International Journal of Ethical AI Application

https://ijeaa.cultechpub.com/ijeaa

E-ISSN: 3093-7736

Vol. 2, No. 1, September 2026

DOI: 10.64229/b98f3723

Cultech Publishing

Article

The Illusion of Private AI: Rethinking Data Privacy in Isolated Context and Private Knowledge Base Systems

Karim Mualla*

School of Computing and Mathematical Sciences, University of Leicester, Leicester, United Kingdom

*Corresponding author: Karim Mualla, kjm49@le.ac.uk

Abstract

Private AI is used in this paper as a deployment-based umbrella term for AI systems whose data and computation are constrained within a defined trust boundary. That boundary may be local or on-device, on-premises or private-network infrastructure, a private or dedicated cloud, or a managed cloud service that provides logical and contractual isolation. These models are not equivalent. This paper is concerned mainly with the final category: cloud-based isolated-context AI services that allow users to upload a bounded collection of documents and query them through one or more foundation models. NotebookLM is used as the main illustrative case, while Microsoft 365 Copilot and ChatGPT Business/Enterprise are comparison points. The central argument is straightforward: source isolation and exclusion from foundation-model training are useful privacy controls, yet neither is equivalent to local processing. The study uses a structured conceptual review of peer-reviewed security and privacy research, authoritative standards and regulator guidance, and current provider documentation. It separates privacy, cybersecurity and governance concerns across five layers: transmission and account boundary; inference; stored and derived representations; training and model improvement; and governance and human access. The paper then converts those layers into a practical zero-trust decision framework for deciding what may be uploaded, what requires an approved enterprise environment, and what should be minimised, anonymised or kept local. Its practical value is that it gives educators, managers and technical staff a way to make a data decision without pretending that a simple label such as private or no training answers every privacy question. The framework is illustrated through the design of an AI Security, Ethics and Management module at the University of Leicester. No empirical validation is claimed at this phase; the educational application is a design case that motivates a planned evaluation.

Keywords

Generative AI, Private AI, NotebookLM, Data privacy, Large language models, AI security, Retrieval-augmented generation, Governance, Zero trust, Higher education

Article History

Received: 30 June 2026
Revised: 26 August 2026
Accepted: 03 September 2026
Available Online: 08 September 2026

1. Introduction

Private AI is an attractive phrase, however, it needs a technical boundary. In this paper, Private AI means an AI deployment in which data access and model processing are intentionally constrained within a defined trust boundary. That boundary may be (1) local or on-device, where processing stays on user-controlled hardware; (2) on-premises or private-network infrastructure controlled by an organisation; (3) a private or dedicated cloud environment; or (4) a managed cloud service that relies on logical isolation, account controls and contractual privacy commitments. A laptop with networking disabled and a managed cloud knowledge service may both be described as private, but they achieve privacy in materially different ways. The paper therefore treats Private AI as a spectrum of deployment and assurance rather than a single architecture.

The distinction matters because generative AI is no longer a niche activity. HEPI Report 199, the Student Generative AI Survey 2026 published on 12 March 2026, surveyed 1,054 full-time UK undergraduates; 95% reported using AI in at least one way and 94% reported using generative AI to help with assessed work [1]. In the workplace, Microsoft and LinkedIn reported in 2024 that 75% of surveyed knowledge workers were already using AI at work [2]. These figures do not measure NotebookLM adoption particularly, and public product-level adoption statistics for notebook-style systems remain limited. They do, however, show the scale at which staff and students are now making everyday decisions about what material is safe to place into AI services.

A common example is NotebookLM. The user uploads source documents to a service, selects which sources should be active, asks questions and receives answers grounded in those materials. Google states that NotebookLM uses uploaded files, generated outputs and chat history to build the user's knowledge base, and it distinguishes this processing from direct training of foundation models [3,4]. For qualifying Workspace and Education accounts, Google also publishes stronger commitments around human review and generative-model training [5]. Those commitments are important. The point of this paper is not to imply that they are false. The point is that they answer only some of the questions that a strict privacy analysis should ask.

A document is not uploaded directly into an LLM as if the model were a file server. It is uploaded to an AI service or application. The service may extract text, divide it into chunks, create numerical representations for retrieval, store source material and conversation state, select relevant passages, construct a prompt, invoke one or more models, filter the output and maintain operational records. The model is one component of a wider application architecture. That architecture is the appropriate unit of analysis when the concern is data privacy.

Having defined the four deployment boundaries at the outset, the present study concentrates on managed cloud AI services with logical and contractual isolation. It calls this category a cloud-based isolated-context AI service. Notebook-style knowledge-base tools are one subtype of that category. Local, on-premises and private-cloud deployments remain important comparison points because they move or narrow the processing boundary rather than merely changing a privacy notice.

The problem is not simply legal or technical. Contracts and privacy notices define permitted processing, retention, review and secondary use. Technical architecture determines where data travels, what derived representations exist and what attack surfaces are introduced. Governance determines whether the right account, approval route and data category are being used. NIST treats generative AI risk as a lifecycle and socio-technical problem [6,7], while UK data-protection guidance expects organisations to examine purpose, data minimisation, security and accountability [8]. The European AI Act adds a wider governance context in which risk management, transparency and human oversight are increasingly important [9].

Recent peer-reviewed surveys similarly separate privacy and security concerns across training, inference and deployment rather than treating them as one model-level problem [10-12]. This matters because the question 'is my data used for training?' is narrower than the question 'what happens to my data when I use this service?'. The first question is worth asking. It is simply not enough by itself.

The research gap addressed here is practical. Existing work studies training-data extraction, prompt injection, retrieval attacks, embedding privacy and AI governance. Standards describe risk-management processes, and providers describe product-specific controls. What is less clear for an educator, manager or ordinary technical decision-maker is how to combine those pieces when deciding whether a particular document belongs in a particular AI environment. The contribution of this paper is a five-layer privacy analysis and a decision framework that translates those layers into an upload decision. It is deliberately not presented as a new cryptographic protection mechanism, nor as evidence that commercial services secretly train on every upload.

Accordingly, the paper has three analytical objectives rather than empirical hypotheses. First, it clarifies what private means across local, on-premises and managed-cloud deployment. Secondly, it identifies which privacy and cybersecurity concerns remain when model weights are not updated. Finally, it develops an auditable decision process for matching data sensitivity, processing purpose, architecture and consequence of failure. These objectives are addressed through a structured conceptual review, cross-system comparison and a pedagogical design application.

2. Background and Literature Review

2.1 Scope and Key Terminology

The terms privacy, security, confidentiality, data isolation and local processing are related but should not be used as synonyms. In this paper, privacy concerns the permitted collection, use, retention, disclosure and secondary use of information. Cybersecurity concerns protection of the service and its data against unauthorised access, manipulation or disruption. Confidentiality is the narrower property that information is not disclosed to unauthorised parties. Data isolation describes a logical or architectural boundary intended to stop one tenant, account or workload from accessing another. Local processing means that the relevant inference and data handling occur inside hardware controlled by the user or organisation rather than on a third-party cloud service.

Within this taxonomy, Private AI is not treated as a binary technical property. It is a deployment and assurance claim whose meaning depends on the boundary being controlled. A managed service such as NotebookLM may provide account, source and contractual isolation while still processing data remotely. The useful follow-up question is therefore not simply 'is this private?', but 'private from whom, within which processing boundary, and by what mechanism?'

For consistency, the phrase cloud-based isolated-context AI service is used throughout the rest of the paper for a managed service that accepts a bounded user or organisational knowledge source and uses that source to ground generated responses. A notebook-style AI service is a user-facing subtype that persists a collection of sources as a notebook or knowledge base. NotebookLM is the primary example because its product design makes the source boundary especially visible. Microsoft 365 Copilot and ChatGPT Business/Enterprise are not identical notebook products, but they provide useful comparison points because they show how different managed AI environments describe organisational data use and training controls.

Table 1 compares the main deployment patterns, their grounding or retrieval boundary, and the different meanings of privacy that follow from them.

Table 1. Comparison of private-AI deployment patterns and managed service examples.

Deployment / example	Processing boundary	Published training / data-use position	Residual privacy or security consideration
Local / on-device LLM	User-controlled device; may be offline	No third-party inference service is required	Endpoint compromise, local logs, model supply-chain risk and physical access remain relevant.
On-premises / private-network LLM	Organisation-controlled servers and network	Organisation determines training and retention policy	Internal administrator access, network segmentation, patching and local governance still matter.
NotebookLM with qualifying Workspace / Education account	Google-managed cloud service; users select notebook sources and responses are grounded in those sources with citations [4]	Google states that uploads, queries and model responses are not human reviewed or used to train AI models under the qualifying protections [5]	Remote processing and storage still occur; account configuration, sharing, retention and service terms remain part of the privacy boundary.
Microsoft 365 Copilot / Copilot Chat with enterprise data protection	Microsoft 365 service boundary; Microsoft 365 Copilot can ground prompts in permissioned Microsoft Graph content, semantic indexing and optional web context	Microsoft states that prompts, responses and Graph-accessed data are not used to train foundation models [13]	Prompts and responses may be logged and retained for audit/eDiscovery depending on the service and plan [14].
ChatGPT Business / Enterprise	OpenAI-managed workspace; Company Knowledge can search and fetch from eligible connected organisational sources and return cited answers	OpenAI states that business inputs and outputs are not used for training by default [15]	Privacy still depends on workspace configuration, retention choices, access controls and the organisation's own data policy.

From a retrieval perspective, the three managed services are not interchangeable. NotebookLM is the most explicitly source-bounded of the three examples: users choose the notebook sources and Google describes responses as grounded in those sources with inline citations [4]. Microsoft 365 Copilot uses a broader permission-driven model in which the Copilot orchestrator can ground a prompt using Microsoft Graph data the user is authorised to access, supported by semantic indexing and, where enabled, web grounding [16].

ChatGPT Business/Enterprise provides a different pattern again. Company Knowledge can search and fetch from eligible connected organisational sources and return citations to those sources [17], while business inputs and outputs are not used for model training by default [15]. Functionally, all three approaches augment generation with external context, but their retrieval corpus differs: a manually bounded notebook, a permissioned organisational graph and web context, or connected company sources. The exact backend retrieval algorithms are proprietary, so this paper does not assume that the three systems use an identical vector store, chunking strategy or textbook RAG implementation.

2.2 Retrieval-Augmented Generation (RAG) and Isolated Context

RAG is a common grounding architecture for using information that is not stored in a model's permanent parameters. In a standard RAG pipeline, source documents are ingested, divided into passages, represented for retrieval, selected in response to a query and then inserted into the LLM context before generation [18]. Notebook-style services may use proprietary variations rather than a textbook RAG pipeline, but the RAG model is useful for understanding why source isolation is not the same thing as model training.

A user can add recent or confidential material without waiting for a foundation model to be retrained. That is part of the attraction. Yet the service may also create extracted text, chunks, embeddings (numerical vectors that encode features or semantic relationships for retrieval), indices, prompts, citations, conversation histories and summaries. Peer-reviewed work on RAG privacy shows that retrieval databases can become a distinct privacy surface rather than a neutral add-on to the model [19]. The relevant object is therefore the end-to-end service, not only the LLM.

2.3 Processing, Model Training and in-Context Adaptation

These three operations need to be separated early because much of the public confusion comes from mixing them. Document processing means that the service reads, transforms, retrieves or otherwise uses the content to complete the requested task. Model training means that an optimisation procedure updates model parameters using training examples. In-context adaptation describes the temporary way in which examples or instructions within the current context can change model behaviour without a permanent parameter update.

Transformers process input via attention and feed-forward operations that produce temporary internal states [20]. Ordinary inference does not require a backward pass or a permanent weight update. That is why a provider can truthfully say that a document is processed but not used to train the foundation model. In-context learning nevertheless allows model behaviour to adapt to examples supplied inside the prompt or context [21-23]. This does not mean that every prompt secretly fine-tunes the model. It means that processing, in-context adaptation and training are technically different claims and privacy wording should not collapse them into one.

2.4 Embeddings, Memorisation and Privacy Risk

An embedding is a numerical vector intended to encode features or semantic relationships in data so that similar items can be retrieved or compared. In an isolated-context service, embeddings can be used to locate relevant passages without sending the whole source collection into every prompt. They are useful precisely because they preserve information about the source. That usefulness is also why they should not be treated as anonymised data by default.

Research on memorisation shows that sensitive training examples can sometimes be extracted from large language models [24]. That result concerns data that entered training and should not be misapplied to ordinary notebook uploads. A closer analogy for retrieval systems is embedding privacy. Morris et al. showed that text embeddings can reveal substantial information about source text [25], and work on private retrieval starts from the premise that neural similarity search can leak information unless the retrieval architecture is designed to protect both queries and database content [26]. These are privacy risks because the consequence is information disclosure or reconstruction. The security mechanisms used to protect embedding stores, by contrast, belong to cybersecurity. The two domains meet when a security failure enables a privacy violation.

2.5 Prompt Injection and Retrieval-Layer Security

Prompt injection is primarily a cybersecurity problem. An attacker or untrusted document causes the LLM application to treat data as an instruction and behave in an unintended way. The privacy consequence appears when that security failure causes sensitive information to be retrieved, exposed or sent to an unauthorised destination. Keeping the attack mechanism and the privacy outcome separate makes the risk easier to reason about.

Greshake et al. demonstrated indirect prompt injection against LLM-integrated applications that process external content [27]. RAG creates additional attack surfaces because the retrieval database can itself be manipulated. Tan et al. showed that crafted content inserted into a retrieval source can alter model behaviour even when the attacker lacks knowledge of the user's exact query or the underlying model parameters [28]. Zeng et al. found that RAG can expose private retrieval data under attack conditions [19]. These studies do not establish that every commercial notebook is vulnerable in the same way, but they show why a private source collection can also become a private attack surface.

The system view is reinforced by recent security work. Debenedetti et al. describe privacy side channels created by components around the model, including preprocessing, monitoring and filtering [29]. OWASP likewise treats prompt injection, sensitive-information disclosure, data/model poisoning and vector/embedding weaknesses as application-level concerns [30,31]. The UK National Cyber Security Centre makes the related point that prompt injection is not simply SQL injection under a new name and that the entire AI-enabled system must be considered [32,33].

2.6 Governance and the Boundary of Assurance

Governance matters here because several privacy claims cannot be inferred from model architecture alone. A user cannot inspect a provider's complete retention pipeline, administrator access model or internal logging simply by observing the generated answer. These claims therefore depend partly on contracts, published controls, audit mechanisms and institutional account settings.

Foundation-model research has long highlighted the gap between broad capability and complete control [34], while risk taxonomies emphasise information hazards and malicious use alongside other social harms [35]. For this paper, the useful governance question is narrower: does the organisation know the processing purpose, data category, deployment boundary, retention position and consequence of failure well enough to authorise the upload? Standards such as NIST AI RMF, ISO/IEC 42001 and ISO/IEC 23894 offer organisational risk-management structures, while OWASP provides a more security-specific application lens [36,37]. The framework proposed later does not replace any of them. It supplies a document-level decision checkpoint that can sit underneath those broader frameworks.

Standard texts on AI and deep learning provide the underlying distinction between learned parameters, inference and generalisation [38,39]. The management question, however, is intentionally simpler: who controls the data, what processing boundary is being relied upon, and what happens if the privacy assumption is wrong?

3. Methodology

This study uses a structured conceptual and design-oriented review. It is not a systematic review and it does not claim a PRISMA-style exhaustive search. Nor does it attempt to reverse engineer a proprietary product. The aim is to combine the evidence needed for a defensible decision framework: peer-reviewed research describing technical mechanisms and attacks; standards and regulator guidance describing governance expectations; and provider documentation describing current product-specific commitments.

The literature search was conducted iteratively across ACL Anthology, USENIX Security proceedings, ScienceDirect, SpringerLink and publisher or institutional repositories for NIST, ISO, OWASP, the ICO and the NCSC. Search concepts included combinations of large language model privacy, RAG privacy, retrieval attack, prompt injection, embedding privacy or inversion, AI governance, local AI, private AI, enterprise AI data protection and generative AI security. Seminal work before 2020 was retained where it defines transformer or deep-learning mechanisms. For the rapidly changing security and governance literature, preference was given to 2023-2026 sources.

Inclusion followed four rules. First, peer-reviewed journal or conference work was preferred for technical claims about attacks, privacy leakage, retrieval behaviour and system security. Secondly, official standards and regulator sources were used for governance and legal-policy framing. Thirdly, provider documentation was used only for claims about that provider's own service, for example whether business content is used to train foundation models. Provider statements were not treated as independent evidence that a product is universally secure. Fourthly, sources were excluded when they were purely promotional, duplicated an already stronger source, or could not be connected to one of the five privacy layers.

The analysis then proceeded in four stages. Stage one separated the broad phrase private AI into deployment boundaries: local/on-device, on-premises/private network, private cloud and managed cloud. Stage two coded the literature into privacy, cybersecurity and governance concerns, while also recording where the three overlap. Stage three mapped those concerns to five lifecycle layers: transmission and account boundary; inference; storage and derived representations; training and model improvement; and governance and human access. Stage four translated the layers into a practical upload-decision framework.

The framework was evaluated analytically rather than empirically. Five criteria were utilised: lifecycle completeness, meaning that the decision should not focus on inference alone; conceptual separation, meaning privacy, security and governance should not be treated as synonyms; actionability, meaning a user must be able to answer the question without privileged provider access; traceability, meaning each question can be linked to an identified risk layer; and transferability, meaning the same logic should remain useful across local, enterprise and consumer deployment patterns. A crosswalk against NIST AI RMF, ISO/IEC 42001/23894 and OWASP is provided in Section 6. This is a form of conceptual validation, not evidence of user effectiveness.

The University of Leicester material in Section 5 is therefore presented as a pedagogical design application rather than an empirical case study. No participants were recruited for this paper, no student data were collected and no learning outcome is claimed. The activities are included to show how the framework can be operationalised in teaching. Empirical evaluation is deliberately reserved for future work.

4. Technical Analysis: Five Privacy Layers

The five layers are not intended to restate generic cloud-security advice. Traditional cloud services already require encryption, access control and retention management. AI-based services add further privacy questions because user

content can be transformed into embeddings and retrieval indices, combined dynamically with prompts, routed through probabilistic model inference, turned into derived summaries, and exposed to new attack classes in which instructions and data share the same textual channel. The privacy outcome is therefore determined by the sensitivity of the data, the processing purpose and the complete service architecture, not simply by whether the network connection is encrypted.

4.1 Layer One: Transmission and Account Boundary

The first layer concerns where the data crosses from a user-controlled environment into a managed service. Encryption in transit protects the network journey, but it does not make remote processing local. The user is now relying on identity management, account permissions, provider infrastructure, tenant isolation, service configuration and contractual controls. This layer is shared with conventional cloud computing. What makes the AI case different is what happens next: the uploaded material can be transformed and repeatedly retrieved as part of an AI workflow rather than merely stored as a file.

4.2 Layer Two: Ingestion, Retrieval and Inference Exposure

Inference is only one component of the service. Before an LLM generates an answer, a cloud-based isolated-context service may parse a file, extract text, divide it into passages, create embeddings, build an index, retrieve selected passages and assemble a context. During model execution, that context becomes part of the computational input and generates temporary activation states. None of this requires a permanent weight update. It is nevertheless real processing of the content. The privacy question at this layer is therefore not only who can access the model, but which application components can access the source, query, retrieved passages and generated output.

4.3 Layer Three: Storage and Derived Representations

A persistent notebook must remember something. Depending on the design, this may include original files, extracted text, chunks, embeddings, retrieval indices, summaries, citations, prompts, responses and chat history. The privacy consequence depends on what is retained, for how long, under which tenant or account, and whether deletion covers derived artefacts as well as the source. An AI-specific concern is that a generated summary can be more convenient to disclose than the source itself because it may condense the most sensitive information into one readable paragraph.

4.4 Layer Four: Training and Model Improvement

Training exclusion is a meaningful control. If the published commitment is honoured, uploaded material should not be used to update the provider's general-purpose foundation-model parameters. That reduces a long-term secondary-use risk. The mistake is to promote this one control into a complete definition of privacy. A document can be excluded from training and still be uploaded, parsed, embedded, indexed, retrieved, logged or exposed through an application vulnerability. Training risk is one layer, not the whole system.

4.5 Layer Five: Governance and Human Access

The final layer concerns people, purpose and accountability. A notebook may be shared too widely. A consumer account may be used instead of an approved institutional account. Feedback can change the treatment of an interaction. Support staff, administrators, subprocessors, legal requests and retention policies may be governed by terms that differ between consumer, education and enterprise products. The privacy outcome is therefore a product of architecture, policy and user behaviour. A technically well-designed service can still be used inappropriately if the wrong data category is uploaded for the wrong purpose.

4.6 Cybersecurity Pathways that Can Become Privacy Failures

The privacy layers are easiest to use when the attack mechanism is made explicit. Account compromise or misconfigured sharing can expose the source collection at layer one. Retrieval poisoning and indirect prompt injection can alter what the service retrieves or how it interprets retrieved content at layer two. Weak access controls around vector stores or embedding services can disclose sensitive representations at layer three. Data or model poisoning concerns can affect training and model-improvement pipelines at layer four. Excessive administrator access, logging, integrations or retention can create layer-five exposure.

In other words, privacy and security should be connected causally rather than treated as interchangeable labels. Prompt injection is a security vulnerability; sensitive-information disclosure is one possible privacy consequence. A compromised vector store is a security failure; reconstruction or disclosure of confidential source content is the privacy consequence. This separation is consistent with OWASP's distinction between prompt injection, sensitive-information disclosure and vector/embedding weaknesses [30,31].

5. Pedagogical Application: AI Security, Ethics and Management at the University of Leicester

The framework is connected to the design of a new module titled AI Security, Ethics and Management at the University

of Leicester. This section is a pedagogical application case, not an empirical evaluation. It describes how the framework is used to structure planned learning activities. No participants were recruited for this study, no student responses were analysed and no claim is made that the activities have already improved learning outcomes.

The design starts from a practical observation. AI security, ethics and management are often taught as separate topics even though one organisational incident can involve all three. Uploading a confidential document to a cloud AI service may be encrypted correctly, contractually prohibited, ethically questionable and poorly governed at the same time. Students therefore need a vocabulary that separates the questions before reconnecting them.

The first activity, the Private Notebook Test, presents five fictional documents: a public lecture transcript, an unpublished research idea, a student-support note, a commercial partner brief and an examination question bank. Students classify each item and justify whether it can be used in a local model, an approved enterprise service or a consumer cloud service. The intended learning outcome is not agreement on one universal answer. It is the ability to justify an upload decision by reference to sensitivity, purpose, processing boundary, provider controls and consequence of failure.

The second activity traces a document across the AI workflow: upload, extraction, chunking, embedding, retrieval, prompt construction, model inference, output generation, citation, storage and deletion. This directly mirrors layers one to three of the framework and makes visible what a friendly interface compresses into a few clicks. It also gives students a precise way to explain why 'no back propagation' is useful information but not a complete privacy statement.

The third activity focuses on the security-to-privacy pathway. Students place harmless conflicting instructions inside a synthetic document and discuss how an LLM application may treat the text as both content and instruction. They then identify the security mechanism (indirect prompt injection), the affected layer (retrieval/inference), and the possible privacy outcome (unintended disclosure). The use of synthetic material keeps the activity educational rather than operationally risky.

This application is consistent with wider work arguing that AI teaching should connect technical capability with assessment design and governance rather than treating generative AI as a single classroom tool [40]. The empirical question–whether the framework actually improves understanding or classification consistency–remains open and is addressed in the Future Work section.

A concrete way to apply the framework is to consider an unreleased assessment question bank that a module leader wants an AI service to review for duplication, clarity and topic coverage. This is a synthetic worked scenario, and no live assessment material or student data were used in this study. The data category is high-consequence assessment content, the purpose is quality assurance, and the failure consequence is assessment compromise. Accordingly, a local or approved on-premises model keeps inference inside the institution-controlled boundary, although endpoint, administrator and model-supply-chain risks still remain.

A managed cloud service changes that boundary. NotebookLM, Microsoft 365 Copilot or ChatGPT Business or Enterprise may offer strong enterprise controls and published restrictions on model training, but those assurances do not themselves authorise the upload of live examination material. Under the five-layer framework, the decision therefore becomes conditional: if institutional policy explicitly approves that service and account type for unreleased assessment data, the managed environment may be acceptable. If not, the safer choice is local processing or a minimised/synthetic version of the material. The same service could still be entirely appropriate for public lecture notes. This is the practical point of the framework: it classifies the data and purpose before it classifies the tool as 'safe' or 'unsafe'.

6. Decision-Making Framework

The framework was derived by translating each lifecycle layer into a question that an ordinary decision-maker can answer before an upload. The order matters. Data sensitivity and purpose come first because the same architecture can be acceptable for public information and unacceptable for a disciplinary record. Processing location then establishes the trust boundary. Provider controls address training, human review and retention. Representation persistence asks what remains after the task. Consequence of failure prevents a low-probability event from being treated the same way regardless of impact.

Table 2 positions this contribution against established governance and security frameworks. The proposed framework is intentionally narrower: it is a document-level decision aid rather than an organisational management system or an attack catalogue.

Table 2. Relationship between the proposed framework and established AI risk/security guidance.

Framework / guidance	Primary level	Main strength	Gap addressed by this paper
NIST AI RMF / GenAI Profile [6,7]	Organisation / lifecycle	Govern, map, measure and manage AI risk across the lifecycle	Does not give a short document-by-document upload decision for isolated-context services.
ISO/IEC 42001 and ISO/IEC 23894 [36,37]	Management system / organisational risk	Formalises AI governance, risk responsibilities and continual improvement	Operates above the level of an individual user deciding whether a particular source should enter a cloud AI service.
OWASP LLM Top 10 [30]	Application security	Identifies concrete LLM application vulnerabilities and mitigations	Security taxonomy does not itself classify data sensitivity, provider policy and consequence of upload.
Five-layer private-AI framework (this paper)	Document / use-case decision	Connects data category, deployment boundary, derived representations, training controls and consequence of failure	Designed as a practical pre-upload checkpoint; it complements rather than replaces the broader frameworks.

Table 3 offers the resulting decision questions and explicitly establishes the link to the five privacy layers.

Table 3. Decision-making framework for assessing data exposure across private AI environments.

Decision area	Key question	Related layer(s)	Lower-risk example	Higher-risk example
Data category and purpose	What is the sensitivity of the content, and why is it being processed?	All layers; determines impact	Published module description used for summarisation	Student complaint, health data, confidential partner material or live assessment content
Processing location	Is processing local, on-premises/private network, private cloud or managed public cloud?	1-2	Local model with no external connection	Remote consumer service with unclear organisational approval
Provider and account controls	Are training, review, sharing and retention controls documented for this account type?	1, 4, 5	Managed education/enterprise account with explicit controls	Consumer account or unclear terms/settings
Representation persistence	What remains after the session, including files, chunks, embeddings, logs, prompts and summaries?	2-3	Temporary processing with verified deletion appropriate to the task	Persistent source, retrieval index, history and derived summaries
Consequence of failure	What happens if the privacy or security assumption is wrong?	All layers	Minor inconvenience involving already-public data	Legal breach, assessment compromise, IP loss, safeguarding harm or major reputational damage

The rule produced by Table 3 is deliberately cautious but not anti-cloud. Public, low-consequence material can usually be processed through mainstream tools, subject to policy. Internal but non-sensitive information may be acceptable when account and retention controls are clear. Confidential, regulated, assessment-related, commercially sensitive or unpublished information should be minimised, anonymised, moved to an approved enterprise environment, processed locally, or not uploaded at all. The framework does not assume that providers are dishonest. It assumes that an important decision should not depend on one reassuring label.

Figure 1 illustrates how the service workflow maps onto the five privacy layers. The user and transmission stages correspond mainly to layer one. Cloud processing combines ingestion, retrieval and inference in layer two. Storage and representation map directly to layer three. The figure's provider-use and access column spans layer four, where training/model improvement is considered, and layer five where administrative access, monitoring and governance controls sit. The output stage is cross-cutting: a generated response can disclose information produced from any earlier layer. This is why the decision framework starts with data sensitivity and consequence rather than treating inference as the sole risk point.

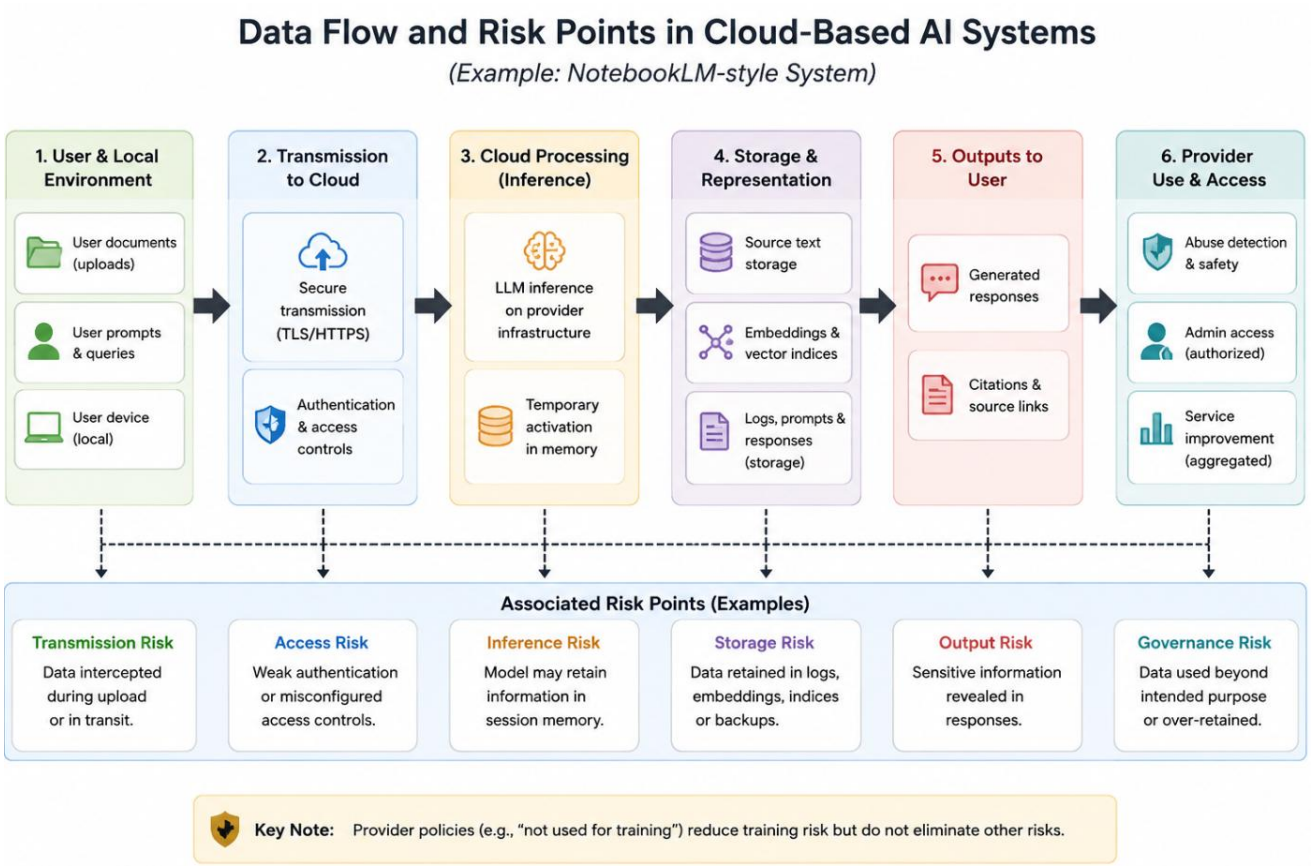

Figure 1. Data flow and risk points across the five privacy layers in a cloud-based isolated-context AI service. Legend: the colours distinguish workflow stages, not six privacy layers. Layer 1 covers the account/transmission boundary and transmission to the managed service; Layer 2 covers ingestion, retrieval and inference; Layer 3 covers storage and derived representations; Layers 4 and 5 are represented in provider use/access as training/model-improvement and governance/human-access concerns. The output stage is cross-cutting rather than a sixth layer.

7. Discussion

7.1 Private by Architecture, Contract or Experience

The word private is doing too much work. To a user, it may mean that nobody else can see the notebook. To a provider, it may mean that the content is logically isolated and excluded from general model training. To a lawyer, it may mean that processing is governed by a contract. To a cybersecurity specialist, it may mean that access, logging, retention and attack paths have been controlled. All of those interpretations can be reasonable, but they are not interchangeable.

The additional distinction introduced in this revision is architectural. Local and on-premises AI do not merely provide a stronger version of the same cloud privacy promise; they move the processing boundary. A user may still have poor local security, weak permissions or an untrusted model supply chain, but the third-party inference boundary can be removed. Managed cloud services instead rely on technical and contractual isolation inside provider infrastructure. The choice between them should be made according to data sensitivity and purpose rather than ideology.

7.2 No Training Is Not No Exposure

The most important distinction in this paper remains the one between no training and no exposure. No training refers to the absence of a particular secondary use: updating future model parameters with submitted content. No exposure, in the strict architectural sense, would require the content to stay within the user-controlled processing boundary. A managed cloud service can credibly promise the first under defined terms. By design, it cannot claim the second in the same sense as an offline local system.

This is not an argument against enterprise AI. In fact, clearer language can strengthen trust. A useful privacy statement would tell the user separately whether data leaves the device, which service components process it, what is retained, what is logged, whether human review can occur, whether derived representations persist, and whether any of that content is used to update general models. A longer explanation is less marketable than the word private, but it supports a better decision.

7.3 What Is Novel about the Five-Layer Framework

The paper's novelty is not that it discovers remote cloud processing, nor that it proposes another high-level AI governance framework. Its contribution is the level at which the decision is made. NIST and ISO operate primarily at organisational and lifecycle levels. OWASP catalogues application threats. Privacy research examines particular leakage mechanisms. The five-layer framework connects those perspectives to one operational question: should this specific information be placed into this specific AI service for this purpose?

That connection also clarifies the relationship between privacy and cybersecurity. A prompt-injection vulnerability is not automatically a privacy breach; it becomes one when the attack causes confidential material to be disclosed. An embedding is not automatically a leak; it becomes a privacy concern when the representation can be accessed, reconstructed or used beyond the intended purpose. The framework therefore keeps mechanisms and consequences separate while ensuring that both influence the upload decision.

7.4 Classification Rather Than Fear

A blanket ban on cloud AI would be unrealistic and, in many settings, counterproductive. These tools can improve accessibility, research synthesis, feedback and productivity. The problem is not cloud processing by itself. The problem is applying the same decision to every document.

Higher education guidance should therefore move beyond 'use AI responsibly'. Public lecture slides are not the same as unpublished examination questions. Synthetic data is not the same as named student records. A managed education account is not the same as an arbitrary consumer account. Classification turns a vague ethical message into an operational control. The same logic can transfer to an organisation deciding between local inference, an approved enterprise platform and a consumer tool.

7.5 Limitations

The first limitation is empirical. The framework has not yet been validated through user studies, expert panels or controlled comparison of commercial infrastructures. The University of Leicester section and the worked assessment scenario are design applications, not evidence of measured effectiveness. They demonstrate how the framework can be used, but not how much it improves decision quality. The present contribution should therefore be read as a conceptual and practice-oriented framework whose effectiveness still needs to be tested.

The second limitation is dependence on publicly available information. Product architectures are partly opaque, and provider documentation can change. As a result, the product comparisons in Table 1 are a dated evidence snapshot based on documentation available in August 2026, not a permanent benchmark. A statement about training, retention, grounding or human review is accurate only for the documented product, account type and date examined. This is exactly why the framework asks users to verify current controls rather than treating one vendor statement as permanent.

Thirdly, the framework simplifies technical and legal detail so that non-specialists can use it. That is a design choice but also a limitation. It cannot replace threat modelling, penetration testing, data-protection impact assessment or specialist legal advice. Jurisdiction also matters: contractual and regulatory requirements differ between the UK, EU and other regions.

Finally, the motivating application is higher education. The decision areas are deliberately generic enough to transfer to healthcare, government or commercial organisations, but that transfer has not yet been empirically demonstrated. Different sectors may need different data categories, approval thresholds and consequence scales.

8. Conclusion

This paper has examined the gap between isolated context and strict data privacy in cloud-based AI services. Its central conclusion is deliberately narrow. A service may be private with respect to source access, account isolation and foundation-model training while remaining remote with respect to processing and storage. Those statements can all be true at the same time.

The five-layer analysis separates transmission and account boundary, ingestion/retrieval/inference, stored representations, training/model improvement and governance/human access. The decision framework translates those layers into five practical questions about data sensitivity and purpose, processing location, provider controls, persistence and consequence of failure. The crosswalk with NIST, ISO and OWASP shows that the framework is not a replacement for established governance or security guidance. It sits below them as a pre-upload decision aid.

The pedagogical application shows how the same framework can organise teaching in AI Security, Ethics and Management without pretending that a classroom activity is empirical validation. Students can trace the complete workflow, distinguish a security mechanism from a privacy consequence and justify why one document belongs in an enterprise service while another should remain local. That is also the wider management lesson. Private AI is better treated as a spectrum of assurance than a binary product label.

9. Future Work

Future work will move from conceptual validation to empirical evaluation in both education and organisational settings. The first study will recruit students and staff who use generative AI and will employ a pre/post design. Participants will complete a short knowledge test covering processing, training, embeddings, logging and local execution before and after a technical intervention based on the five-layer model. Improvement will be measured using change in test scores and confidence ratings.

A second study will test decision consistency. Participants will classify a fixed set of public, internal, confidential and regulated scenarios and justify the selected processing environment. Agreement can be measured using percentage agreement and an inter-rater measure such as Cohen's kappa where the design permits. Additionally, the analysis will examine whether participants identify the same risk layer and consequence of failure, rather than merely choosing the same final answer.

A third study will compare local/on-premises and managed enterprise workflows for the same bounded tasks. Candidate measures include task completion time, usability, answer quality, perceived trust, correct identification of privacy controls, and the amount of source material that needs to leave the user-controlled boundary. The purpose will not be to prove that local is always safer or enterprise cloud is always better, but to measure the trade-offs that the conceptual paper currently describes.

Finally, an organisational pilot will seek assessment from information-security, data-protection and AI-governance practitioners outside higher education. Expert review can test whether the five questions are complete enough for real approval workflows and whether sector-specific additions are required. This would provide the evidence needed to refine the framework from a teaching and decision aid into an empirically validated organisational tool.

Acknowledgements

No individuals or organisations are acknowledged beyond the sources cited in the manuscript.

Funding

This research received no external funding.

Ethics Statement

This study did not involve human participants, animals, or the collection or analysis of personal data; therefore, ethical approval was not required.

Data Availability Statement

No new data were generated or analysed in this study. The work is based on published literature, standards, regulatory guidance and provider documentation cited in the manuscript.

Author Contributions

K.M. conceived the study, developed the conceptual framework and pedagogical application, conducted the literature review and interpretation, prepared the manuscript and figures, and reviewed and approved the final version.

Conflict of Interest Statement

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Generative AI Usage Statement

Gemini, Claude and ChatGPT were used only as editorial support tools during manuscript development and revision. They assisted with organising the author’s existing notes and arguments, language and structural editing in the Introduction, Literature Review, Methodology, Discussion and Conclusion, reference formatting, and preparation of responses to editorial and peer-review comments. The research question, conceptual argument, five-layer framework, pedagogical application, interpretation of sources, selection of substantive claims and final scholarly responsibility are the author’s own. No empirical data were generated or analysed by generative AI. The author reviewed the final manuscript and accepts responsibility for all cited sources and factual claims.

Figure Copyright Statement

Figure 1 was created specifically for this manuscript under the author's direction and is not reproduced from a third-party publication. The author confirms that he holds the rights required to submit and publish the figure and is not aware of any copyright infringement issue.

References

[1] Stephenson R, Armstrong C. Student generative artificial intelligence survey 2026. London: Higher Education Policy Institute. 2026. (HEPI Report 199). Available from: https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/ (accessed on 07 September 2026).

[2] Microsoft, LinkedIn. 2024 Work trend index annual report: AI at work is here. Now Comes the Hard Part. 2024. Available from: https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part/ (accessed on 07 September 2026).

[3] Google. Privacy and terms of use in NotebookLM. NotebookLM Help. Available from: https://support.google.com/notebooklm/answer/17004255 (accessed on 07 September 2026).

[4] Google. Learn about NotebookLM. NotebookLM Help. Available from: https://support.google.com/notebooklm/answer/16164461 (accessed on 07 September 2026).

[5] Google Workspace. Generative AI in Google Workspace privacy hub. Google Workspace Help. Available from: https://knowledge.workspace.google.com/admin/gemini/generative-ai-in-google-workspace-privacy-hub?hl=en (accessed on 07 September 2026).

[6] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0), NIST AI 100-1. 2023. DOI: 10.6028/NIST.AI.100-1

[7] National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative Artificial Intelligence Profile, NIST AI 600-1. 2024. DOI: 10.6028/NIST.AI.600-1

[8] Information Commissioner's Office. Guidance on AI and data protection. 2023. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ (accessed on 07 September 2026).

[9] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. 2024. Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed on 07 September 2026).

[10] Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 2024, 4(2), 100211. DOI: 10.1016/j.hcc.2024.100211.

[11] Kibriya H, Khan WZ, Siddiqa A, Khan MK. Privacy issues in Large Language Models: A survey. Computers and Electrical Engineering, 2024, 120, 109698. DOI: 10.1016/j.compeleceng.2024.109698.

[12] Zhang R, Li HW, Qian XY, Jiang WB, Chen HX. On large language models safety, security, and privacy: A survey. Journal of Electronic Science and Technology, 2025, 23(1), 100301. DOI: 10.1016/j.jnlest.2025.100301

[13] Microsoft. Enterprise data protection in Microsoft Copilot and Microsoft Copilot Chat. Microsoft Learn. Available from: https://learn.microsoft.com/en-us/microsoft-365/copilot/enterprise-data-protection (accessed on 07 September 2026).

[14] Microsoft. Frequently asked questions about Microsoft Copilot Chat. Microsoft Learn. Available from: https://learn.microsoft.com/en-us/copilot/faq (accessed on 07 September 2026).

[15] OpenAI. Business data privacy, security, and compliance. 2026. Available from: https://openai.com/business-data/ (accessed on 07 September 2026).

[16] Microsoft. Microsoft 365 Copilot architecture and how it works. Microsoft Learn. Available from: https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-architecture (accessed on 07 September 2026).

[17] OpenAI. Company knowledge in ChatGPT (Business, Enterprise, and Edu). OpenAI Help Center. https://help.openai.com/en/articles/12628342-company-knowledge-in-chatgpt-business-enterprise-and-edu (accessed on 07 September 2026).

[18] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33. DOI: 10.48550/arXiv.2005.11401

[19] Zeng S, Zhang J, He P, Xing Y, Liu Y, Xu H, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024, 4505-4524. DOI: 10.18653/v1/2024.findings-acl.267

[20] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Long Beach, CA, USA: Curran Associates Inc. 2017, 5998-6008. DOI: 10.48550/arXiv.1706.03762.

[21] Akyürek E, Schuurmans D, Andreas J, Ma T, Zhou D. What learning algorithm is in-context learning? Investigations with linear models. arXiv, 2022 DOI: 10.48550/arXiv.2211.15661

[22] von Oswald J, Niklasson E, Randazzo E, Sacramento J, Mordvintsev A, Zhmoginov A, et al. Transformers learn in-context by gradient descent. In: Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 2023. Available from: https://proceedings.mlr.press/v202/von-oswald23a.html (accessed on 07 September 2026).

[23] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020. DOI: 10.48550/arXiv.2005.14165

[24] Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. In: 30th USENIX Security Symposium, 2021. Available from: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting (accessed on 07 September 2026).

[25] Morris J, Kuleshov V, Shmatikov V, Rush A. Text embeddings reveal (Almost) as much as text. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023, 12448-12460. DOI: 10.18653/v1/2023.emnlp-main.765

[26] Zyskind G, South T, Pentland A. Don't forget private retrieval: Distributed private similarity search for large language models. Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, 2024, 7-19. Available from: https://aclanthology.org/2024.privatenlp-1.2/ (accessed on 07 September 2026).
[27] Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 2023. DOI: 10.1145/3605764.3623985
[28] Tan Z, Zhao C, Moraffah R, Li Y, Wang S, Li J, et al. Glue pizza and eat rocks–exploiting vulnerabilities in retrieval-augmented generative models. Proceedings of EMNLP 2024, 1610-1626. DOI: 10.18653/v1/2024.emnlp-main.96
[29] Debenedetti E, Severi G, Carlini N, Choquette-Choo CA, Jagielski M, Nasr M, et al. Privacy side channels in machine learning systems. 33rd USENIX Security Symposium, 2024. Available from: https://www.usenix.org/conference/usenixsecurity24/presentation/debenedetti (accessed on 07 September 2026).
[30] OWASP GenAI Security Project. OWASP Top 10 for LLM Applications 2025. Available from: https://genai.owasp.org/llm-top-10/ (accessed on 07 September 2026).
[31] OWASP GenAI Security Project. LLM08: 2025 Vector and Embedding Weaknesses. 2025. Available from: https://genai.owasp.org/llmrisk/llm082025-vector-and-embedding-weaknesses/ (accessed on 07 September 2026).
[32] UK National Cyber Security Centre. Prompt injection is not SQL injection (it may be worse). 2025. Available from: https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection (accessed on 07 September 2026).
[33] UK National Cyber Security Centre. Thinking about the security of AI systems. 2023. Available from: https://www.ncsc.gov.uk/blog-post/thinking-about-security-ai-systems (accessed on 07 September 2026).
[34] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. 2021. DOI: 10.48550/arXiv.2108.07258
[35] Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, et al. Ethical and social risks of harm from Language Models. arXiv, 2021. DOI: 10.48550/arXiv.2112.04359
[36] International Organization for Standardization. ISO/IEC 42001:2023, Information technology–Artificial intelligence–Management system, 2023. Available from: https://www.iso.org/standard/42001 (accessed on 07 September 2026).
[37] International Organization for Standardization. ISO/IEC 23894:2023, Information technology–Artificial intelligence–Guidance on risk management, 2023. Available from: https://www.iso.org/standard/77304.html (accessed on 07 September 2026).
[38] Russell SJ, Norvig P. Artificial intelligence: A modern approach, 4th ed. NJ: Pearson, 2021. Available from: https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780137505135 (accessed on 07 September 2026).
[39] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press, 2016. Available from: https://www.deeplearningbook.org/ (accessed on 07 September 2026).
[40] Mualla KJ, Mualla W. Leveraging AI in Higher Education: Contemporary Approaches for Teaching, Learning and Assessment Design. In: Higher Education in the Arab World: Artificial Intelligence, Springer, Cham, 2026, 169-191. DOI: 10.1007/978-3-031-99068-7_8