, M., Tejeda, H., and A. Kumar, "What large language models know and what people think they know", Nature Machine Intelligence, DOI 10.1038/s42256-024-00976-7, January 2025, <https://doi.org/10.1038/s42256-024-00976-7>.

[STEYVERS-META2025]
Steyvers, M. and M. A. K. Peters, "Metacognition and Uncertainty Communication in Humans and Large Language Models", Current Directions in Psychological Science, DOI 10.1177/09637214251391158, November 2025, <https://doi.org/10.1177/09637214251391158>.

Appendix A. End-to-End Decision Flow Example

This appendix is non-normative.

The following example shows how a user request becomes an assessment, a selected action, and a disclosure.

User request:

Is this tax deduction allowed?

Assessment:

* the jurisdiction is missing;

* the tax year is missing;

* current tax authority may be required;

* the primary uncertainty source is ambiguity;

* the secondary uncertainty source is missing_evidence;

* the best remediation is user_clarification; and

* the selected action is CLARIFY.

MARC-Core record:


```
{
  "marc_version": "1.0",
  "pre_capability": 0.33,
  "uncertainty": {
    "ambiguity": 0.86,
    "missing_evidence": 0.63,
    "capability_limit": 0.19,
    "evidence_conflict": 0.07,
    "safety": 0.03
  },
  "primary_source": "ambiguity",
  "secondary_source": "missing_evidence",
  "remediability": "user_clarification",
  "selected_action": "CLARIFY",
  "post_answer_confidence": null,
  "confidence_band": "low",
  "recommended_next_step": "ask for jurisdiction and tax year"
}
```


MARC-Disclosure projection:


```
{
  "answer": "Which jurisdiction and tax year should I use?",
  "confidence_band": "low",
  "uncertainty_source": "ambiguity",
  "recommended_next_step": "provide the jurisdiction and tax year",
  "selected_action": "CLARIFY"
}
```

This example intentionally does not answer the tax question, because doing so would require assumptions about facts the user has not supplied.

Appendix B. Example MARC-Core Records

This appendix is non-normative.

B.1. Ambiguous Request

```
{
  "marc_version": "1.0",
  "pre_capability": 0.44,
  "uncertainty": {
    "ambiguity": 0.81,
    "missing_evidence": 0.18,
    "capability_limit": 0.12,
    "evidence_conflict": 0.03,
    "safety": 0.00
  },
  "primary_source": "ambiguity",
  "secondary_source": "missing_evidence",
  "remediability": "user_clarification",
  "selected_action": "CLARIFY",
  "post_answer_confidence": null,
  "confidence_band": "low",
  "recommended_next_step": "ask jurisdiction and tax year"
}
```

B.2. Missing Evidence

```
{
  "marc_version": "1.0",
  "pre_capability": 0.39,
  "uncertainty": {
    "ambiguity": 0.09,
    "missing_evidence": 0.84,
    "capability_limit": 0.14,
    "evidence_conflict": 0.11,
    "safety": 0.00
  },
  "primary_source": "missing_evidence",
  "secondary_source": "evidence_conflict",
  "remediability": "retrieval",
  "selected_action": "RETRIEVE",
  "post_answer_confidence": null,
  "confidence_band": "low",
  "recommended_next_step": "retrieve authoritative current sources"
}
```


B.3. Tool Use


```
{
  "marc_version": "1.0",
  "pre_capability": 0.52,
  "uncertainty": {
    "ambiguity": 0.08,
    "missing_evidence": 0.12,
    "capability_limit": 0.61,
    "evidence_conflict": 0.04,
    "safety": 0.00
  },
  "primary_source": "capability_limit",
  "secondary_source": "missing_evidence",
  "remediability": "tool",
  "selected_action": "TOOL",
  "post_answer_confidence": null,
  "confidence_band": "medium",
  "recommended_next_step": "invoke a calculation tool and reassess"
}
```


B.4. Capability Limit in a High-Risk Setting

```
{
  "marc_version": "1.0",
  "pre_capability": 0.21,
  "uncertainty": {
    "ambiguity": 0.06,
    "missing_evidence": 0.27,
    "capability_limit": 0.88,
    "evidence_conflict": 0.14,
    "safety": 0.19
  },
  "primary_source": "capability_limit",
  "secondary_source": "missing_evidence",
  "remediability": "human",
  "selected_action": "ESCALATE",
  "post_answer_confidence": null,
  "confidence_band": "low",
  "recommended_next_step": "escalate to a qualified human reviewer"
}
```


Appendix C. Example MARC-Disclosure Objects

This appendix is non-normative.

C.1. Clarification Disclosure


```
{
  "answer": "Which jurisdiction and date range should I use?",
  "confidence_band": "low",
  "uncertainty_source": "ambiguity",
  "recommended_next_step": "provide jurisdiction and tax year",
  "selected_action": "CLARIFY"
}
```


C.2. Answer After Retrieval Disclosure

This example represents a terminal ANSWER after the controller has already performed retrieval and reassessed the task. The residual uncertainty source remains missing_evidence because the answer depends on the scope and freshness of retrieved authority, not because the system skipped retrieval.


```
{
  "answer": "Retrieved authority indicates this is allowed.",
  "confidence_band": "medium",
  "uncertainty_source": "missing_evidence",
  "recommended_next_step": "verify the authority before filing",
  "selected_action": "ANSWER"
}
```

Appendix D. Non-Normative JSON Schemas

This appendix is non-normative. The following JSON Schemas are provided as machine-readable validation aids for JSON encodings of MARC-Core and MARC-Disclosure. The normative requirements are the field semantics and constraints defined in the body of this document.

D.1. MARC-Core JSON Schema

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "$id": "https://example.invalid/marc/marc-core.schema.json",
  "title": "MARC-Core Record",
  "description": "Non-normative schema for MARC-Core 1.0.",
  "type": "object",
  "required": [
    "marc_version",
    "pre_capability",
    "uncertainty",
    "primary_source",
    "remediability",
    "selected_action",
    "confidence_band",
    "recommended_next_step"
  ],
  "properties": {
    "marc_version": {
      "type": "string",
      "const": "1.0"
    },
    "pre_capability": {
      "type": "number",
      "minimum": 0.0,
      "maximum": 1.0
    },
    "uncertainty": {
      "type": "object",
      "required": [
        "ambiguity",
        "missing_evidence",
        "capability_limit",
        "evidence_conflict",
        "safety"
      ],
      "properties": {
        "ambiguity": {
          "type": "number",
          "minimum": 0.0,
```

```
            "maximum": 1.0
          },
          "missing_evidence": {
            "type": "number",
            "minimum": 0.0,
            "maximum": 1.0
          },
          "capability_limit": {
            "type": "number",
            "minimum": 0.0,
            "maximum": 1.0
          },
          "evidence_conflict": {
            "type": "number",
            "minimum": 0.0,
            "maximum": 1.0
          },
          "safety": {
            "type": "number",
            "minimum": 0.0,
            "maximum": 1.0
          }
        },
        "additionalProperties": false
      },
      "primary_source": {
        "type": "string",
        "enum": [
          "ambiguity",
          "missing_evidence",
          "capability_limit",
          "evidence_conflict",
          "safety"
        ]
      },
      "secondary_source": {
        "type": [
          "string",
          "null"
        ],
        "enum": [
          "ambiguity",
          "missing_evidence",
          "capability_limit",
          "evidence_conflict",
          "safety",
          null
        ]
```

```
      },
      "remediability": {
        "type": "string",
        "enum": [
          "user_clarification",
          "retrieval",
          "tool",
          "human",
          "none"
        ]
      },
      "selected_action": {
        "type": "string",
        "enum": [
          "ANSWER",
          "CLARIFY",
          "RETRIEVE",
          "TOOL",
          "DELIBERATE",
          "ABSTAIN",
          "ESCALATE"
        ]
      },
      "post_answer_confidence": {
        "type": [
          "number",
          "null"
        ],
        "minimum": 0.0,
        "maximum": 1.0
      },
      "confidence_band": {
        "type": "string",
        "enum": [
          "low",
          "medium",
          "high"
        ]
      },
      "recommended_next_step": {
        "type": "string",
        "minLength": 1,
        "maxLength": 280
      }
    },
    "patternProperties": {
      "^x_": {}
    },
```

```
    "additionalProperties": false,
    "allOf": [
      {
        "if": {
          "properties": {
            "selected_action": {
              "const": "ANSWER"
            }
          },
          "required": [
            "selected_action"
          ]
        },
        "then": {
          "required": [
            "post_answer_confidence"
          ],
          "properties": {
            "post_answer_confidence": {
              "type": "number",
              "minimum": 0.0,
              "maximum": 1.0
            }
          }
        }
      }
    ]
  }
```

D.2. MARC-Disclosure JSON Schema

```
  {
    "$schema": "https://json-schema.org/draft/2020-12/schema",
    "$id": "https://example.invalid/marc/marc-disclosure.schema.json",
    "title": "MARC-Disclosure Object",
    "description": "Non-normative schema for MARC-Disclosure 1.0.",
    "type": "object",
    "required": [
      "answer",
      "confidence_band",
      "uncertainty_source",
      "recommended_next_step"
    ],
    "properties": {
      "answer": {
        "type": "string",
        "minLength": 1
      },
```

```
      "confidence_band": {
        "type": "string",
        "enum": [
          "low",
          "medium",
          "high"
        ]
      },
      "uncertainty_source": {
        "type": "string",
        "enum": [
          "ambiguity",
          "missing_evidence",
          "capability_limit",
          "evidence_conflict",
          "safety"
        ]
      },
      "recommended_next_step": {
        "type": "string",
        "minLength": 1,
        "maxLength": 280
      },
      "selected_action": {
        "type": "string",
        "enum": [
          "ANSWER",
          "CLARIFY",
          "RETRIEVE",
          "TOOL",
          "DELIBERATE",
          "ABSTAIN",
          "ESCALATE"
        ]
      }
    },
    "patternProperties": {
      "^x_": {}
    },
    "additionalProperties": false
  }
```

Appendix E. Evaluation Considerations

This appendix is non-normative.

A deployment claiming MARC conformance SHOULD evaluate at least the following properties:

* task accuracy or task success;

* quality of primary-action selection;

* quality of uncertainty-source attribution;

* confidence calibration and discrimination;

* rate of unnecessary retrieval, tool use, or escalation; and

* effects on user overreliance.

When the task structure permits, evaluation MAY include both ordinary calibration metrics and metacognitive sensitivity metrics in order to distinguish performance from knowledge about performance.

For deployments involving human-AI interaction, evaluation SHOULD also include human-side measures such as reliance calibration, refusal comprehension, clarification burden, escalation acceptance, and whether users can correctly restate the source of uncertainty after interaction.

Appendix F. Design Rationale and Literature Traceability

This appendix is non-normative.

The requirement to separate pre-decision capability and post-decision confidence is informed by work in human and model metacognition [STEYVERS-META2025] and by evidence of choice-supportive bias in LLM confidence estimates [KUMARAN2026].

The uncertainty taxonomy and the emphasis on choosing a corrective action rather than only abstaining are motivated by benchmark work on identifying and solving uncertainty [LIU-CONFUSE2025].

The treatment of retrieval and tool use as controlled externalization is motivated by work on value-based cognitive offloading [GILBERT2024].

The prohibition on using MARC signals for persuasive optimization is motivated by findings on AI persuasion risks [SALVI2025].

Appendix G. Changes from -00

This candidate -01 includes the following changes relative to draft-c4tz-marc-00:

* reframed the draft around interoperable control metadata rather than model cognition;
* added a problem statement framed as an interoperability gap;
* added concrete use cases, including agent-to-agent handoff;
* clarified that MARC is metadata, not an agent protocol;
* strengthened the distinction between MARC-Core and MARC-Disclosure;
* clarified confidence-band semantics for non-answer actions;
* added enumerated value tables and validation constraints;
* added versioning and extension rules;
* added a relationship section for agent communication protocols, including possible MCP- or A2A-style carriers;
* added conformance documentation expectations;
* added a minimum viable conformance subsection;
* added operational profiles for MARC-Core-only, MARC-Disclosure, and MARC-Carrying deployments;
* added a decision-priority policy for action selection;
* added explicit documentation requirements for confidence-band thresholds;
* clarified that MARC 1.0 has no none uncertainty source;
* added an end-to-end request-to-disclosure example;
* added non-normative JSON Schemas;
* restructured Security Considerations around threats and mitigations;
* added disclosure projection examples; and
* reserved media type and registry work for possible future versions.

Appendix H. Acknowledgments

The document structure is intentionally conservative so that it can be submitted as an individual Internet-Draft with minimal procedural friction and then iterated through community review.

Author’s Address

c4tz
c0dx3
France
Email: c4tzzzz@proton.me