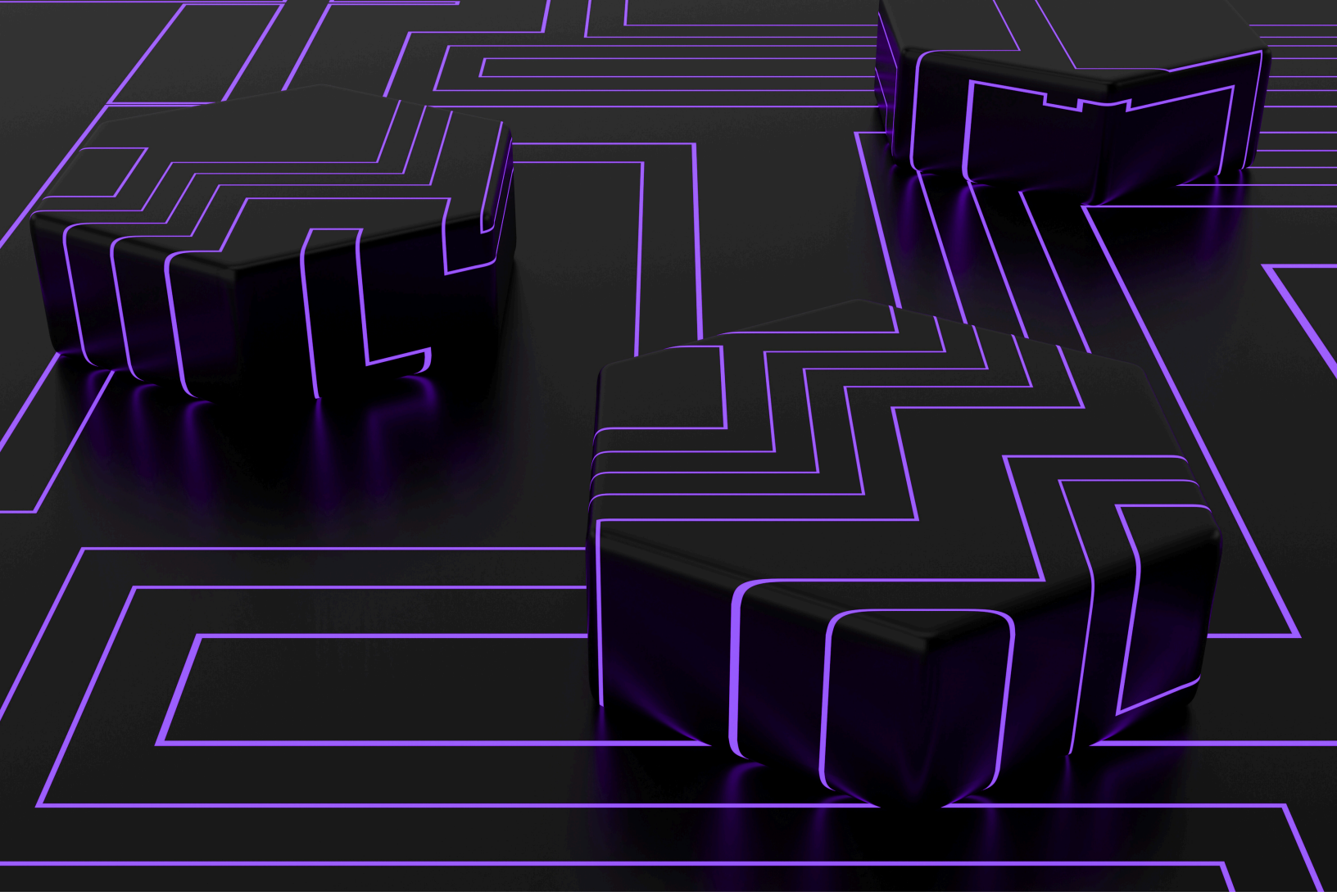


RECOGNIZING AND MITIGATING RISKS WHEN USING LLMS

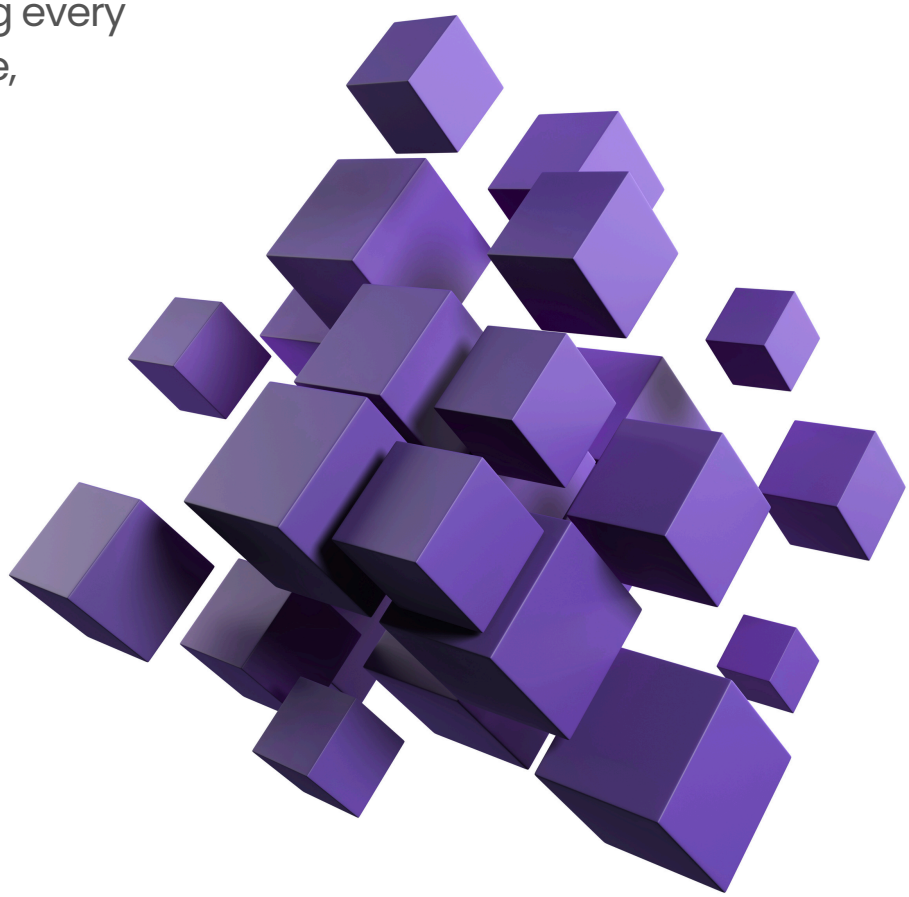
GERT VAN ASSCHE

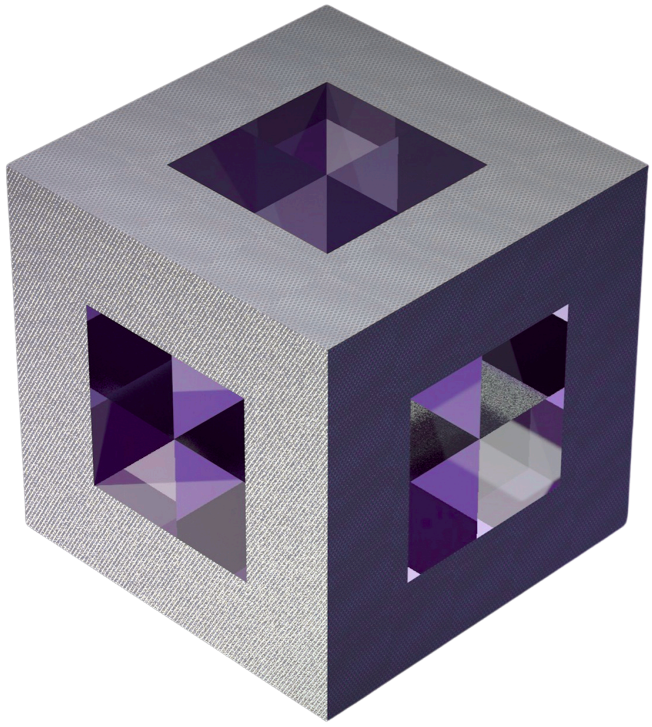
Executive Summary

The increased application of large language models (LLMs) across industries, brings with it new challenges. Specifically, when using LLMs to innovate, it's crucial to recognize the risks they bring, and equally important to have a partner that can help mitigate them.

DATAmundi specializes in helping companies manage and reduce these risks through smart data solutions, expert annotation, and responsible AI deployment strategies.

In this paper, we outline nine key risk categories and how DATAmundi's services directly address each one, pairing every challenge with actionable, proven solutions.





Index

1. Regulatory and Compliance Risk
2. Model Risk
3. Data Risk
4. Operational Risk
5. Security Risk
6. Ethical and Reputational Risk
7. Strategic Risk
8. Legal Risk
9. Human Factors Risk
10. LLMs Going Agentic
11. Conclusion

1. Regulatory & Compliance Risk

LLMs can unintentionally violate regulations such as MiFID II, GDPR, or SEC/FINRA rules. This often stems from the disclosure of sensitive data, insufficient audit trails for automated decisions, or biased outputs that invite regulatory scrutiny.

Solution:

To mitigate these risks, we offer PII Redaction & Sensitive-Term Filtering to prevent exposure of client-specific or personal data. We also help create Golden Datasets that enforce compliance boundaries during LLM output generation.

For example, when an LLM is asked about medical self-treatment, the model is trained to redirect the user to a licensed professional -- thanks to annotations that flag inappropriate advice and ensure compliance with safety standards.

2. Model Risk

LLMs are prone to hallucinations (confident but incorrect answers), non-determinism (inconsistent outputs), lack of explainability, and model drift over time. These issues can undermine trust and usability.

Solution:

To reduce hallucinations and unreliability, we deploy a Dual Annotation & Disagreement Workflow, where both AI and human annotators review outputs, flag inconsistencies, and surface low-confidence responses early.

Our Retrieval Fine-Tuning Services improve RAG systems by enhancing document recall with relevance annotations -- ensuring that retrieved evidence is accurate, useful, and trustworthy. We also support custom benchmarking to detect model drift and maintain performance over time.

In the past we worked on projects reducing hallucinations in financial and legal RAG systems by annotating relevance, trustworthiness, and source type (e.g., avoiding blogs in medical queries but accepting them for travel advice). This helps ensure factual grounding based on query context.

3. Data Risk

LLMs may ingest non-compliant or copyrighted training data, become vulnerable to prompt injection or adversarial inputs, or leak information during inference. Prompts containing non-anonymized PII are particularly dangerous.

Solution:

We help companies manage data risks with PII Detection and Redaction, using both manual annotation and automated pipelines to identify and remove sensitive information.

Where real PII is needed for realistic simulation, we generate synthetic PII that mimics real-world formats.

We also annotate for sensitive phrases and intents, flagging potentially dangerous or inappropriate prompts and responses before deployment.

4. Operational Risk

Failure of LLM-based systems during critical business processes -- such as trading or client reporting -- can have far-reaching consequences. Overreliance on third-party APIs and performance issues during high-load moments are common pain points.

Solution:

To reduce operational fragility, we support the creation of Golden Test Sets for stress-testing LLM systems under simulated load, latency conditions, and edge cases.

Our annotation teams can create scenarios that mimic real-world failures, helping businesses validate LLMs before production rollout.

Additionally, we provide ongoing monitoring data annotation to keep systems aligned with operational needs post-deployment.

5. Security Risk

APIs and prompt interfaces can become entry points for cyberattacks. Sensitive information may leak via compromised models, and insiders may misuse tools by inputting confidential data into unvetted platforms.

Solution:

Security begins with visibility. We provide security-focused prompt annotation to train models to recognize and reject risky prompts.

Our data workflows also include insider misuse detection, helping organizations train models that can identify and block suspicious patterns or inputs.

Finally, we support red-teaming datasets to simulate attacks and strengthen system robustness.

6. Ethical & Repulational Risk

A single biased or inappropriate model output can lead to serious PR issues. Lack of transparency or ethical guardrails around LLM usage may result in client distrust, especially regarding fairness and accountability.

Solution:

We conduct bias detection projects -- including gender bias, ethnic bias, and fairness audits -- using carefully designed multilingual data.

Our teams annotate text, translations, and model outputs across cultures and scenarios to uncover and reduce systemic bias.

We also help create explanation-supportive training data, enabling LLMs to offer clearer rationales for their answers, which helps restore client trust and increase transparency.

7. Strategic Risk

Many LLM pilot projects never scale due to misalignment with business goals or unclear governance. Companies may also over-promise ROI based on unrealistic expectations of model performance.

Solution:

We enable more informed strategic decisions with benchmarking datasets and human-in-the-loop evaluations, giving leaders clarity on model capabilities and limitations.

With our scenario-driven data creation, we help teams test how models behave in their actual business context -- before they commit to scaling.

This reduces the risk of failed deployments and supports sustainable growth.

8. Legal Risk

LLMs might generate content that violates intellectual property laws or contractual agreements. In regulated environments, this can expose the business to litigation or fines, especially when output is used in client-facing tools.

Solution:

Our annotation services help identify copyrighted material, detect derivative content risks, and maintain source attribution integrity. By using vetted datasets with clear licensing, we help organizations avoid legal grey zones.

We also provide vendor oversight data workflows, enabling companies to validate and monitor third-party LLM behavior for contractual compliance.

9. Human Factors Risk

Users may over-trust AI outputs or stop developing essential skills as AI becomes more prevalent. Without training, employees may misuse or misunderstand these tools.

Solution:

We help build training sets for responsible AI use, offering examples of good and bad interactions. These are tailored to your organization's context and used in onboarding or upskilling programs.

We also support feedback loop data annotation, allowing models to learn from user interactions and flag dangerous patterns -- like blind acceptance of risky outputs.

LLMs Going Agentic

When LLMs sit at the heart of agentic, automation-driven workflows — making decisions, generating communications, or triggering downstream actions in real-time — being in control is no longer optional; it becomes mission-critical. In such systems, even a single faulty output can cascade into erroneous decisions, client dissatisfaction, regulatory violations, or financial loss.

Trust in LLM output isn't just about model quality; it's about knowing why the model produced a specific result, under what conditions, and how confident you can be in that result. This is especially important when outputs are consumed without human review, or when multiple agents interact autonomously.

That's why the ability to control, validate, and monitor LLM behavior — through transparent data, rigorous evaluation, and targeted fine-tuning — is essential. At DATAmundi, we help organizations build that trust from the ground up. From gold-standard annotations and red-teaming datasets to relevance fine-tuning and explainability support, we equip you with the data infrastructure and workflows needed to put the LLM on a leash — smart, useful, but never unsupervised.

Because automation at scale is only valuable if it is reliable, accountable, and aligned with your goals. And that starts with data you can trust, and a partner who helps you stay in control.

Conclusion

Every LLM-related risk comes with a matching opportunity for risk reduction — and that's where DATAmundi excels. By embedding robust data practices, annotation expertise, and contextual evaluation into the AI lifecycle, we help companies innovate confidently and responsibly.

Learn more about our services on our Data Services page.

datamundi.ai/ai-data-services