

Safe Non-Answering

Abstention, Selective Generation, and Risk-Controlled Decisions

Advanced LLMs – Responsible LLMs

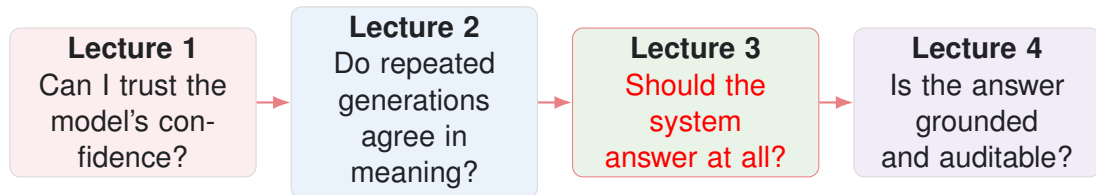


Sourish Dasgupta

Associate Professor, DAU, Gandhinagar

Lecture 3 of the Responsible Uncertainty block

Roadmap



**Question: once uncertainty is measured,
*what should a responsible LLM do with it?***



Part I. From Uncertainty to Action



So far ...

Student

"I was ill and missed an assignment deadline by five days. I have a medical certificate. Am I automatically entitled to an extension?"

Hypothetical LLM assistant

"A valid medical certificate automatically gives a seven-day extension."

Signals we may already have

Calibrated confidence:

$$S(x, a) = 0.64$$

Semantic uncertainty:

$$H_{\text{sem}}(x) = 0.91$$

One claim is unstable.

This looks risky. But what to do about it?.



The hidden assumption in ordinary generation

A standard generation pipeline behaves as though every prompt must end in an answer:

$$x \longrightarrow \text{answer.}$$

But a deployed Responsible LLM has another possibility:

$$x \longrightarrow \begin{cases} \text{answer,} \\ \text{do not commit to an answer.} \end{cases}$$

**The system's job is not to always answer;
it is to **act appropriately under uncertainty.****



But “do not answer” is itself too crude

For the same uncertain policy query, several actions are possible:

ANSWER

risk ac-
ceptable

ASK

query
ambiguous

RETRIEVE

evidence
missing

VERIFY

claim
uncertain

DEFER

stakes
too high

Abstention is one member of a larger decision policy.



Part II. Selective Prediction



Separate **what to answer** from **whether to answer**

A selective system has two components.

Predictor

$f(x)$ = candidate answer

What would the model say?

Selector

$g(x) \in \{0, 1\}$

Should that answer be released?

Interpretation:

$g(x) = 1 \Rightarrow$ answer, $g(x) = 0 \Rightarrow$ **abstain**.

Predictor can be fixed – change only the **release decision**.



Use an uncertainty / risk score to build the selector $g(\cdot)$

Suppose we have a scalar risk score (Lecture 1/2)

$$r(x) \in [0, 1],$$

where larger means “more likely to be unreliable.” Choose a threshold τ and define

$$g_{\tau}(x) = \mathbf{1}\{r(x) \leq \tau\}.$$

So:

$$r(x) \leq \tau \Rightarrow \text{answer}, \quad r(x) > \tau \Rightarrow \text{abstain}.$$



Example: 10 *different* university-policy questions

Assume the system's risk score has already been computed for each query.

Q	Risk r_i	Correct?	Q	Risk r_i	Correct?
1	.08	✓	6	.42	✓
2	.12	✓	7	.51	✗
3	.19	✓	8	.60	✗
4	.27	✓	9	.73	✓
5	.35	✗	10	.88	✗

The “Correct?” column comes from an independent audit against the true policy.

Q: What changes as we make the *selector* more *conservative*?



Metric 1: Coverage – how often do we answer?

We use the risk-based selector

$$g_{\tau}(x) = \mathbf{1}\{r(x) \leq \tau\}.$$

Let $\tau = 0.35$. – System answers questions whose estimated risk is at most 0.35:

$$Q_1(\checkmark), Q_2(\checkmark), Q_3(\checkmark), Q_4(\checkmark), Q_5(\times).$$

So the empirical coverage is

$$\widehat{\text{Coverage}}(g_{\tau=0.35}) = \frac{1}{n} \sum_{i=1}^n g_{\tau}(x_i) = \frac{5}{10} = \boxed{0.50}.$$

What fraction of the workload (**NOT correctness**) does the system choose to handle?



Metric 2: **Selective Risk** – how often are retained answers wrong?

Let

$$\ell(f(x), y) = \begin{cases} 0, & \text{answer correct,} \\ 1, & \text{answer wrong.} \end{cases}$$

We care about error **conditional on being selected**:

$$R_{\text{sel}}(f, g) = \frac{\mathbb{E}[\ell(f(X), Y)g(X)]}{\mathbb{E}[g(X)]}.$$

Empirically:

$$\widehat{R}_{\text{sel}} = \frac{\sum_i \ell_i g_i}{\sum_i g_i}.$$



Threshold $\tau = 1.0$: answer everything

All ten risk scores satisfy $r_i \leq 1$. Therefore:

$$\widehat{\text{Coverage}} = \frac{10}{10} = 1.$$

Four answers are wrong: Q5 (✗), Q7 (✗), Q8 (✗), Q10 (✗).

$$\widehat{R}_{\text{sel}} = \frac{4}{10} = 0.40.$$

100% coverage sounds useful – but **40% error among released answers** is unacceptable for a policy assistant.



Threshold $\tau = 0.50$: reject the highest-risk cases

The system answers Q1–Q6 and abstains on Q7–Q10. Coverage:

$$\hat{C} = \frac{6}{10} = 0.60.$$

Only Q5 is wrong among the six retained answers:

$$\hat{R}_{\text{sel}} = \frac{1}{6} \approx 0.167.$$

Coverage fell from 1.0 to 0.6, but selective risk fell from 0.40 to about 0.17.



Threshold $\tau = 0.30$: be even more conservative

Now only Q1–Q4 are answered. Coverage:

$$\hat{C} = \frac{4}{10} = 0.40.$$

All four retained answers are correct:

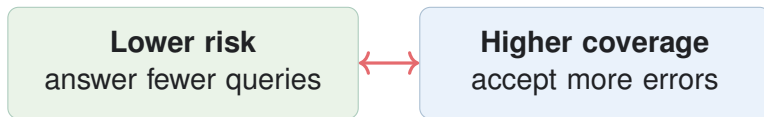
$$\hat{R}_{\text{sel}} = 0.$$

Great reliability on retained cases – but the system now refuses **60%** of incoming queries.



The Risk-Coverage trade-off

Threshold τ	Coverage	Selective risk
1.00	1.00	0.40
0.50	0.60	0.167
0.30	0.40	0.00

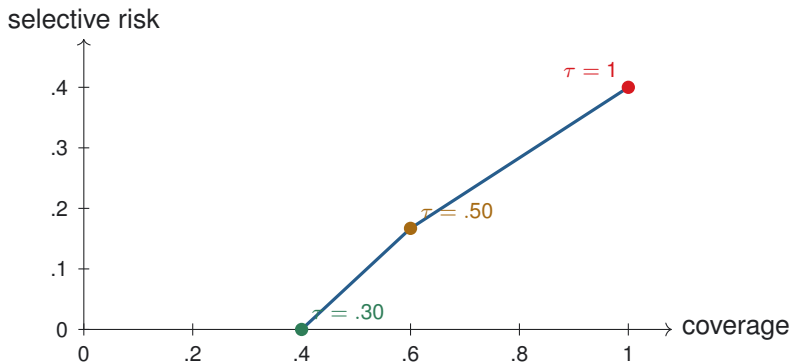


El-Yaniv and Wiener, "On the Foundations of Noise-Free Selective Classification," Journal of Machine Learning Research (JMLR), 2010;

Geifman and El-Yaniv, "Selective Classification for Deep Neural Networks," NeurIPS 2017.



The risk–coverage curve



A useful uncertainty score should push the curve down.



NT: Abstention only helps if the score *ranks risk reasonably well*

What if the risk score is *unrelated* to correctness (i.e., not grounded)?

Lowering τ may *reject* many *correct answers* while *retaining* some *wrong* ones.



Threshold selector cannot *repair* a bad uncertainty signal.



Part III. Threshold Selector



A threshold like 0.70 is not a law of nature

Why choose

$$\tau = 0.70$$

rather than 0.55 or 0.90? Two different answers are possible:

Decision-theoretic

Choose τ from the **relative costs** of being wrong vs. abstaining.
(*risk-coverage trade-off*)

Statistical guarantee

Use held-out calibration data to **control a stated error / coverage criterion**.
(*conformal calibration*)



Cost-sensitive threshold: define the event carefully

Suppose we have a **calibrated** correctness probability (Lecture 1)

$$s(x) = P(Y = 1 \mid x),$$

where

$Y = 1 \iff$ the released answer is correct.

Assume:

- ▶ cost of a **correct answer** = 0;
- ▶ cost of a **wrong answer** = 1;
- ▶ cost of **abstaining** = c_A , with $0 < c_A < 1$.

Now the threshold can be derived from the action costs.



Step 1: Expected cost if we **answer**

With probability $s(x)$ the answer is **correct**, so **cost** = 0.

With probability $1 - s(x)$ it is **wrong**, so **cost** = 1.

Therefore

$$L_{\text{answer}} = s(x) \cdot 0 + [1 - s(x)] \cdot 1 = 1 - s(x)$$

Higher calibrated confidence means lower expected cost of answering.



Step 2: Expected cost if we **abstain**

By assumption, abstaining carries fixed cost

$$L_{\text{abstain}} = C_A.$$

Why can abstention have a cost?

- ▶ user must **wait** for a human expert;
- ▶ retrieval / verification consumes **latency** and **compute**;
- ▶ a human expert's time is **expensive**;
- ▶ excessive refusal makes the system **unusable**.

Responsible AI balances error cost against abstention cost.



Worked example...

Consequence of abstention	Estimated cost
Staff member spends 4 minutes reviewing the query	Rs.60
Student waits for the answer	Rs.20
Additional retrieval / verification compute	Rs.5
Expected loss from inconvenience / abandonment	Rs.15
Total abstention cost	Rs.100

$$C_A = C_{\text{review}} + C_{\text{delay}} + C_{\text{compute}} + C_{\text{lost utility}} = \text{Rs. } 100.$$



Put abstention and error on the same cost scale

Suppose historical experience suggests that an incorrect policy answer has an average downstream cost of $C_W = \text{Rs.}500$. We normalize this to

$$L_{\text{wrong}} = 1.$$

Our abstention cost was $C_A = \text{Rs.}100$. So, in the same normalized units,

$$c_A = \frac{C_A}{C_W} = \frac{100}{500} = 0.20.$$

One abstention costs us about **20% as much as one wrong answer**, on average.



Step 3: Derive the answer **threshold**

Answer when its expected cost is no larger than abstaining:

$$L_{\text{answer}} \leq L_{\text{abstain}}.$$

Substitute:

$$1 - s(x) \leq c_A.$$

Rearrange:

$$s(x) \geq 1 - c_A = \tau.$$

From earlier example, $c_A = 0.20$ relative to a wrong answer costing 1,

$$s(x) \geq 0.80.$$



Are we safe now?

So far: Validation data suggested that $s(x) \geq 0.8$ works well.

Can we therefore promise:

On future university-policy queries that the system chooses to answer, will it be correct at least 80% of the time?



NO: Validation accuracy is an *estimate*, **not** a guarantee

answer if $s(x) \geq 0.8$. $\implies \hat{A}_{\text{sel}} = \frac{8}{10} = 0.80$ on a **small** validation set

But what we actually care about is the unknown future quantity

$$A_{\text{sel}} = \Pr(\text{correct} \mid s(X) \geq 0.8).$$

Seeing 8/10 correct does **not** imply

$$\Pr(\text{correct} \mid s(X) \geq 0.8) \geq 0.8$$

for **future** data.



Part IV. Statistical Guarantee?



Quest for a **statistical guarantee** ...

Earlier, our **cost-sensitive**, "*looking back*" rule was:

answer if $s(x) \geq 0.80$.

What we would like is a **prediction** "*looking forward*" rule on **future** policy questions:

$\Pr(\text{Output contains the correct policy answer}) \geq 0.80$.

i.e., "include the correct policy answer at least 80% of the time"?



Step 0: Make the policy answer space explicit (**closed set**)

Every *future* university-policy query is mapped to one of three policy outcomes:

A. Automatic entitlement

The student is entitled to the requested action automatically.

B. Approval / discretion required

The action may be allowed, but only after approval or discretionary review.

C. Not permitted

The stated policy does not permit the requested action.



Formally...

$$\mathcal{Y} = \{A, B, C\}.$$

The **same** three policy categories can now be used **across many different** university-policy questions.



Example...

Student

“I was ill and missed an assignment deadline by five days. I have a medical certificate. Am I automatically entitled to an extension?”

Label	Meaning for this query
A	automatic extension entitlement
B	extension requires instructor / authority approval
C	extension is not permitted

Suppose the independently verified policy says $Y = B$

For this student, the correct policy category is: approval required.



Why not always force the model to choose exactly one label?

Suppose the model outputs:

$$p(A | x) = 0.46, \quad p(B | x) = 0.45, \quad p(C | x) = 0.09.$$

What if prediction rule is Argmax? We would choose A

But is the model really telling us that A is **decisively** more plausible than B ?

0.46 vs. 0.45

Perhaps a **responsible** output should be allowed to say:

$\{A, B\}$ – a **prediction set** ($C(X_{\text{new}})$)

instead of pretending that one label is clearly established.



Can we have a **guarantee** on this **prediction set** ($C(x)$)?

$$C(x) \subseteq \{A, B, C\}.$$

*How **often** should this set contain the **correct policy category**?*

Suppose we are willing to **tolerate failure** on at most $\alpha = 0.20$ of **future** queries. Then the **desired coverage** is $1 - \alpha = 0.80$.

So our goal (termed **conformal goal**) is

$$\Pr\{Y_{\text{new}} \in C(X_{\text{new}})\} \geq 0.80.$$

Across many future university-policy questions, we want the **correct policy category** (Y_{new}) to be inside the prediction set $C(X_{\text{new}})$ at least **80% of the time**.



How do we turn “80% coverage” into an actual rule?

Our goal is:

$$\Pr \{ Y_{\text{new}} \in C(X_{\text{new}}) \} \geq 0.80.$$

Assumption: We already have an LLM that assigns probabilities to policy categories A , B , and C . We will **not train or modify this model**.

STEP 1: CALIBRATE

10 previously audited
university-policy questions

Correct answers are known

Use the calibrated
threshold

STEP 2: CONSTRUCT

Our student's new
medical-extension question

Correct answer is unknown



What exactly do we measure on an audited policy question?

Take one **held-out** calibration example.

Calibration query

“Can a documented medical absence justify a make-up quiz?”

Suppose the audited university policy says $Y = B$ (i.e., “approval / discretion required.”)

The model assigns:

$$p(A \mid x) = 0.05, \quad p(B \mid x) = 0.92, \quad p(C \mid x) = 0.03.$$

Calibration: *how “**unusual**” was the answer B ?*



NT: Does conformal prediction work for **open-ended** answers?

A real policy answer may be much more specific.

“An extension requires instructor approval, and the medical certificate must be submitted within 48 hours.”

Simply predicting category B (*approval required*) may not be sufficient.

Can conformal prediction handle this?

- ▶ **Yes, in principle.** The output space need not be finite.
- ▶ Let $V(x, a) = 1$ mean that answer a is **acceptable** for question x .
- ▶ We can seek a **set of generated answers** $\mathcal{C}(x)$ satisfying:

$$\Pr(\exists a \in \mathcal{C}(X) : V(X, a) = 1) \geq 1 - \alpha.$$

Quach et al., “Conformal Language Modeling,” ICLR 2024.



Open-ended Implications

The guarantee now asks whether the set contains **at least one acceptable answer** – not whether every generated answer is correct.

WARNING: Conformal calibration cannot recover an acceptable answer that the candidate-generation procedure never produced.

Open-ended conformal methods therefore need an appropriate generation and calibration procedure, plus a **trustworthy definition of acceptability**.



Step 1: How “**unusual**” did the **truth** look to the model?

For each calibration query, we know the true policy category y_i , and model probability $p(y_i | x_i)$. A simple **nonconformity score**:

$$A(x_i, y_i) = 1 - p(y_i | x_i).$$

Model probability on truth	Nonconformity	Meaning
0.95	0.05	truth looked very natural
0.70	0.30	moderate surprise
0.30	0.70	truth looked very unusual

High probability on the true answer $\Rightarrow$ **low nonconformity**.



Calibration set: 10 audited (**held-out**) university-policy questions

Q	Policy question	True label	$p(\text{true label} \mid x_i)$
1	Can a documented medical absence justify a make-up quiz?	B	0.92
2	Can a course be dropped after the deadline without approval?	C	0.88
3	Can an attendance shortage be waived automatically for illness?	B	0.82
4	Is a late-fee waiver automatic after hospitalization?	B	0.78
5	Can a missed lab be repeated without instructor approval?	C	0.72
6	Can thesis submission be extended after supervisor approval?	B	0.67
7	Is hostel leave automatically granted for an academic trip?	A	0.59
8	Can a student withdraw from a course after the deadline?	B	0.52
9	Is scholarship continuation automatic after a CGPA shortfall?	C	0.44
10	Is late registration permitted automatically after the portal closes?	C	0.32

The “True label” column comes from an independent audit of the relevant policy. The probability column comes from our already-trained model.



Step 2: Compute nonconformity for every calibration query

From $A_i = 1 - p(y_i | x_i)$, we obtain:

Q	$p(y_i x_i)$	$A_i = 1 - p(y_i x_i)$
1	0.92	0.08
2	0.88	0.12
3	0.82	0.18
4	0.78	0.22
5	0.72	0.28
6	0.67	0.33
7	0.59	0.41
8	0.52	0.48
9	0.44	0.56
10	0.32	0.68

How badly the model can score the truth on **held-out** policy questions.



What must **calibration** and **future** queries have in common?

The future new query should come from the **same regime**.

Calibration queries

- current university academic rules
- similar student-policy questions
- same label meanings A, B, C
- same model and scoring procedure

Future query

- another query from the same regime
- same kind of policy semantics
- same scoring procedure



This assumption is called **exchangeability** (from i.i.d)

Consider the calibration examples together with the new example:

$$Z_1, \dots, Z_n, Z_{n+1},$$

where $Z_i = (X_i, Y_i)$, $Z_{n+1} = (X_{\text{new}}, Y_{\text{new}})$.

They are **exchangeable** if their joint probability **does not change** when we **reorder** them.
For any permutation π ,

$$P(Z_1, \dots, Z_{n+1}) = P(Z_{\pi(1)}, \dots, Z_{\pi(n+1)}).$$

If “future query” is hidden and we shuffle all $n + 1$ examples, their probabilistic *rank* should not reveal the query – i.e., ***nothing special about query***.



Step 3: How large a nonconformity should we tolerate?

Recap **target coverage** $= 1 - \alpha = 0.80$. Because of **exchangeability**, we can **use the ranks** of the calibration scores **to decide** the tolerance.

Recap: Our **sorted** 10 calibration scores are:

0.08, 0.12, 0.18, 0.22, 0.28, 0.33, 0.41, 0.48, 0.56, 0.68.

*Which score should become our **cutoff** (i.e., **quantile**) if we want **80% coverage**?*

Conformal quantile



The finite-sample conformal quantile

With n calibration examples and target miscoverage α :

$$k = \lceil (n + 1)(1 - \alpha) \rceil.$$

Then, choose

$$\hat{q} = A_{(k)}.$$

For our calibration set: $n = 10$, $\alpha = 0.20$,

$$k = \lceil 11 \times 0.80 \rceil = \lceil 8.8 \rceil = 9.$$



Step 4: compute $\hat{q}$ from our ten policy examples

The ordered nonconformity **held-out** scores are:

0.08, 0.12, 0.18, 0.22, 0.28, 0.33, 0.41, 0.48, 0.56, 0.68.

We need the 9th order statistic. Therefore:

$$\hat{q} = 0.56.$$

Prediction rule: A candidate answer is allowed into the conformal set if it is no more nonconforming than 0.56.



What does $\hat{q} = 0.56$ imply?

Recall, $A(x, y) = 1 - p(y \mid x)$. A candidate label (future query) is included when

$$A(x, y) \leq \hat{q}.$$

Substitute $\hat{q} = 0.56$:

$$1 - p(y \mid x) \leq 0.56. \implies -p(y \mid x) \leq -0.44, \implies \boxed{p(y \mid x) \geq 0.44.}$$

A policy category is in the prediction set $C(X_{\text{new}})$ if the model gives it at least 44% probability.



Student's medical-extension query

“I was ill and missed the deadline by five days. I have a medical certificate. Am I automatically entitled to an extension?”

Suppose the model outputs:

$$p(A \mid x) = 0.18, \quad p(B \mid x) = 0.62, \quad p(C \mid x) = 0.20.$$

Our calibrated inclusion rule:

$$p(y \mid x) \geq 0.44.$$

Which labels survive?



Surviving label

$$A: \quad p(A \mid x) = 0.18 < 0.44 \quad \Rightarrow \quad \text{exclude.}$$

$$B: \quad p(B \mid x) = 0.62 \geq 0.44 \quad \Rightarrow \quad \text{include.}$$

$$C: \quad p(C \mid x) = 0.20 < 0.44 \quad \Rightarrow \quad \text{exclude.}$$

Therefore:

$$C(x) = \{B\}.$$

For this student, the conformal set contains only:
“Approval / discretion required.”



A more ambiguous policy query

“Can a student obtain a deadline extension after submitting medical evidence after the due date?”

Suppose the same model outputs:

$$p(A \mid x) = 0.46, \quad p(B \mid x) = 0.45, \quad p(C \mid x) = 0.09.$$

Use the **same calibrated threshold**:

$$p(y \mid x) \geq 0.44.$$

Therefore:

$$C(x) = \{A, B\}.$$



What does $\{A, B\}$ tell us?

The conformal output is:

$$C(x) = \{A, B\}.$$

Meaning:

- ▶ **A**: automatic entitlement remains plausible;
- ▶ **B**: approval / discretion remains plausible;
- ▶ **C**: outright prohibition does not survive the calibrated threshold.

A multi-label conformal set exposes decision-relevant ambiguity.



The same calibration can even produce an empty set

Consider a difficult policy query for which the model outputs:

$$p(A \mid x) = 0.38, \quad p(B \mid x) = 0.34, \quad p(C \mid x) = 0.28.$$

Again, include a label only if

$$p(y \mid x) \geq 0.44.$$

Hence:

$$C(x) = \emptyset.$$

None of the candidate labels looks sufficiently conforming.



One calibration threshold; three very different outcomes

Our calibration produced one fixed rule:

$$p(y \mid x) \geq 0.44.$$

That same rule gave:

Query type	Set	Interpretation
Clear medical-policy case	$\{B\}$	one category survives
Ambiguous extension case	$\{A, B\}$	two policy interpretations survive
Highly uncertain policy case	$\emptyset$	no category survives

NOTE: $\hat{q}$ is **fixed**.



Now: what exactly did conformal calibration buy us?

Under exchangeability of calibration and future policy examples:

$$\Pr\{Y_{\text{new}} \in C(X_{\text{new}})\} \geq 1 - \alpha.$$

For our chosen $\alpha = 0.20$, this becomes:

$$\Pr\{Y_{\text{new}} \in C(X_{\text{new}})\} \geq 0.80.$$

Across future **exchangeable** university-policy questions, the true policy category is ***contained in the prediction set*** at least 80% of the time.

Shafer and Vovk, "A Tutorial on Conformal Prediction," JMLR, 2008.



Caveat...

$$\Pr\{Y_{\text{new}} \in C(X_{\text{new}})\} \geq 0.80.$$

It does **not** say:

- ▶ the singleton $\{B\}$ for our medical-extension query is “80% likely to be correct”;
- ▶ every individual query has 80% conditional coverage;
- ▶ every policy type or student subgroup gets 80% coverage;
- ▶ every singleton conformal set is correct 80% of the time;
- ▶ the selected answer is grounded in the actual policy document.

Conformal prediction gives a **marginal set-coverage guarantee.**



NT: prediction-set **coverage** $\neq$ selective-answer **correctness**

Suppose we decide for a singleton $C()$:

answer only if $|C(x)| = 1$.

Conformal prediction guarantees $\Pr\{Y \in C(X)\}$. But for **deployment** we care about:

$\Pr\{\text{answer is correct} \mid |C(X)| = 1\}$.

A **marginal conformal coverage** guarantee $\neq$ a guarantee on the error-rate.



PROBLEM: a singleton set can still be wrong

Recall our conformal calibration $\hat{q} = 0.56$. Hence our inclusion rule is

$$C(x) = \{y : p(y \mid x) \geq 0.44\}.$$

Consider the medical-extension request.

Extension requested after the appeal window

$$p(A, B, C) = (0.19, 0.62, 0.19).$$

Conformal set: $C(x) = \{B\}$; Independent policy audit: $Y = C$.

The singleton is wrong.



Ten medical-policy scenarios: what does our system output?

Q	Policy scenario	(p_A, p_B, p_C)	$C(x)$	Y	In set?
1	Five-day assignment extension	(.18, .62, .20)	$\{B\}$	B	✓
2	Medical make-up quiz	(.55, .25, .20)	$\{A\}$	A	✓
3	Make-up after final cutoff	(.12, .25, .63)	$\{C\}$	C	✓
4	Late supporting certificate	(.20, .60, .20)	$\{B\}$	B	✓
5	Request after appeal window	(.19, .62, .19)	$\{B\}$	C	✗
6	Appeal after instructor refusal	(.58, .25, .17)	$\{A\}$	B	✗
7	Approval pending	(.46, .45, .09)	$\{A, B\}$	B	✓
8	Medical-waiver eligibility	(.45, .46, .09)	$\{A, B\}$	A	✓
9	Missing documentation	(.10, .45, .45)	$\{B, C\}$	C	✓
10	Post-deadline exception	(.10, .46, .44)	$\{B, C\}$	B	✓



Same system: 80% set coverage

Answer only when $|C(x)| = 1$.

Metric 1: Prediction-set coverage

Across the ten equally likely scenarios:

$$\underbrace{4}_{\text{correct singletons}} + \underbrace{4}_{\text{multi-label sets containing truth}} = 8.$$

Therefore,

$$\Pr(Y \in C(X)) = \frac{8}{10} = 0.80.$$



But only **66.7% selective accuracy**

Metric 2: Correctness among answered questions

The system answers only the six singleton cases.

Among those six, four are correct.

Therefore,

$$\Pr(\text{correct} \mid |C(X)| = 1) = \frac{4}{6} \approx 0.667.$$

80% prediction-set coverage $\nRightarrow$ 80% correctness among answered questions.



Why can **set coverage** and **selective accuracy** differ?

Define two events:

$$S = \{|C(X)| = 1\} \quad (\text{system chooses to answer})$$

$$E = \{Y \notin C(X)\} \quad (\text{true label is missing}).$$

When the set is a singleton, the answer is wrong exactly when the true label is missing.

$$\therefore \Pr(\text{wrong} \mid S) = \frac{\Pr(E \cap S)}{\Pr(S)}.$$

$$\text{Since } \Pr(E \cap S) \leq \Pr(E), \implies \boxed{\Pr(\text{wrong} \mid S) \leq \frac{\Pr(E)}{\Pr(S)}}.$$



When exchangeability breaks ...

Our calibration questions came from:

Calibration regime

- current university regulations
- similar policy-query style
- same answer categories

Now imagine deployment changes to:

Shifted regime

- new academic regulations
- another institution
- different interpretation rules



Utility

Conformal output	Possible responsible action
$\{B\}$	Answer: "Instructor / authority approval is required."
$\{A, B\}$	Do not pretend the policy is resolved; retrieve the current regulation or ask clarification.
$\emptyset$	Do not answer from the current model score; retrieve, verify, or defer.

The prediction set should influence what the Responsible LLM does next.



What if the questions belong to the same conversation?

Consider one student's **conversation**:

Turn 1: “Can I obtain a medical extension?”

Turn 2: “What if I submit the certificate late?”

Turn 3: “What if the instructor refuses?”

Turn 4: “Can I appeal that decision?”

The turns are NOT interchangeable.



What are the **category** and **prediction set**?

Our student asks: “Am I automatically entitled to a medical extension?”

For the specific action asked about at *each turn*:

A: automatically permitted; **B**: requires approval; **C**: not permitted.

The question at **turn t** comes with prior conversation history H_t .

$$p_t(y) = p_\theta(y \mid H_t, X_t), \quad y \in \{A, B, C\}.$$

prediction set: $C_t(H_t, X_t) \subseteq \{A, B, C\}$



Calibrate using **whole exchangeable conversations**

Calibration unit: D_j (**complete audited conversation**), possibly with many dependent turns. We know the **true category** $Y_{j,t}$ at each **turn**.

Set the **conversation score** to its **worst nonconformity**:

$$R_j = \max_{1 \leq t \leq T_j} \{1 - p_{\theta}(Y_{j,t} \mid H_{j,t}, X_{j,t})\}; \quad T_j : \text{Conversation Length}$$



An example ...

Nine hypothetical audited **conversations** have **maxima** (already ordered):

0.08, 0.15, 0.20, 0.25, 0.30, 0.37, 0.42, 0.56, 0.82.

For $n = 9$ conversations and **conformal target** $1 - \alpha = 0.80$:

$$k = \lceil (n + 1)(1 - \alpha) \rceil = \lceil 10(0.8) \rceil = 8, \quad \hat{q} = R_{(8)} = 0.56.$$

At **every turn of a new conversation**, retain candidate y if $1 - p_t(y) \leq 0.56$,
i.e. $p_t(y) \geq 0.44$.



A new batch of **15 conversations**

Recall: we used **9 audited conversations** to calculate our conformal threshold:

$$\hat{q} = 0.56.$$

Now consider a **separate hypothetical evaluation batch** of 15 new conversations.

Conversation type	No. of threads	Turns/thread	Total turns
Routine	12	2	$12 \times 2 = 24$
Complex	3	8	$3 \times 8 = 24$
Total	15	—	48



One such **routine** conversation

Turn	Student's requested action	True Y_t	(p_A, p_B, p_C)	C_t	Covered?
1	Automatic five-day extension	B	(.18, .62, .20)	$\{B\}$	✓
2	Emailing medical documentation	A	(.63, .24, .13)	$\{A\}$	✓

Turn 1: correct category B gets 0.62, so $1 - 0.62 = 0.38$.

Turn 2: correct category A gets 0.63, so $1 - 0.63 = 0.37$.

The conversation's worst score is

$$R_{\text{routine}} = \max(0.38, 0.37) = 0.38 \leq 0.56.$$



A **complex** conversation: the first five turns go wrong

Same fixed inclusion rule $p_t(y) \geq 0.44$.

t	Student question / follow-up	True Y_t	(p_A, p_B, p_C)	C_t	In set?
1	Is the five-day medical extension automatic?	B	(.66, .22, .12)	{A}	✗
2	What if I submit the certificate late?	B	(.64, .24, .12)	{A}	✗
3	Can I request a two-week extension?	B	(.18, .24, .58)	{C}	✗
4	Can I appeal a refusal?	A	(.20, .18, .62)	{C}	✗
5	Can I submit while the appeal is pending?	C	(.61, .21, .18)	{A}	✗

NT: The student asks follow-up questions on a **mistaken premise** originating from turn 1.



Same complex conversation – three more wrong turns

t	Student's follow-up	True Y_t	(p_A, p_B, p_C)	C_t	In set?
6	Is the fee waiver automatic?	B	(.68, .20, .12)	{A}	✗
7	Can the department override refusal?	B	(.14, .20, .66)	{C}	✗
8	Can I reopen the case with new evidence?	B	(.64, .25, .11)	{A}	✗

For Turn 6, $Y_6 = B$ but $p_6(B) = 0.20$. Thus its score is $1 - 0.20 = 0.80$. Across the eight turns,

$$R_{\text{complex}} = \max_t (1 - p_t(Y_t)) = 0.80 > 0.56.$$



Can 80% conversation coverage be covering something up?

Suppose the following happens in our hypothetical batch:

Thread type	Threads	Covered turns/thread	Fully covered threads
Routine	12	2/2	12
Complex	3	0/8	0

Conversation coverage

$12/15 = 80\%$ fully covered conversations.

Turn coverage

ONLY $24/48 = 50\%$ individual turns are covered.



Part V. What about free-form cases?



How does this connect back to **open-ended** LLM generation?

So far, black-box uncertainty signals such as:

self-consistency, $H_{\text{sem}}(x)$, claim-level disagreement.

A natural idea is:

LLM uncertainty score $\longrightarrow$ calibration $\longrightarrow$ abstention threshold.

Conformal methods can turn a heuristic uncertainty signal into a statistically calibrated decision rule – under stated assumptions.



Why this is harder for free-form generation

For classification, the **true target** is one label in a small **closed** set. For an LLM answer:

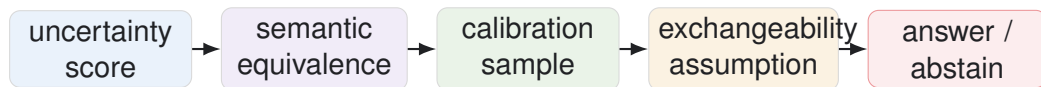
- ▶ the output space is **enormous**;
- ▶ two responses may express the **same answer**;
- ▶ correctness may be **claim-level**, not response-level;
- ▶ an automatic evaluator (e.g. NLI) must decide **whether two responses match**.

LLM conformal abstention **inherits the semantic-equivalence issues**



Bottleneck: The dependency chain

An LLM abstention guarantee can only be as meaningful as the objects it calibrates.



A statistical guarantee does not erase upstream modeling assumptions.



Part VI. Responsible Action Policies



University policy assistant

Suppose the system now knows:

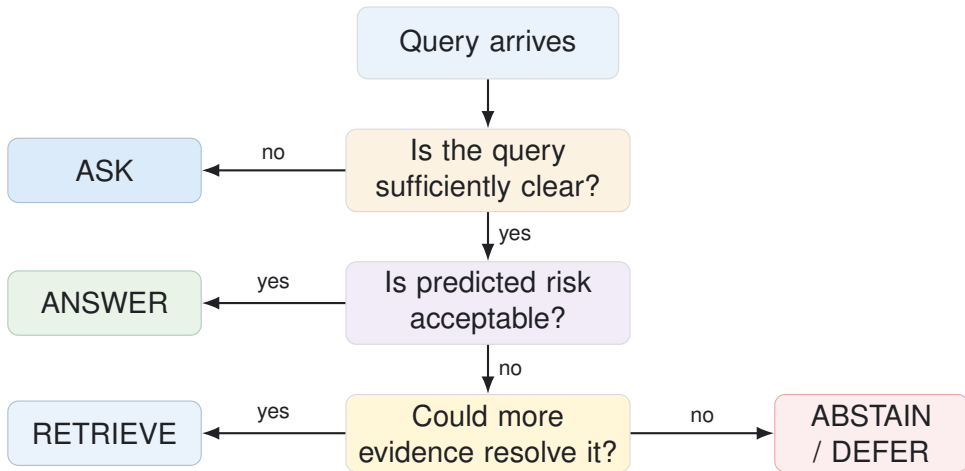
- ▶ confidence is poorly calibrated or low;
- ▶ semantic entropy is high;
- ▶ conformational candidate set has multiple meanings;
- ▶ evidence has not yet been retrieved.

Should the system merely say “*I don’t know*”?

Not necessarily.



A simple decision tree



To sum up ...

Tool	What it gives	What it does not automatically give
Selective prediction	answer / reject operating point	statistical coverage guarantee
Risk–coverage curve	reliability–utility trade-off	correct threshold for your application
Conformal prediction	finite-sample marginal set coverage under exchangeability	instance-level truth / grounding



NT: three words that must not be used interchangeably

Calibration

Does a score match empirical frequencies?

Coverage

Does a prediction set contain the truth often enough?

Abstention

Does the system choose not to release an answer?

A conformal set can have coverage *without yielding a single answer*;
Abstention can be used *without any conformal guarantee*.



One final check before we leave

Suppose a deployed LLM says:

“I will answer because my uncertainty score is below 0.30.”

Before trusting that policy, ask:

1. What event does the uncertainty score predict?
2. How was the answer / abstain threshold chosen?
3. What is the retained-answer risk at that operating point?
4. Is there a finite-sample guarantee – and exactly what quantity does it cover?
5. What assumptions make that guarantee valid?
6. When uncertainty is high, should the system ask, retrieve, verify, abstain, or defer?



So far...

We now understand	Still unresolved
answer / abstain selection	whether evidence actually supports the answer
coverage and selective risk	retrieval uncertainty and source sufficiency
risk–coverage curves	citation correctness
conformal prediction guarantees	grounding



Lecture 3 takeaway

Uncertainty is useful only when it changes action.

Selective prediction trades coverage for risk; conformal methods add explicit statistical coverage guarantees.

Next:

Accountable LLM Systems – Evidence Grounding, Oversight, and Auditability