ELLM

Query Governance in eLLM

The article discusses query governance in eLLM, a system that monitors and controls questions asked to AI assistants, ensuring harmful content is flagged or blocked. It details configuration options, detection layers, and reporting features for administrators.

AUTHOR
Paul Flanders

PUBLISHED
8 Jun 2026

DOCUMENT
InfoHub — Knowledge, news, and updates from EssingtonITS

Query Governance in eLLM

Paul Flanders · 8 Jun 2026

What this product is for

When staff or students use an AI assistant, most organisations want assurance that it is not being used to seek harmful information, and evidence that they are monitoring for it. Query governance is built for exactly that, especially in education, defense, and other regulated settings. It is the inbound counterpart to eLLM's existing privacy controls: instead of governing the answer, it governs the question. The result is a system you can put in front of users while staying in control of how it is used, and being able to prove it during an inspection or safeguarding review.

Main features

- Screens every question, across chat, document-grounded answers, image questions, and the external API, in one place.
- Two actions per category: **block** (refuse with a safe message, no model cost) or **flag** (answer normally but record an alert).
- **Keyword detection** that always runs, matching terms and patterns even when disguised (for example spacing, look-alike letters, or number-for-letter swaps).
- Optional dynamic detection that catches paraphrased or novel wording the keyword list never listed, and helps discover new slang.
- Exempt groups (such as administrators) who are never screened.
- A dedicated **Governance** console with configuration on one page and alerts, trends, and reports on another.

- Per-user safeguarding analysis that shows whether someone's attempts look persistent and intent-driven, written up by the assistant for a safeguarding lead.
- Every block and flag is written to the tamper-proof activity record and can be exported for evidence.

How it works

Blocking and flagging

Each governance category is set to one of three actions:

- **Block:** the question is refused with a safe, neutral message and the model is never called, so a blocked query costs nothing. The event is recorded.
- **Flag:** the question is answered as normal, but an alert is recorded so the activity is visible to administrators. The user is not interrupted.
- **Allow / off:** the category does not act.

If a question matches more than one category, the strongest action wins, so a block always takes precedence over a flag.

The two layers of detection

Detection works in two layers, and an administrator chooses how far to take it:

- **Keyword and pattern matching (always on):** each category holds a list of terms or patterns. eLLM checks the question against them, and against a cleaned-up version of the question, so common disguises, extra spaces, look-alike characters, and number-for-letter swaps still match. This layer is instant and uses no AI processing.
- **Dynamic detection (optional):** when enabled, eLLM also compares the meaning of a question to example phrases an administrator has provided, catching paraphrases and slang the keyword list would miss. A local model can then review borderline cases after the answer has been sent, judging intent, removing false alarms, and surfacing newly discovered phrases that an administrator can add to the keyword list with one

click.

All of this runs on your own servers using your local model, so the screening itself never sends questions to an outside provider.

The Governance console

Governance has its own section in the admin console, separate from general settings, with two pages:

- **Configuration:** a master on/off switch, the categories (each with its action, its terms, and its example phrases), the dynamic-detection options, the list of exempt groups, and the message shown when a query is blocked. A set of sensible default categories is provided to start from.
- **Alerts & reports:** the monitoring view, described next.

Alerts and reports

- The reports page evidences how the assistant is being used over a period you choose:
- Headline figures for blocked and flagged activity, plus a weighted risk score.
- A trend chart over time and a breakdown by category.
- The users with the most governance hits, so repeat activity stands out.
- A panel of newly discovered phrases, each with a one-click button to add it to a category.
- A recent-events table you can filter (by action, category, user, date range, or free-text search) and export to Excel or CSV. Each row links straight to the exact tamper-proof record behind it.
- **AI safeguarding analysis**, shown in a centred pop-up: ask the assistant to summarise the whole period, or focus on one user. The per-user view includes a **persistence breakdown** that classifies activity as Persistent, Recurring, or Isolated and shows how the attempts are spread over time, so a safeguarding lead can judge

whether someone looks intent-driven or had a one-off lapse.

Common tasks

- **Turn governance on:** open the Governance console, switch the master toggle on, and review the default categories.
- **Block a topic outright:** set that category's action to block; the user gets a safe message and the model is never called.
- **Monitor without interrupting users:** set a category to flag, so questions are answered but recorded for review.
- **Catch disguised or paraphrased wording:** turn on dynamic detection and add a few example phrases to the category.
- **Prepare for a safeguarding or inspection review:** set the period, read the trend and per-user analysis, and export the events.
- **Act on discovered slang:** use the one-click add on the suggested-terms panel to fold a new phrase into a category.

Things to know

- Governance is off by default. Nothing is screened until an administrator turns it on.
- Screening happens before any document search or model call, so it covers chat, document-grounded answers, image questions, and the external API uniformly.
- Exempt groups (administrators by default) are never screened.
- The detection uses your local model only; questions are never sent to a cloud provider for screening.
- Flags raised by the dynamic layer are recorded quietly after the answer, by design, so they do not add any delay for the user.
- It is built to evidence monitoring for an Ofsted or safeguarding review, but it is useful anywhere you need oversight of how an AI assistant is being used.

Troubleshooting

- **Nothing is being blocked or flagged:** check the master switch is on, that the category has an action set, and that the user is not in an exempt group.
- **A harmless question was caught:** a keyword may be too broad. Refine the category's terms, or rely on dynamic detection, which filters many false alarms.
- **Disguised wording is getting through:** turn on dynamic detection and add example phrases that capture the intent rather than exact words.
- **I want monitoring but not refusals:** set the categories you care about to flag rather than block.
- **A report looks empty:** widen the period, or confirm governance has been on long enough to record activity.

Frequently asked questions

Does a blocked question cost anything?

No. A block refuses before the model is ever called, so there is no model cost.

Will users know they were flagged?

No. Flagged questions are answered normally; the alert is recorded for administrators only.

Does screening send our questions to the cloud?

No. Detection runs on your own servers with your local model.

Can I prove we are monitoring?

Yes. Every block and flag is written to the tamper-proof activity record, and the reports can be exported to Excel or CSV as an evidence pack.

How do I tell a persistent attempt from a one-off?

Use the per-user analysis. It classifies activity as Persistent, Recurring, or Isolated and shows how it is spread over time.

Summary

Query governance gives you oversight of what people ask eLLM. Categories you control either block harmful questions outright, at no model cost, or quietly flag them for review. Keyword matching always runs, optional dynamic detection catches disguised and novel wording, and a dedicated console turns the activity into trends, per-user safeguarding analysis, and exportable evidence. It is a practical way to offer an AI assistant in education, defence, and other sensitive settings while staying in control and able to demonstrate it.

> Need assistance navigating the complexities of eLLM? EssingtonITS offers expert guidance and tailored IT solutions to help you succeed. Visit EssingtonITS.co.uk for comprehensive support.