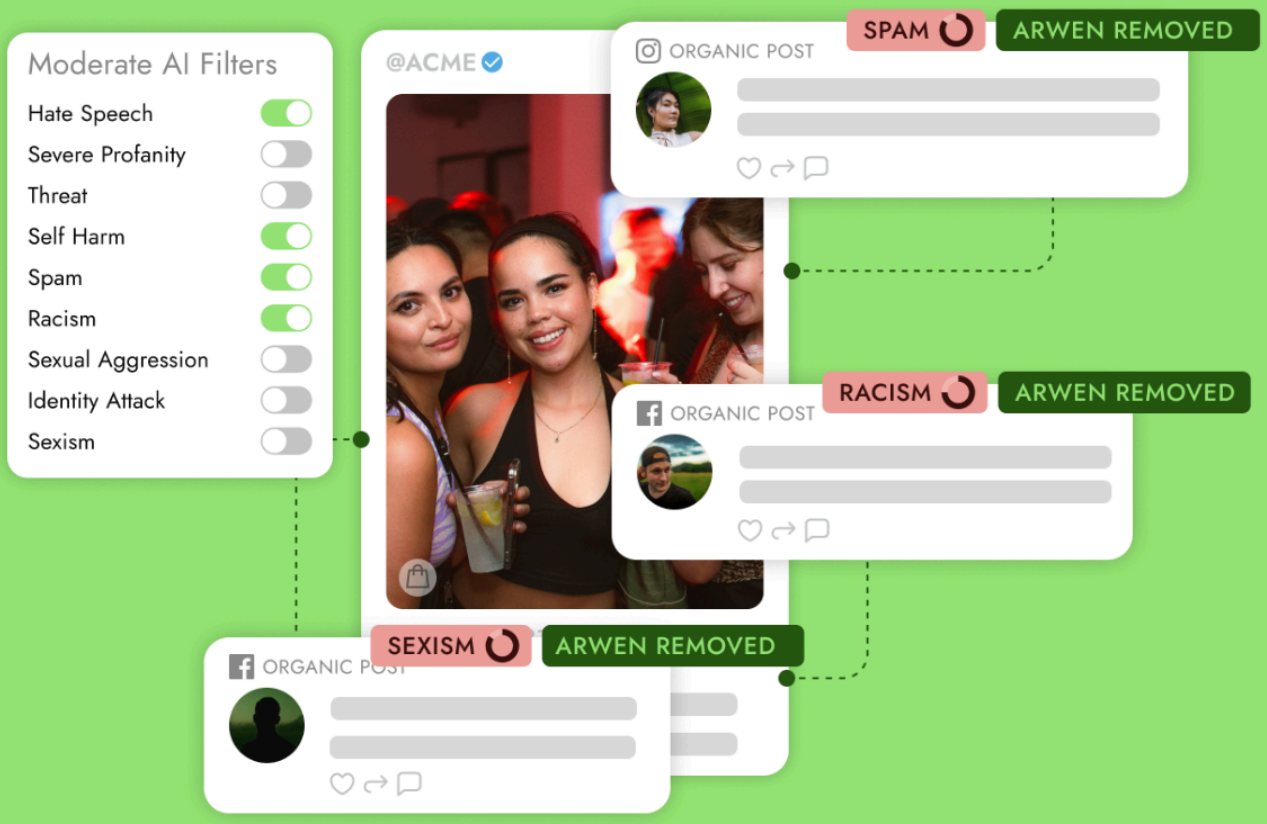
Buyers Guide to AI auto-moderation solutions

What's an LLM? How is it different to Classical Machine Learning? Do keywords still matter? Is it a magic wand? **We bust the myths, explain the jargon, and summarise the solutions.**



arwen

You AI Co-pilot for social media

The Buyers Guide to AI Auto-Moderation Services

Over the last few years, an abundance of automated moderation techniques have entered the market - Closed AI, Open AI, Public AI, LLM, Generative AI, Keyword Filters etc etc. It's enough to make your head spin. Some less scrutable vendors are using this confusion to peddle some distinctly sub-optimal solutions. Plenty of our customers came to us saying they were fully covered, only to discover that they weren't really - and that many tricks were being missed. This is just what happens when new stuff is being developed really quickly!

So, if you're confused, fear not! Everyone is 😊

In this article we're going to demystify the world of AI auto-moderation techniques. Suffice to say, all have pros and cons that need to be considered before adoption.

What are the main considerations when selecting an AI auto-moderation solution?

The three main considerations when looking at any AI auto-moderation technique are:

1. **What comment collection technique is being used?** Are comments being scraped, or are they being collected via an official API? Each has pros and cons
2. **What moderation technology is being used?** Is it really AI, or an older technology just badged as AI? A good gauge of the AI's sophistication is "how much do humans have to oversee the AI?" If the answer is "a lot" then it's probably not a great set up.
3. **Other things to consider:** What languages are covered? What formats can it handle (text, gifs, emojis, video)? What ethical and environmental considerations are important to the provider?

First we'll look at the different moderation technologies, then we'll look at the different comment collection techniques, and we'll close with a few other considerations.

An overview of different moderation techniques

Type 0: Human moderation

Okay calling it Type 0 is a bit tongue-in-cheek. Human moderation was once the only way. Someone would literally have to view each comment in turn and decide whether it needed to be moderated. Obviously this is labour intensive and quite expensive. Assembling a team of people to be able to 24/7 spot all the different types of toxicity and spam, across every language, in real time is an expensive business. Which is what has driven all the innovation we'll be looking at.

Key point though is: all moderation systems, no matter how "magical" the AI, need human involvement. Don't let anyone tell you their moderation system is fully autonomous. Humans have to be involved or things can quickly get out of hand. So the real question for any provider is: what is the balance between machine and human in your system?

- Do humans do most of the moderation, with a machine handling the obviously toxic comments? This solution will work perfectly well but will cost you a lot
- Do they review every or some AI decisions? If they're using more basic moderation technology
- They review patterns and false positive rates and use that insight to train the AI? This is the future - the moderator makes sure the machine is working properly, manually intervenes when it isn't, but - crucially - trains the AI so the next decision is better. Each moderator may be more expensive per hour, but they are highly skilled moderators + trainers.

Pros:

- When you have a great moderator, accuracy is amazing. You can't beat it - even with the best of what current technology has to offer.

Cons:

- Humans have to sleep, so you'll need a provider with people available around the clock, likely in all the countries with languages you want covered
- Humans are inconsistent, get tired and are prone to change their opinions over time
- Humans don't share the same knowledge, make mistakes and have bias - unconscious or otherwise. A good AI will usually have a lower error rate than a human (this feels

counter-intuitive. It's called the [Appeal to Nature or Naturalistic Fallacy](https://en.wikipedia.org/wiki/Appeal_to_nature)).

- It's a pretty horrible job - leading to high turnover of staff, which leads to a big hiring and retraining overhead
- It's the most expensive technique we'll cover

Conclusion:

I love humans. I am one! But there are some jobs humans weren't designed to do. Arwen reviews thousands of comments a day, moderating 25 types of toxic comments across 30 languages, in real time, without ever needing a day off. Asking a group of humans to do this when we no longer have to, feels a bit ick.

Type 1: Keyword filtering

This has historically been the technique of choice. Human moderators analyze the domain to be protected and develop long lists of keywords. These are loaded into a pretty primitive moderation technology, which then moderates comments that use one or more of the keywords.

Pros:

- Simple to implement with strong domain knowledge.
- Built-in systems in most networks allow for keyword lists, and many social media management tools offer equivalent features.
- Useful for unique, niche words related to your brand or products.
- Automatically hides comments containing the specified keywords.
- Low tech cost but high resource costs.

Cons:

- High error rate or false positives when keywords are used in non-toxic contexts.
- Users can bypass filters by misspelling words or using emojis, known as algospeak.
- Maintaining keyword lists becomes labor-intensive.
- Cannot moderate images, gifs, and videos.
- Ineffective for languages other than the one the list is maintained in.
- Ineffective against spammers who continually change their language patterns.
- Relies heavily on human oversight, increasing labor intensity.

Conclusion:
Still prevalent but unsophisticated compared to other technologies. Expensive to run and maintain.

Type 2: Pattern Matching

Pattern matching goes a few steps beyond keyword filtering. It checks comments for the presence of known patterns, handling some obvious misspellings.

Pros:

- Better than keyword filtering, catching basic misspellings.
- Reduced risk of false positives, requiring fewer human reviews.

Cons:

- Still requires maintenance of keyword lists.
- Struggles with algospeak.
- High error rate.
- Ineffective against spammers who continually change their patterns.

Conclusion:
Only marginally more accurate than keyword filtering, with minimal resource benefits. It's also easier to manage and can catch some behaviors keywords cannot.

Type 3: Machine learning models

Machine learning models are able to exploit vast amounts of unstructured data usually to do one thing really well. In the moderation space, this means detecting specific types of harmful content, like racism, spam, or homophobia with a very low error rate. With the advent of Large Language Models (LLMs), these are often now called "Classical machine learning models" or "Classification models", to distinguish them from LLMs, which are also Machine learning models, just much more advanced.

Pros:

- More accurate than keyword lists, with expert teams continually training the models.
- They need a smaller data set to learn from than the LLMs we'll come onto, and the way they process comments is simpler (relatively speaking) so they can be easier to build and maintain
- Can analyze entire comments rather than relying on individual keywords.
- Not as energy intensive as LLMs, so better environmentally.

Cons:

- They're not so good at taking context into account - so can struggle with regional nuance if they haven't been trained on comments from that region.
- They aren't good at detecting things outside of their area of expertise - so a model that's great at detecting anti-semitism might be useless at detecting racism. Any machine learning moderation solution will have to have multiple models at work. For instance, at Arwen we have 25 models.
- Ongoing training of the AI is critical - ie telling them when they make a mistake so they won't do it again - as they aren't built to self-learn, in the same way that LLMs can.
- Slow to change. You often have to rely on a longer feedback loop to the team that maintains them to improve them.
- Most models are not multi-modal - ie they only work only on one format eg text. So you'll need separate models for images, gifs, and videos.
- High integration cost due to limited compatibility with social media management tools. Most moderation providers will do that integration under the bonnet of their own solution.

Conclusion:

Highly accurate for specific types of content at a low cost. If a provider you're looking at is using them, make sure they have a good and close relationship with their model providers. If they have built their own models, find out how they work.

Type 4: Large Language Models

This is the most recent, very sexy stuff that has set the world on fire. We'll first cover LLMs in general and then break them out into the two subsets: Open and Closed.

Large Language Models use complex deep learning systems which allow them to learn from huge datasets. Most have spent months / years crawling the internet, devouring content, to develop enormous datasets, with an ever growing number of parameters. To give you a sense of scale, one training set that OpenAI used is called Common Crawl, which is made up of over 250 billion web pages spanning 17 years, to which 3–5 billion new pages are added each month. This "knowledge" makes them multi-functional, and able to analyse, detect and generate content with high, often uncanny, accuracy.

Pros:

- Highly accurate moderation decisions.
- Quick and easy to change for specific customers.
- Prompt engineering makes them highly accessible and means you can get them to perform decisions in a very tailored way
- Can handle context - so they can go beyond "is this a toxic comment?" to ask "Is this a toxic comment in the context of this brand / this moment / this post / this image?"

Cons:

- Prone to hallucination, where the AI generates incorrect information. If your provider is using them, make sure they have protocols in place to reduce this risk, such as Retrieval Augmented Generation (RAG).
- Substantially more expensive per decision than Machine Learning Models.
- Very energy intensive. They eat carbon. So best used sparingly.
- Relies on the moderation provider having good third party management with their LLM provider. LLMs are expensive to build. We're talking billions of dollars. No moderation provider will own its own LLM, so it's important to understand which one they're using, why and how well they are using it. If that underpinning LLM were to change dramatically, that will inadvertently impact the service they provide you.
- Often overkill for 90% of the bread-and-butter moderation work. Like using a Ferrari to do the simple job of detecting a swear word.
- Requires expert human oversight for accuracy, usually from a trained and experienced Data Scientist. Make sure they have them.
- Quicker to change, as the Prompt Engineer can modify prompts to make the LLM change behaviour quicker than traditional machine learning models

Conclusion:
LLMs offer amazing new opportunities for moderation but come at a considerably higher cost - both in cash terms and to the environment. Also don't be fooled into thinking their magical and faultless, the risk of hallucination is high. Think of LLMs as the kooky but very knowledgeable professor you go to to check those weird outlier comments - they know a lot but are prone to very confidently giving eccentric results. In comparison your machine learning models are your dogged, trusted, low-energy workers tirelessly detecting 90% of the toxicity with very high accuracy.

4a: Open AI LLMs

Open AI, or "Public AI", is a type of AI model that is publicly accessible and can be modified by anyone. It's also made more confusing because the maker of ChatGPT is actually called Open AI! But there are others like Microsoft's Co-Pilot and Google's Genesis. All are in their own way Open AIs.

The main thing here is that Open AI's are sometimes perceived to not offer the levels of security required by many enterprises. It's important to note that open AI doesn't mean insecure and closed AI doesn't mean secure. However perceptions continue to shift across the industry, so if in doubt, check with your Information Management and Security team.

4b: Closed AI LLMs

Closed AI aren't publicly accessible, though they may be trained on publicly available data. Organisations like Microsoft now offer their LLM Bing Chat as either an Open or Closed version. The latter is preferred by enterprises because they offer better data security and confidentiality, particularly when it comes to confidential business data. It's less of a consideration with moderation, but is critical in other areas. At Arwen we also offer an Engage product, which automatically refers to internal "facts" to generate recommended replies to comments. Again, check with your Information Management and Security team for guidance.

Type 5: RAG (Retrieval Augmented Generation)

RAG uses "guardrails" to ensure AI retrieves organizational facts, like moderation guidelines, for more accurate decisions.

Pros:

- Ensures AI decisions are informed by agreed facts.
- Generates content in the client's brand tone and voice
- Essential for handling nuanced toxic comments.

Cons:

- Can be complex to implement.

Conclusion:

RAG reduces the risk of hallucination in an LLM, so prevents AI from making strange moderation decisions. It's very valuable if you want nuanced moderation.

Type 6: Integrated approach

This technique is the meta technique! It combines all of the aforementioned techniques into one integrated solution, using the right technology for each part of the moderation task.

Pros:

- Uses keywords for unique words targeted at your brand.
- Pattern Matching supports spam and URL detection
- Classical machine learning models provide accurate, energy-efficient and low cost detection of routine toxic comments.
- LLMs make decisions on the tougher, more nuanced comments, where context and intention need to be taken into account.
- RAG ensures moderation decisions align with organizational facts and brand values.

Cons:

- Integrating multiple techniques can be complex and resource-intensive, so make sure your supplier has the right operating model to manage it all

- Any team working this way needs deep tech chops - we're talking degree-level data science, prompt engineering etc, ideally with decent reference-able experience on-the-ground in moderation

Conclusion:
Combining techniques offers the best moderation solution, balancing quality and cost. It means the provider can tailor a solution to you. You don't have to buy a Ferrari when you only need a Ford.

An overview of different collection techniques

Type 1: Scraping comments

Many providers scrape comments from social networks, then pass them through their technology for moderation.

Pros:

- No need for individual social media profile owners to authorize collection of comments.

Cons:

- Collects only a small percentage of comments.
- Slower collection hampers quick moderation.
- Information management risks from third-party handling.
- Legal risks in regions where scraping is prohibited.
- Scraping often lacks useful metadata.
- Violates most social network rules, risking status.
- Reliance on unreliable third-party tools.
- Separate process required for instructing the network to moderate a comment.
- Generally more expensive in the long-term.

Type 2: Collecting comments via an approved API

This method uses an authorized API to collect comments in real-time.

Pros:

- Collects all comments at best instantly or at least quickly.
- Legally compliant in all regions.
- Compliant with social network rules.
- Collects comment metadata.
- Reliable and robust, minimizing surprises.
- No third-party information management risks.
- Uses API to notify the network to moderate comments.

Cons:

- Users must authorize their profiles, which can be seen as a barrier.
- Requires Instagram Business Accounts linked to Facebook pages for moderation.

Conclusion:

API-based collection is the preferred method, ensuring completeness, compliance, and reliability.

Other considerations

So we've covered the main points, now let's look at some secondary considerations:

1. **Languages.** If your community is global, you need a moderation solution that can travel with you. If your community is smaller and nation-specific, this is less important. Global buyers should look for two things
    a. Ability to detect toxicity and spam comments in multiple languages**. Translation is pretty straightforward machine-learning stuff, so most good providers should be able to do most languages
    b. Take into consideration different cultural references and nuance**. Usually this can only be achieved by training the AI on large amounts of regional- and culturally-specific comments. So the more experience the provider has in a geography, the better.
2. **Formats**. Social media users express themselves in comments through a range of media - text, gifs, emojis and video. As moderation software has improved, the more committed

trolls have moved to algo-speak (where words are intentionally mis-spelled or hidden by emojis) or to putting words in images and gifs. A few specific considerations:

   a. **How are gifs and videos moderated?** A good provider will be able to handle them. Usually by taking a number of frames from the gif / video and analysing them for offensive words and imagery. These technologies are getting very accurate, but be aware, they usually come with higher processing costs. So if video and gifs aren't a major issue for you, it's less of a concern
   b. **How are comments treated?** Does the solution look at individual characters and words, or does it review the whole comment in combination before making a decision? The latter is better. A combination of the two is best.

3. **Approach to innovation.** The world is moving fast - and even faster because AI is an arms race, with bad actors like spammers, bots, government sponsored troll farms and even your run-of-the-mill bedroom troll using AI to beat the moderation software. So it's critical that the provider you choose is investing in R&D to stay ahead and continually closing gaps as they emerge. Providers who have evolved from being human moderation providers (i.e. large teams of people reviewing comments) have a fundamentally different business model underpinning them compared to a natively AI provider.
4. **Ethics**. Let's be honest, moderation is a nasty business. It's like removing the trash - a necessary evil, but no-one wants to do it. It's also historically been an expensive business (because humans need to be paid) with a real drive to continually lower costs. In response to this, we've seen some pretty unethical solutions. Here are a few things to look out for:

   a. **If humans are going to do your moderation, where are they?** Using a provider who is paying people in Lagos dollars-a-day to moderate your community's racist comments is not a good look. Especially when those moderators report burnout and maltreatment. Make sure your provider has a clear ethical moderation supply chain. At Arwen one of our pledges is "no humans were harmed in the making of this decision", so we make sure we have a wellbeing programme for anyone who is routinely exposed to severe comments.
   b. **If your solution is going to rely on AI, how energy intensive is it?** AI is recognised to be energy hungry. Those processors are under glaciers for a reason. It's the responsibility of all of us to reduce our carbon consumption. It's hard to trace impact through global systems, but a rule of thumb is that Open AI is more energy intensive than single-purpose classical machine learning models -

so if you're relying on Open AI it's probably overkill. Like using a Ferrari to find a swear word. Ideally your solution should use a balance of techniques, appropriate to the job.

5. **Transparency**. If you moderate without transparency, you will create a vacuum. And as I'm sure we all remember from physics class, "nature abhors a vacuum". If you leave it, it will be filled by rumours of censorship. This is a very common fear for clients. You may feel uncomfortable "coming out" as moderating your social community, because it's been the wild west for years and your users are used to having the run of the town. But trust me when I say: it's always better to be transparent. Explain somewhere what you will and won't tolerate, and tell your community about it. Look out for providers who help you through this process. At Arwen we include this consultancy and drafting service in our onboarding service.

Conclusion

Hopefully this article gives you some useful guidance and checkpoints so you can make a confident decision when selecting an AI auto-moderation solution.

My one take-away would be: don't listen to anyone who tells you that their AI is "better" or that someone else's is "the worst". Everyone is doing things differently, with relative merits depending on what you need. Don't be fooled into buying a Ferrari when actually you only need a reliable Ford. Here are some scenarios:

"I only get the occasional profanity that I want removed within the day"

Make full use of each network's own moderation features, as they are pretty good at the basics in some geographies, just slow. If you want to go further, download a profanity keyword list and add them as keywords to your social listening tool (OFCOM in the UK has a good one).

"I get lots of racism and I need it removed super fast"

Go for a provider who uses a good single-purpose anti-racism AI moderation model and make sure they use an API to collect your comments, not scraping. You shouldn't need an Open AI model unless you get lots of weird edge-cases.

"I can never anticipate what I'm going to get, but we get trolled and spammed globally in really creative ways and I need it continually removed in real time"

Make sure your provider uses a fully integrated approach, with Open AI, Closed AI and keywords working in parallel, collecting comments by API. And make sure they have strong global language coverage.

You get the picture. Different solutions for different needs. Different courses, different horses.

Technology is always changing, so we'll be returning to update this as it does. But if you spot that we're missing something, please do get in touch with us at info@arwen.ai

We're proud to be an integrated and ethical AI service provider, collecting comments via robust and reliable APIs, and moderating using a combination of technologies, so that each one plays to its strengths. It's what we believe makes us a world leader, with such high client retention rates and happy team members.

If you want to explore social media moderation further, get in touch at info@arwen.ai or visit our website to book a time and we can have a chat.

www.arwen.ai