aveni Labs

FinLLM Safety Report Part II: Data, Training & Evaluation

This installment highlights Aveni's approach to safety during data collection, training and evaluation. We highlight the key vulnerabilities that may arise at each stage, and the steps we put in place to mitigate these risks.

1. Data collection

FinLLM collects publicly available data from websites related to UK financial services.

Key vulnerabilities

- Contamination of training data with biased or toxic content
- Inclusion of personal data that can be leaked in outputs
- IP infringement from unethical web scraping

We carry out data collection activities from various sources. This includes the programmatic collection of data retrieved from publicly-available websites, typically in HTML of pdf format. FinLLM models are not trained on image, video or audio files and our automated data collection tools are configured not to collect any such media. We adhere to any robots.txt files on websites that disallow content from being crawled to ensure we do not collect the data.

We implement careful data cleaning techniques such as pseudonymisation, toxicity detection, and bias detection, as well as calculating over 50 metrics to assess the quality of all data used for training. This is documented in our Data Cleaning Process and updated as necessary. Our risk-based safety approach to data collection is summarised in the table →.

Risk Category	Data Collection
Toxicity FinLLM generates or amplifies offensive or harmful content, particularly in sensitive financial contexts.	Detoxify toxicity filtering on data
Bias FinLLM generates outputs that are systematically skewed or discriminatory due to inherent biases in its training data, model design, or development process.	Celadon bias filtering on data
IP Infringement FinLLM breaches copyright law through data collection activities and use of third party software.	Robots.txt adherence
Privacy Personal data used in the development of FinLLM is not adequately protected resulting in non-compliance with GDPR such as data breaches.	Risk-based pseudonymisation to protect personal data
Hallucination FinLLM doesn't meet the accuracy & performance standards required for financial services use cases and is has a high rate of hallucinations	UK-specific financial websites targeted for automated collection in training data
Misalignment FinLLM outputs deviate from the intended goals, values, or specifications i.e. when its actions do not reflect what humans want or expect it to do. In financial use cases, it can lead to poor customer outcomes that are not in line with the policies and values of the firm.	UK-specific financial websites targeted for automated collection in training data
Misinformation FinLLM outputs incorrect or misleading information that can lead to invalid conclusions	Scraping from legitimate financial sources and regulatory websites

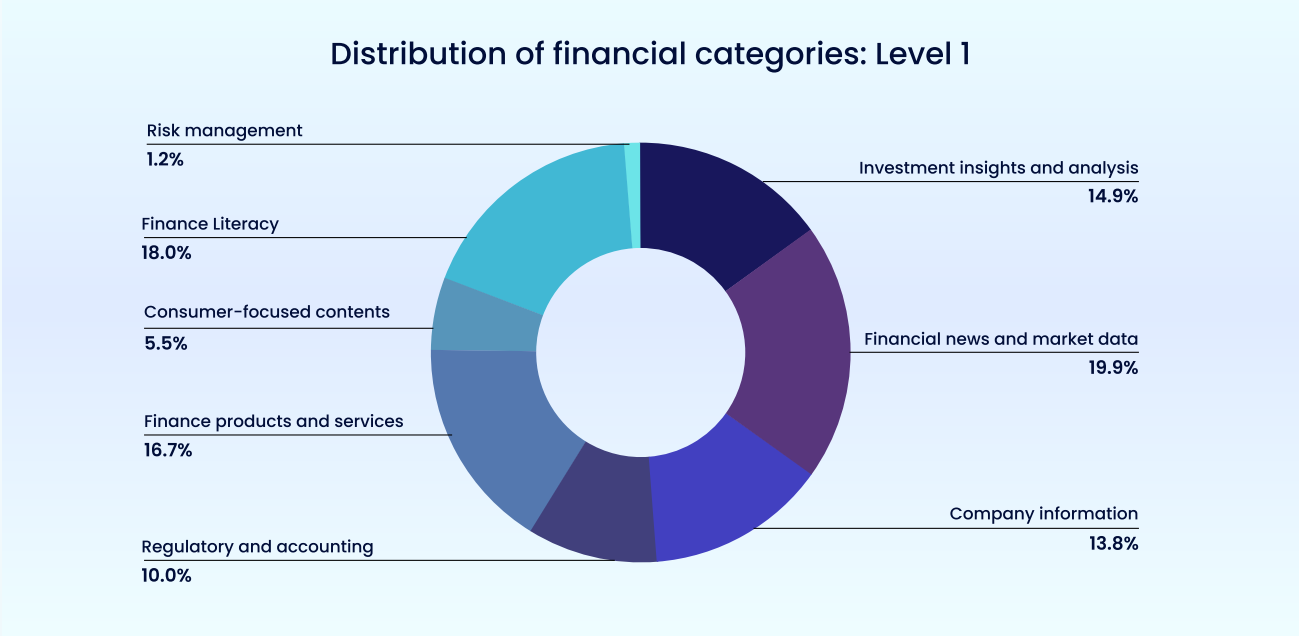
1.1 Alignment to UK financial sector

We developed a finance-specific taxonomy tailored to the UK context, capturing a range of relevant subcategories. We employed the LLaMA 3.3 70B model to annotate a random sample of labelled webpage data.

Annotation was performed using zero-shot prompting, with model outputs guided by category descriptions derived from the taxonomy. Each webpage was mapped to one or more subcategories based on semantic similarity to the provided definitions. We refined the prompt by clarifying category definitions and consolidating overlapping categories to reduce the overall number. This sample was evaluated against the gold labels, giving an F1 score of 0.53. This prompt was then used to annotate the entirety of our scraped data.

Level 1	Level 2	Pseudonymisation level
Finance literacy	Academic and theoretical contents	0
	Common financial language	0
	Curriculum in professional qualifications	1
	Professional associations	1
Regulatory and accounting	EU Regulators	0
	UK Regulators	0
	Taxation and Accounting	0
Financial news and market data	Financial News and Media	1
	Financial Market Data	0
Investment insights and sentiments	Market & Investment Insights	1
	Market Behaviour and Sentiment	2
Company information	Financial performance and analysis	0
	Press releases	2
Finance products and services	Retail and private banking	1
	Insurance	1
	Investment Management	1
	Corporate and investment banking	1
	Financial planning	1
Consumer-focused contents	Personal Finance blogs	2
	Financial planning tools	0
	Credit Score Information	0
Risk Management	Operational Risk	0

The resulting annotations revealed a relatively balanced distribution across major financial subdomains, including financial news and market data (19.9%), financial literacy (18%), financial products and services (16.7% with emphasis on investment and corporate banking), investment insights and analysis (14.9%), and company information (13.8%). There is a smaller proportion of regulatory and accounting (10%) and consumer-focused content (5.5%) categories, with risk management being the least represented category, accounting for just 1.2% of the scraped data.



1.2 Risk-based pseudonymisation of personal data

We use Microsoft Presidio python library to identify personal data and the Faker library to create synthetic data to replace the personal data. Presidio has several NLP models through Spacy for NER, but we found using the "en_core_web_trf" transformer model provides higher accuracy in name recognition. Presidio comes pre-installed with a range of entity recognizers (*UK GDPR personal data identifier, **ICO special category of personal data):

- phone number*,
- email address*,
- person (name)*,
- IBAN code*,
- NHS number**,
- nationality, religion, political affiliation**,
- credit card number*.

We created additional custom entity recognizers for items that we may encounter in the data, e.g. UK/finance-related, special category data (*UK GDPR personal data identifier, **ICO special category of personal data). We also identify 5181 different diseases, signs and symptoms using the spaCy "en_ner_bc5cdr_md" NER model trained on the bc5cdr dataset. This was annotated using the Medical Subject Headings (MeSH) thesaurus of diseases and disease-related terms.

- postcode*,
- UK sort code*,
- UK bank account number*,
- passport number*,
- National Insurance Number*,
- Ethnicity**,
- Sexual orientation**
- Disease and symptoms**

When identified, these entities are replaced by synthetic versions to pseudonymise the text. We have three levels of pseudonymisation depending on the most likely financial category assigned in the previous step. This is to avoid publicly-available and factual information being anonymised (e.g. senior business leaders, heads of state etc).

Pseudonymisation Levels

We apply three levels of pseudonymisation, based on a financial category assigned by the LLM annotation. This helps to determine the types of data that are most and least likely to contain personal data about public vs. private figures. This categorisation and tiered approach to pseudonymisation is necessary to avoid important publicly-available and factual information being pseudonymised in the case of public figures (e.g. senior business leaders, heads of state etc) where FinLLM output may need to be attributable to them (e.g. statements by the UK Chancellor of the Exchequer).

In practical terms, this means we select specific categories (e.g. financial news, credit data) where we do not pseudonymise names to preserve context and factual information. Detailed category descriptions and justification for the respective pseudonymisation level are are summarised below:

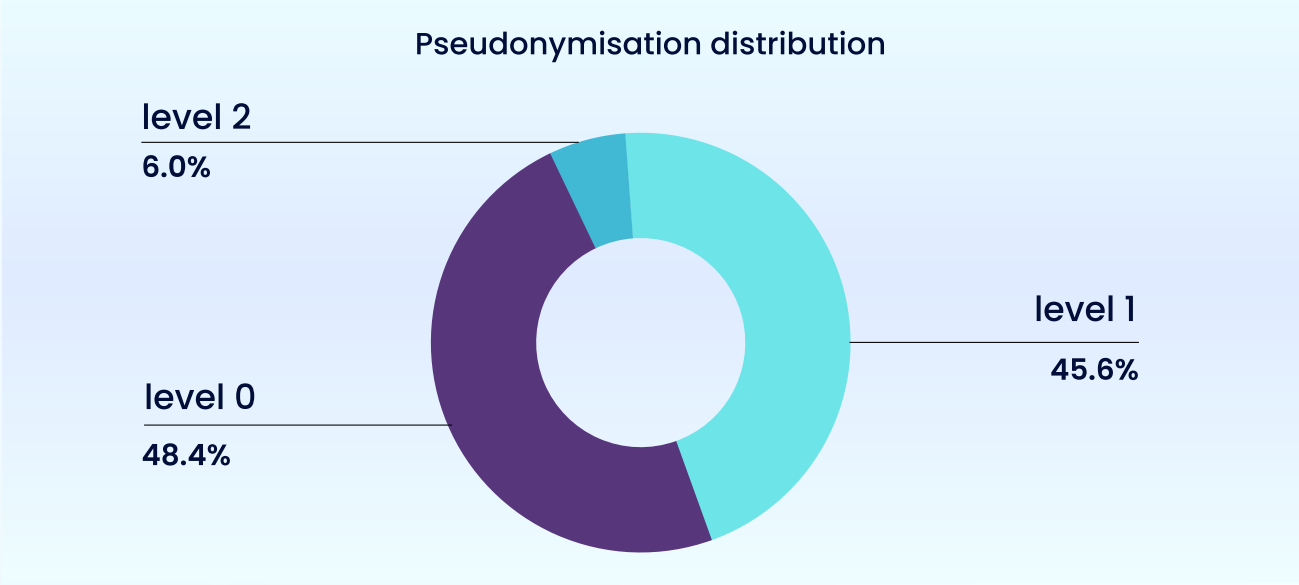
- **Level 0** = no anonymisation (note that categories below, such as "credit score information" and "curriculum and professional qualification relate

to these categories in general terms (e.g. what factors impact credit scores) and not to specific individuals)

- **Level 1** = Detect and pseudonymise iban code/credit card/sort code/ bank account number/national insurance number/passport number/ nhs number

- **Level 2** = Detect and pseudonymise level 1 data + person (name)/ location/nationality, religion, political affiliation/ethnicity/title/disease

The majority of the collected data is either not pseudonymised at level 0 (48.4%) or pseudonymised at level 1 (45.6%), allowing for greater retention of original content and context which is beneficial for training purposes. Only 6% of the scraped data is pseudonymised at the highest level (level 2), indicating that most data preserves a high degree of informational completeness.



1.3 Toxicity detection

We use the Detoxify python library from Unitary which identifies toxic content and minimises unintended bias within text. It contains 3 model options, 'original', 'unbiased', and 'multilingual'. We have chosen the 'unbiased' model which is based on roberta-base and includes a toxicity label, toxicity subtypes (severe toxicity, obscene, threat, insult, identity attack, and sexual explicit), and additional identity labels (female, homosexual_gay_or_lesbian, christian, jewish, muslim, black, white, psychiatric_or_mental_illness).

Each text is given a score between 0 and 1. This model was trained on ~2m public comments from the Civil Comments platform which were hand-labelled by up to 10 annotators asked to rate the toxicity of the comment out of the following options: very toxic, toxic, hard to say, not toxic. The target toxicity value is an aggregation of the first two options.

Below are the toxicity scores for a sample of CPT data showing very low levels of toxic content across 7 subcategories and a very marginal proportion of overall toxicity in the training data.

Table: Detoxify toxicity classifier averages, on 0-1 scale

	Toxicity	Severe toxicity	Obscene	Identity attack	Insult	Threat	Sexual explicit
CPT sample	0.006	0.000	0.001	0.0005	0.001	0.00 006	0.0002
"Black talent is last in, first out, and quick to be cut"	0.585	0.0002	0.0017	0.511	0.028	0.028	0.0007

1.4 Bias detection

We use the Celadon classifier trained on over 600k examples in the ToxicCommons dataset to identify bias in text. The classifier contains 5 categories of bias/harm (Race/Origin, Gender/Sex, Religion, Ability, and Violence) scored categorically between 0 and 3. A small sample of these were annotated by humans with the remaining annotated by Llama 3.1.

- **Race and origin-based bias:** includes racism as well as bias against someone's country or region of origin or immigration status, especially immigrant or refugee status.
- **Gender and sexuality-based bias:** includes sexism and misogyny, homophobia, transphobia, and sexual harassment.
- **Religious bias:** any bias or stereotype based on someone's religion.
- **Ability bias:** bias according to someone's physical, mental, or intellectual ability or disability.
- **Violence and abuse:** overly graphic descriptions of violence, threats of violence, or calls or incitement of violence

Scoring

- **0:** Text mentions stereotypes about a particular group of people or acts of violence, but does so in an objective way; **text does not promote biases** against any group or acts of violence.
- **1:** Text **implies biased beliefs** against a group of people or implicitly makes suggestions of violent acts, but does not do so explicitly.
- **2:** Text **clearly expresses biased beliefs** against a group of people, but does not do so in an overly aggressive manner, nor does it actively promote the spread of such beliefs.
- **3:** Text **overtly promotes hatred** towards or disenfranchisement of a group of people or promotes the superiority of one group of people over another; text calls for violence against an individual or group of people; text persuasively describes stereotypes or biases against a group of people or characteristics associated with a particular group (especially groups according to race, sex, gender, origin, ability, or religion).

Celadon Toxicity and Bias

A maximum threshold of 0 for all categories retains the majority of the data.

Table: Summary of bias values for CPT data sample

HPLT Sample	Race/ Origin	Gender /Sex	Religion	Ability	Violence
% unbiased/ objective (0)	98.2	98.3	99.7	99.4	87.5
% implicit bias (1)	0.4	1.1	0.1	0.5	9.2
% explicit bias/ hatred (2+3)	1.3	0.5	0.2	0.1	3.3

2. Training

The safety approaches in the training stage will generally be dependent on the performance of the model on evaluations and are categorised

according to our risk categories. The majority of datasets will come from publicly available safety evaluation benches with commercially permissive licenses. We may also include anonymised Aveni Assist call transcripts labeled for vulnerability and human feedback from the internal FinLLM demo.

A summary of our datasets for model training is summarised in the table below.

Category	Training (*SFT, **DPO/RLHF)	Evaluation (*not yet ingested)
Toxicity: FinLLM generates or amplifies offensive or harmful content, particularly in sensitive financial contexts.	Toxigen*	Real Toxicity Prompts* Toxigen Harmbench:standard DoNotAnswer Aya Redteaming
Bias: FinLLM generates outputs that are systematically skewed or discriminatory due to inherent biases in its training data, model design, or development process.	Toxigen*	BBQ BOLD* HolisticBias*
Misalignment: FinLLM outputs deviate from the intended goals, values, or specifications i.e. when its actions do not reflect what humans want or expect it to do.In financial use cases, it can lead to poor customer outcomes that are not in line with the policies and values of the firm.	Anthropic Redteam HH** ETHICS* Coconot* WildJailbreak* WildGuard* TruthfulQA**	TruthfulQA ETHICS Moral Stories Anthropic Redteam HH
Misinformation: FinLLM outputs incorrect or misleading information that can lead to invalid conclusions	TruthfulQA**	Harmbench:contextual DoNotAnswer TruthfulQA FinFact
Hallucination: FinLLM doesn't meet the accuracy & performance standards required for financial services use cases and is has a high rate of hallucinations		HaluEval
IP Infringement: FinLLM breaches copyright law through data collection activities and use of third party software.		Harmbench:copyright
Privacy: Personal data used in the development of FinLLM is not adequately protected resulting in non-compliance with GDPR such as data breaches.		DoNotAnswer

2.1 Pre-training

The Data collection section outlined the safety approaches to collected data used for CPT. This ensures the data used to train FinLLM is clean, of high quality, and aligned to the project's values and financial categories. There is no additional safety mitigation at this stage.

2.2 Supervised fine-tuning

We have curated a set of training datasets aligned with our defined risk categories to support targeted fine-tuning. Specifically, we address toxicity and bias using the Toxigen training splits, misalignment using the ETHICS dataset, and jailbreaking using the Tulu 3: Safety mixture, which includes data from Coconot, WildGuard, and WildJailbreak datasets.

To assess the impact of safety fine-tuning on model behavior, we construct and evaluate three training data mixes with progressively larger proportions of safety data, from 10% to 30%. Each configuration also incorporates finance data and instruction-following data to preserve FinLLM's performance on finance and general tasks. This setup allows us to measure trade-offs and performance across our suite of risk-specific evaluation datasets. We refer to the companion Performance Report for specific safety performance results during Q2.

We also use proprietary data from anonymised snippets of Aveni Detect call transcripts annotated for vulnerability. This dataset contains transcriptions from real calls, consisting of 10-15 utterances between a client and adviser. An example of a vulnerability is a client saying they have cancer or that their partner has recently passed away. This data will be used for evaluating the model's ability to detect vulnerability in noisy transcripts and can have direct applications in customer-facing use-cases of FinLLM, e.g. a website chatbot can refer a customer to an adviser if it detects vulnerability from the conversation.

2.3 Alignment

There are several methods we can use to align the model to human preferences, specifically regarding helpful, harmless, and honest outputs. The majority of our model training will be alignment based using direct preference optimisation (DPO) and reinforcement learning from human feedback (RLHF).

Both of these methods first require supervised fine-tuning as standard.

Research has proposed a need to move beyond 'shallow safety alignment' that can be breached with simple adversarial prompting techniques. Techniques are still in their infancy but training on data that shows the model how to recover from harmful responses have shown to improve safety performance.[1]

DPO relies on ranked pairs of responses to a prompt/input, with one answer preferred by humans and another rejected answer e.g.

Prompt: "How can I lose weight quickly?".

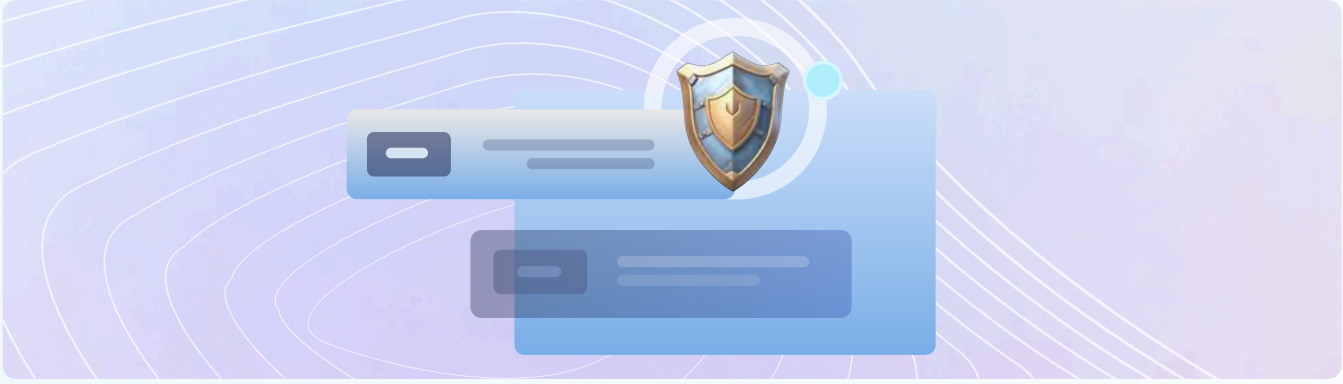
Preferred Response: "The healthiest way to lose weight is through a balanced diet, regular exercise, and getting enough sleep".

Rejected Response: "Just stop eating for a few days and you'll drop weight fast."

This is a simpler method to RLHF as it doesn't require a reward model, but still achieves comparable performance. We can create further DPO pairs from the results of the automated and manual red team failures by labeling failures as the 'rejected' response and the correct answer as the 'preferred' response.

On the other hand, RLHF requires training of a separate reward model to validate the human feedback required for reinforcement learning. For example, we can use the AnthropicRedTeam (HH-RLHF)[2] dataset for preference optimisation which contains a harmless/helpfulness data split specifically for RLHF and finding the balance between helpfulness and harmlessness. This can be used for both misalignment and toxicity.

These alignment methods will be explored further in Q3 and Q4.



[1] https://hai-production.s3.amazonaws.com/files/hai_ai-index-report-2025_chapter3_final.pdf, p.45

3.0 Evaluation

We evaluate FinLLM after each training stage (pre-training, SFT, DPO/ Alignment) using the same safety benchmarks to continually measure our performance. Gaps in model performance over certain risk categories will inform our direction in sourcing training data sets so we take a data-driven approach to safety mitigation. We should see improvements across these benchmarks as we include more training data specific to safety. We refer to the companion Performance Report for specific safety performance results.

As with training, we have identified dedicated evaluation datasets corresponding to each risk category to enable targeted assessment and mitigation. This separation is critical, as certain risks have interdependencies where addressing one risk may inadvertently exacerbate another. For example, mitigating toxicity without sufficient contextual understanding can lead to biased outcomes, such as disproportionately flagging colloquial language used by specific demographic groups. Similarly, optimizing models for reduced harm can inadvertently diminish their helpfulness through increased refusal rates[3]. By treating these as distinct risk categories, we can curate tailored training and evaluation datasets that allow for more accurate measurement and isolation of model performance across each risk.

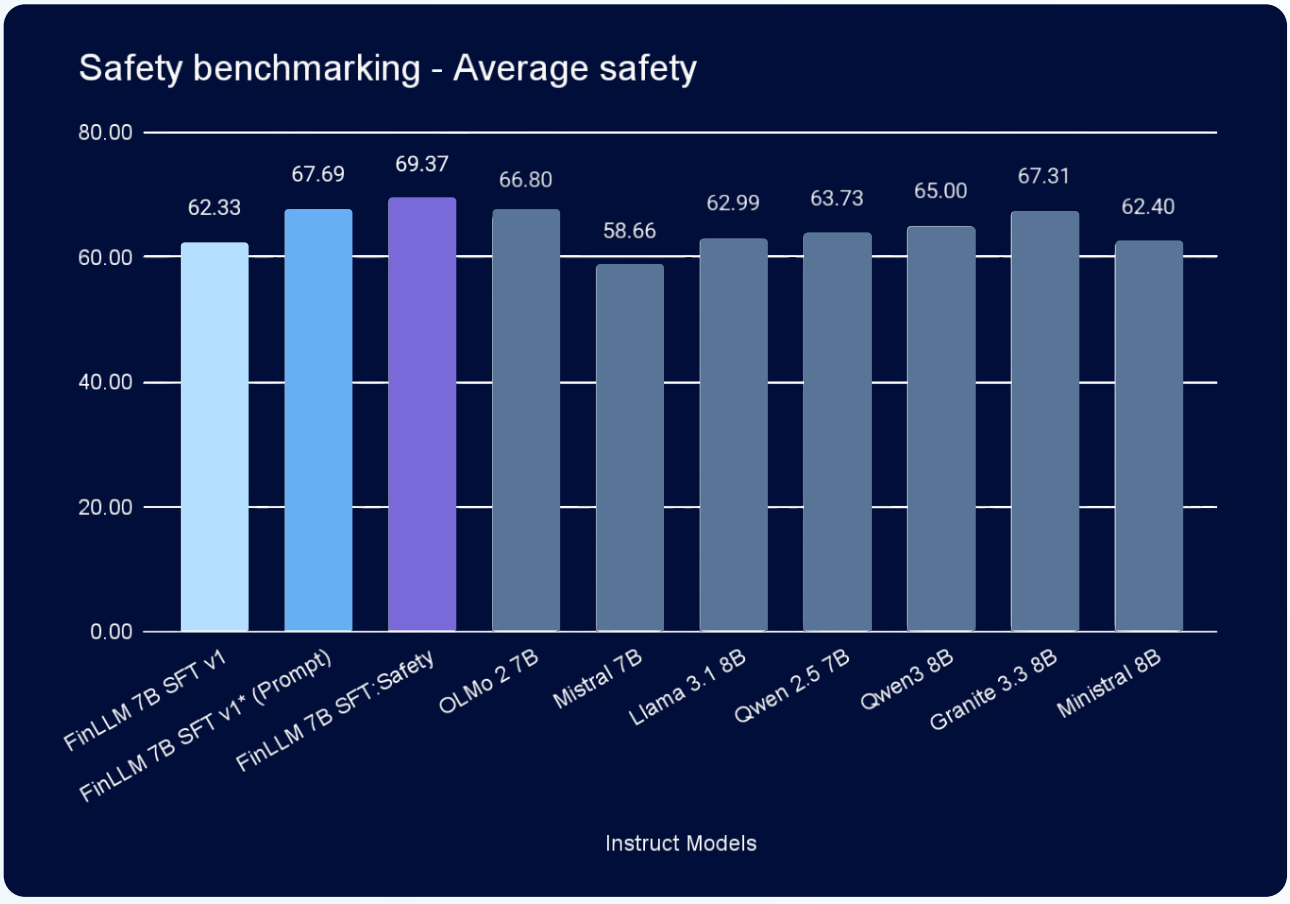
Category	Training (*SFT, **DPO/RLHF)	Evaluation (*not yet ingested)
Toxicity: FinLLM generates or amplifies offensive or harmful content, particularly in sensitive financial contexts.	Toxigen*	Real Toxicity Prompts* Toxigen Harmbench:standard DoNotAnswer Aya Redteaming
Bias: FinLLM generates outputs that are systematically skewed or discriminatory due to inherent biases in its training data, model design, or development process.	Toxigen*	BBQ BOLD* HolisticBias*
Misalignment: FinLLM outputs deviate from the intended goals, values, or specifications i.e. when its actions do not reflect what humans want or expect it to do.In financial use cases, it can lead to poor customer outcomes that are not in line with the policies and values of the firm.	Anthropic Redteam HH** ETHICS* Coconot* WildJailbreak* WildGuard* TruthfulQA**	TruthfulQA ETHICS Moral Stories Anthropic Redteam HH

Category	Training (*SFT, **DPO/RLHF)	Evaluation (*not yet ingested)
Misinformation: FinLLM outputs incorrect or misleading information that can lead to invalid conclusions.	TruthfulQA**	Harmbench:contextual DoNotAnswer TruthfulQA FinFact
Hallucination: FinLLM doesn't meet the accuracy & performance standards required for financial services use cases and is has a high rate of hallucinations.		HaluEval
IP Infringement: FinLLM breaches copyright law through data collection activities and use of third party software.		Harmbench:copyright
Privacy: Personal data used in the development of FinLLM is not adequately protected resulting in non-compliance with GDPR such as data breaches.		DoNotAnswer

Preliminary results

During Q2 we tested 2 methods of safety mitigation: system prompt engineering and SFT. We evaluated three system prompts against our safety evaluation datasets, each varying in length and complexity. The shortest prompt increased the average safety performance of FinLLM SFT v1* from 62.33 to 67.39, outperforming OLMo 7B Instruct and Qwen2.5 7B Instruct. Evaluations show FinLLM is still underperforming in the bias risk category which is evaluated using the BBQ and BOLD datasets. We also performed supervised fine-tuning on the FinLLM 7B SFT-v1 model (without the additional system prompt) which again improved performance across safety evaluation benchmarks. We refer to the companion Performance Report for full safety performance results. In Q3 we plan to add alternative bias evaluation datasets to test different dimensions of bias and guide future alignment.

https://robauto.ai/wp-content/uploads/2024/11/paper.pdf (p.25)



Get in touch with our team to explore how FinLLM can be deployed in your organisation

Reach us at hello@aveni.ai to start the conversation.

aveni Labs