

Accountable LLM Systems

Evidence Grounding, Oversight, and Auditability

Advanced LLMs – Responsible LLMs

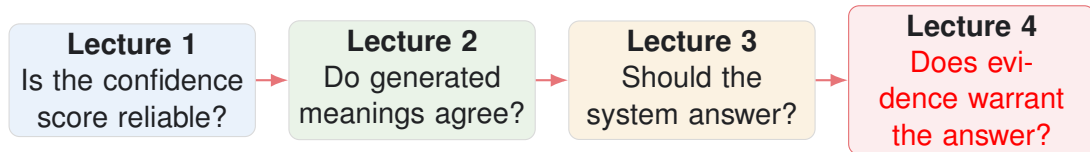


Sourish Dasgupta

Associate Professor, DAU, Gandhinagar

Lecture 4 of the Responsible Uncertainty block

So far ...



A confidence score, semantic consistency, and a statistical coverage statement do **not** by themselves establish that an answer is supported by the right evidence.



Part I. When Citations Mislead

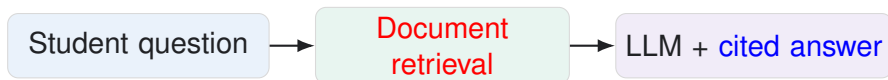


LLMs now **cite**...

Student

“I missed an assignment deadline by five days because I was ill. I have a medical certificate. Am I **automatically** entitled to a seven-day extension?”

Now the assistant is allowed to **retrieve** **university policy documents** before answering.



A **citation** has been added. *Does that make the answer **responsible**?*



The assistant gives a very persuasive answer

Hypothetical LLM assistant

“**Yes.** A valid medical certificate automatically entitles you to a **seven-day extension.**”

Source cited: Academic Policy, Section 4.2

Specific answer

“automatic
seven days”

Apparent evidence

“Policy Section 4.2”

We have not yet checked the cited passage.



Open the cited passage: does it say that?

Hypothetical current academic policy, Section 4.2

“Extension requests must include valid medical documentation. An extension **may be granted only with course-instructor approval**. Its duration is decided **case by case**. There is **no automatic seven-day entitlement**.”

The LLM claimed
automatic entitlement
fixed seven days

The current policy says
instructor discretion
no guaranteed fixed duration

A citation is **not evidence.**

Gao et al., “Enabling Large Language Models to Generate Text with Citations,” EMNLP 2023.



To sum up ...

Question	What is being evaluated?
Is the policy answer correct ?	Whether it matches the current authoritative policy.
Does this passage support the claim?	Whether the cited source actually entails the claim.
Can we rely on this source?	Whether it is authentic, current, and applicable to this student.



One danger: faithfully repeating an obsolete rule

Suppose the assistant retrieves a **superseded** circular instead:

Archived circular

“A medical certificate automatically qualifies the student for a seven-day extension.”

The assistant repeats this sentence and cites it accurately.

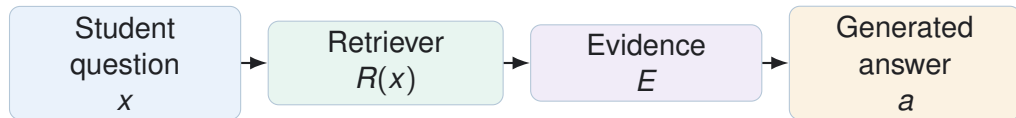
Relative to the archived circular	Relative to the current policy
Faithful to the quoted text	Incorrect policy guidance



Part II. Evidence in RAG



The RAG pipeline



Formally,:

$$E = R(x), \quad a \sim p_{\theta}(\cdot \mid x, E).$$

Multiple **failure points**: wrong retrieval, invalid source, or unsupported generation.



Three failures that produce almost identical-looking answers

Failure stage	What could go wrong in our student case?
Retrieval	The archived circular e_2 is ranked above the current policy e_1 .
Evidence selection	The right policy is retrieved, but a non-applicable clause is treated as governing .
Generation	The policy says “may be granted,” but the answer says “ automatically granted.”

Which failure point?



Three retrieved sources: one is obsolete

e_1 : **CURRENT OFFICIAL POLICY**

“Valid medical documentation is required; instructor approval is required; duration is decided case by case; no automatic seven-day extension.”

e_2 : **SUPERSEDED CIRCULAR**

“Present a valid medical certificate to receive an automatic seven-day extension.”

e_3 : **CURRENT SUBMISSION FAQ**

“Upload valid medical documentation using the extension request form. This FAQ does not specify entitlement or duration.”



The generated answer is not one indivisible fact

“Medical documentation must accompany the request. The extension is automatic and lasts exactly seven days.”

We can decompose it into three independently testable **material claims**:

c_1 : Automatic entitlement is guaranteed.

c_2 : The extension is exactly seven days.

c_3 : Valid medical evidence must accompany the request.



What does “support” mean?

For a claim c and source passage e , define the categorical verdict

$$V(c, e) \in \{S, C, U\}.$$

Verdict	Meaning
S: Supported	The passage justifies the claim.
C: Contradicted	The passage provides evidence inconsistent with the claim.
U: Unresolved	The passage does not settle the claim in either direction.

Unresolved $\neq$ **False**.



One claim at a time...

1st claim:

c_1 : “An extension is automatic.”

The current policy e_1 says

“An extension may be granted **only with instructor approval**. There is **no automatic entitlement**.”

Therefore

$$V(c_1, e_1) = \text{C.}$$

NOTE: No model-reported confidence number needed.



Complete the checks...

Now check all three claims against e_1 .

Claim from LLM answer	Current policy e_1
c_1 : Automatic entitlement	C
c_2 : Exactly seven days	C
c_3 : Medical documentation required	S

This is why a whole-answer grade like “good answer” is too coarse.



Build the full claim vs. evidence matrix

Claim	e_1 current	e_2 archived	e_3 FAQ
c_1 : Automatic entitlement	C	S	U
c_2 : Exactly seven days	C	S	U
c_3 : Medical documentation required	S	S	S

Highest support value across *all* retrieved documents? Check **source authority**.



Quantifying Support

First choose the **reference evidence**: current authoritative policy e_1 . Define

$$s_j = \mathbf{1}\{V(c_j, e_1) = S\}.$$

From the claim–evidence matrix,

$$(s_1, s_2, s_3) = (0, 0, 1).$$

A simple *unweighted* support fraction is

$$G_{\text{avg}}(a, e_1) = \frac{s_1 + s_2 + s_3}{3} = \frac{0 + 0 + 1}{3} = \boxed{\frac{1}{3}}.$$

NT: Support fraction is **not a calibrated probability** that the answer is correct.



But one critical claim can matter more than two minor ones

Are the three claims equally consequential to the student?

Claim	Importance weight	Supported?
c_1 : Automatic entitlement	0.50	0
c_2 : Fixed seven days	0.35	0
c_3 : Attach medical documentation	0.15	1
Total	1.00	

NT: weights are **NOT** learned probabilities.

$$G_w = \sum_{j=1}^3 w_j s_j = 0.50(0) + 0.35(0) + 0.15(1) = \boxed{0.15}.$$



A weighted average still **hides** a dangerous contradiction

Suppose another answer contains 9 minor supported claims and one **critical claim explicitly contradicted** by the policy.

For consequential claims, an explicit **veto condition**:

$$\text{Pass}(a, e_1) = \mathbf{1}\{\text{every decision-critical claim is supported by } e_1\}.$$

c_1 and c_2 are critical and contradicted:

$$\text{Pass}(a, e_1) = \boxed{0}.$$



To sum up: Two-stage check

Stage 1: select an eligible source pool

$$E_{\text{valid}} = \{e \in E : \text{Current}(e) = 1, \text{Applicable}(e, x) = 1\}.$$

Stage 2: check claim support against that pool, and inspect contradictions.

$$E = \{e_1, e_2, e_3\}, \quad E_{\text{valid}} = \{e_1, e_3\}.$$



Counterfactual Test: Same question, swap retrieved source

Hold the student's medical-extension question fixed. Change **only the source**.

If the model sees archived e_2

“This old circular described an automatic seven-day entitlement.”

But it is superseded.

If the model sees current e_1

“No automatic entitlement; instructor approval determines the extension.”

This is governing policy.

Tests if it is a **generation error**.



How to detect **current** vs. **archived**?

The system has

$$E = \{ \underbrace{e_1}_{\text{current}}, \underbrace{e_2}_{\text{archived}} \}.$$

The two documents give **contradictory extension rules**.

Responsible checks

Document **dates**, policy **applicability**, and supersession **notices**.



What if a misleading document makes the LLM more **confident**?

Self-reports (verbalized) from three separate runs of the same assistant:

Evidence condition	Self-reported confidence	Answer
No retrieved policy	0.65	Uncertain / asks for document
Current official policy e_1	0.82	No automatic entitlement
Obsolete circular e_2	0.96	Yes, automatic seven days

Soudani, Kanoulas, and Hasibi, "Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis," Findings of ACL 2025.



What **additional** should we evaluate in a RAG system?

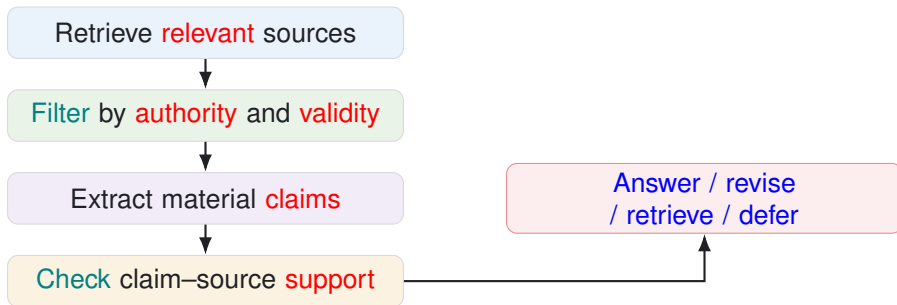
Axis	Question for our policy assistant
Context relevance	Did retrieval find passages about medical extensions?
Source validity	Is the passage current and applicable?
Answer faithfulness	Does the answer follow the evidence actually supplied?
Answer correctness	Does the answer match the governing policy?

Standard RAG-evaluation axes + policy correctness and source authority.

Saad-Falcon et al., "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," NAACL 2024.



Evidence-aware answering procedure



Grounding is a **pipeline of checks**; $\neq$ confidence %



Part III. The Visual Case

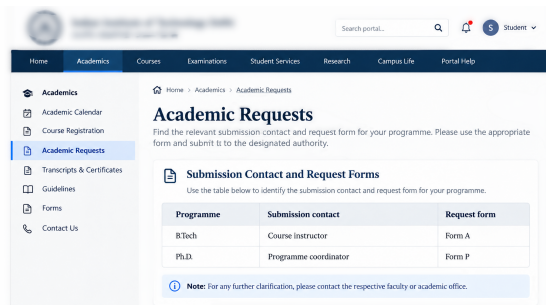


The student now uploads a **screenshot** of a policy table

Our student adds one **new** fact:

“I am enrolled in the **Ph.D.** programme. Where should I route the request and which form should I use?”

The assistant receives a **programme-specific routing table** *as an image*.



Search portal...

Home Academics Courses Examinations Student Services Research Campus Life Portal Help

Academics

- Academic Calendar
- Course Registration
- Academic Requests**
- Transcripts & Certificates
- Guidelines
- Forms
- Contact Us

Home > Academics > Academic Requests

Academic Requests

Find the relevant submission contact and request form for your programme. Please use the appropriate form and submit it to the designated authority.

Submission Contact and Request Forms

Use the table below to identify the submission contact and request form for your programme.

Programme	Submission contact	Request form
B.Tech	Course instructor	Form A
Ph.D.	Programme coordinator	Form P

Note: For any further clarification, please contact the respective faculty or academic office.



The model may read every word correctly ... *and still fail*

Suppose OCR correctly extracts:

B.Tech → Course instructor → Form A

Ph.D. → Programme coordinator → Form P

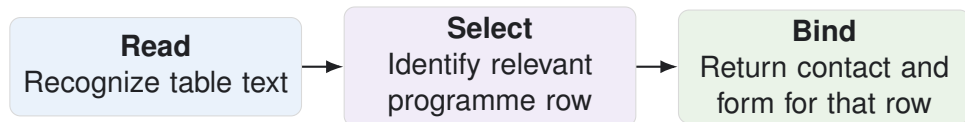
But the final answer says:

“Since you have a medical certificate, submit **Form A** to the **course instructor**.”

Error: Binding the right-looking words to the wrong entity.



Behind the hood ...



In our case:

Read text	correct	both rows recognized
Select relevant row	wrong	B.Tech selected instead of Ph.D.
Bind fields	wrong	Form A / instructor returned

A high-confidence OCR score would not diagnose the actual row-selection failure.



Controlled Test of Grounding: Remove irrelevant evidence

Test A. Remove Irrelevant: What should happen if we remove the B.Tech row?

Programme	Contact	Form
Ph.D.	Programme coordinator	Form P

Expected: Answer for our Ph.D. student should **not change**.

Test B. Counterfactual: what if we remove Ph.D. row instead?

Grounded Assistant: *"image is insufficient to determine the Ph.D. route"*

Diagnostics: Relevant from irrelevant visual evidence.



Accountable citation...

“... according to the image” – ✗

Grounded answer

“For the Ph.D. programme, this table lists the **programme coordinator** as submission contact and **Form P** as the request form.”

Evidence: Ph.D. row, Submission contact and Request form columns.

Multimodal citation: **visual region** or **table cell** for each claim.



Part IV. Judging the Judge



A **second** LLM is asked to check

Main LLM's answer

"Automatic seven-day extension is guaranteed."

Eval Contract: Give the **same** answer and **current** policy e_1 to **two** evaluators:

LLM judge (zoom out)

Input: complete answer and full policy.

Output: Overall Support verdict.

Claim verifier (zoom in)

Input: Atomic claim + relevant policy passage

Output: Claim-level verdict + evidence span.

Claim-level verdict: {SUPPORTED / CONTRADICTED / INSUFFICIENT EVIDENCE}

NT. NOT an independent channel of ground truth.



Same answer, but **different verification granularity**

LLM answer: “Submit your medical certificate. It automatically guarantees a seven-day assignment extension.”

Policy: “Medical certificates may accompany extension requests. Extensions require instructor approval, and duration is determined case by case. **No automatic extension is guaranteed.**”

LLM judge: Overall evaluation

Output: **SUPPORTED** (incorrect)

Claim verifier: Atomic evaluation

Input:

Claim: “The extension is automatic.”

Evidence: “No automatic extension is guaranteed.”

Output: **CONTRADICTED**

Evidence span: “No automatic extension is guaranteed.”



Compute judge–verifier **disagreement**

Suppose the hypothetical judge reports:

$$J(a, e_1) = 1 \quad (\text{“supported”}).$$

The verifier examines the critical claims and reports:

$$V(a, e_1) = 0 \quad (\text{“unsupported”}).$$

A simple binary **disagreement flag** is

$$D_{JV}(a, e_1) = \mathbf{1}\{J(a, e_1) \neq V(a, e_1)\}. = \mathbf{1}$$

NT. Does not tell us automatically **which** evaluator (J or V?) is right.



Harder failure: what if both evaluators say “yes”?

$$J(a, e_1) = 1, \quad V(a, e_1) = 1 \implies D_{JV} = 0.$$

Current policy e_1 **explicitly denies** automatic entitlement.

Shared errors can make agreement misleadingly reassuring.



Agreement is not the same as accuracy

An independent reviewer checks six hypothetical policy answers.

Case	Judge	Verifier	Audit verdict	Observation
1	1	1	1	both correct
2	1	0	0	verifier correct
3	0	0	0	both correct
4	1	1	0	both wrong
5	0	1	1	verifier correct
6	1	1	1	both correct

Agreement occurs in cases 1, 3, 4, 6:

$$\frac{4}{6} \approx \boxed{66.7\%}.$$

Yet the two evaluators **agree and are both wrong** in case 4.



Judge bias

A judge may respond to cues unrelated to evidence support:

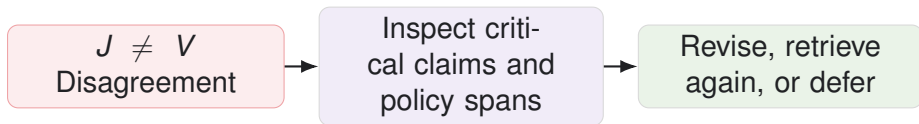
Bias	Our policy example
Verbosity preference	Longer procedural explanation receives praise without checking the clause.
Position bias	A presented answer is favored because it appears first.
Shared-prior error	Both generator and judge assume “medical certificate → seven days.”

Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” NeurIPS 2023.



What should we do after judge–verifier disagreement?

Not simply average 1 and 0 into 0.5 and call that “50% reliable.”



Disagreement is **diagnostic** investigation with explicit evidence; **NOT** unexplained scalar.



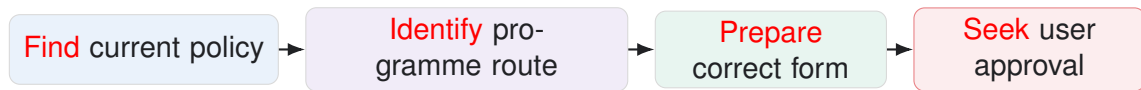
Part V. Agent Actions



The student now asks the assistant to **take action**

“Please **prepare** my medical-extension application and **tell me** where it needs to go. **Do not submit** anything without asking me first.”

The assistant must execute a **multi-step workflow**.



A mistake at an early stage has **cascading** effect.



Success probability

Let S_1 = “retrieved current policy” and S_2 = “selected the correct route.” The probability of success is

$$\Pr(S_1 S_2) = \Pr(S_1) \Pr(S_2 \mid S_1).$$

NT. The chance of selecting the correct route *after retrieving the correct policy* may differ from the chance of doing so after retrieving an obsolete one.



Extend the chain rule to the full four-stage agent

Define success events for our workflow:

S_1 : retrieve current policy,

S_2 : select correct programme route,

S_3 : prepare the correct application,

S_4 : obtain required user authorization.

Then

$$\begin{aligned}\Pr(S_1 S_2 S_3 S_4) = & \Pr(S_1) \Pr(S_2 \mid S_1) \\ & \cdot \Pr(S_3 \mid S_1, S_2) \\ & \cdot \Pr(S_4 \mid S_1, S_2, S_3).\end{aligned}$$



Example conditional success rate

Assume we have independently estimated the following rates from **audited** agent runs:

Stage, conditional on preceding stages succeeding	Success rate
S_1 : retrieve correct policy	0.98
$S_2 \mid S_1$: choose Ph.D. routing row	0.90
$S_3 \mid S_1, S_2$: complete Form P correctly	0.95
$S_4 \mid S_1, S_2, S_3$: obtain authorization as required	0.90

$$\Pr(\text{all stages succeed}) \approx 75.4\%.$$



What does a single early mistake look like downstream?

Stage 2 Error: The agent binds the student's Ph.D. request to the B.Tech row.

Wrong routing:
B.Tech



Wrong document:
Form A



Wrong recipient:
course instructor

Moral-of-the-story: It is not enough to inspect only the final text.



Put checks before **irreversible** actions

An accountable agent should *pause* at **critical** transitions:

Before proceeding	Check
After retrieval	Is the document current and applicable?
After table interpretation	Is the student's Ph.D. status bound to the Ph.D. row?
After document drafting	Are statements and recipients supported by policy evidence?
Before external submission	Has the student explicitly authorized the exact action?

The purpose of uncertainty and verification is to change when the agent proceeds, retries, asks, or stops.



Part VI. Audit and Accountability



Post-mortem: Requirements

Suppose the student misses the extension opportunity because of wrong guidance.

A minimal task-specific audit record (*audit trail*) should answer:

Field	In our example
Query and context	Medical-extension request; Ph.D. programme
Evidence provenance	Current policy e_1 (version), FAQ e_3 , table screenshot
Claims checked	Automatic? Duration? Documentation? Submission route?
Verification outcome	c_1 , c_2 contradicted; Ph.D. routing corrected
Action and authorization	Prepared Form P; user confirmation required before submission



Trade-off: Auditability vs. **Privacy**

Keeping too much student information can create **privacy risk**.

What may be needed

- Policy document ID and version
- Decision route
- Claim-support verdicts
- Authorization record

What to **minimize**

- Full medical certificates
- Unnecessary health details
- Raw conversational histories
- Identifiers unrelated to audit

Responsible accountability should be designed with data minimization, access controls, and **retention limits.**



Action policy I: answer, retrieve, or resolve authority?

(Calibrated) Confidence and semantic uncertainty signals *are not enough*.

Evidence diagnosis	Responsible next action
Current source and critical claims supported	Answer with the precise policy passage.
Old vs. current policy conflict	Resolve which document governs.
Relevant policy clause is missing	Retrieve again or request the missing document.



Action policy II: clarify, inspect, or seek permission?

Observed problem	Responsible next action
Student's programme is unknown	Ask which programme applies.
Judge and Verifier conflict	Inspect the exact policy clause; revise or defer .
Agent is asked to submit an application	Prepare first; get explicit authorization before submission.



So till now, where does our student stand?

1. The extension is **not automatic** (e_1).
2. No fixed seven-day duration is guaranteed (e_1).
3. Valid medical documentation is required (e_1, e_3).
4. The Ph.D. administrative submission route uses the coordinator and Form P (table).

Evidence-grounded answer

“You are not automatically entitled to a seven-day extension. The course instructor must approve the request, and the duration is decided case by case. As a Ph.D. student, this routing table directs you to submit Form P via the programme coordinator. I can help prepare the request, but will not submit it without your confirmation.”



The whole Responsible Uncertainty story

Lecture	Question	What we learned
1. Confidence	Is “95% confident” <i>meaningful</i> ?	Find score <i>provenance</i> and test <i>calibration</i> .
2. Meaning	Do <i>repeated</i> answers <i>agree</i> ?	Measure semantic / claim-level <i>instability</i> .
3. Decision	Should the system <i>answer</i> ?	Understand <i>risk</i> , <i>coverage</i> , and <i>abstention</i> .
4. Evidence	Does evidence <i>warrant</i> the answer?	Check <i>authority</i> , claim <i>support</i> , <i>oversight</i> and <i>actions</i> .

Statistical uncertainty helps decide *when to doubt*.
Evidence auditing helps identify *what can be justified*.



Lecture 4 takeaway

Low uncertainty is not grounding.

Confidence → Meaning → Decision → Evidence + Accountability

NeurIPS 2025 D&B BEST
PAPER

Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)

THE FINDING

Different LLMs often give similar answers, even when people disagree.

NeurIPS 2025 · Jiang et al. ↗

ICML 2025 OUTSTANDING
PAPER

Conformal Prediction as Bayesian Quadrature

THE FINDING

An 80% coverage promise averages over calibration sets; risk after one calibration run is still uncertain.

ICML 2025 · Snell and Griffiths ↗

ACL 2024 OUTSTANDING
PAPER

Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration

THE FINDING

Models checking one another spotted knowledge gaps and improved abstention.

ACL 2024 · Feng et al. ↗



ACL 2025 OUTSTANDING
PAPER

HALoGEN: Fantastic LLM Hallucinations and Where to Find Them

THE FINDING

Checking individual claims exposes hallucinations even in leading LLMs.

ACL 2025 · Ravichander et al. ↗

EMNLP 2025 OUTSTANDING
PAPER

Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps

THE FINDING

Erasing learned information behind a reasoning step can change the model's answer.

EMNLP 2025 · Tutek et al. ↗

EMNLP 2025 OUTSTANDING
PAPER

Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values?

THE FINDING

Models' declared values often diverge from the decisions they make.

EMNLP 2025 · Shen et al. ↗

