Short answer: Mistral only started training LLMs in 2023 (the company didn't exist in 2017), so "since 2017" really means "from 2023–2025." Below I go through *every* Mistral LLM family I could find, with a focus on:

- How each model was **aligned** (SFT, DPO, RL, reward models, moderation stacks, etc.)
- Which **benchmarks** they used
- How their **values alignment story evolved** from one generation to the next

I'll flag clearly where the public record is detailed vs. where it's vague.

---

1. Sources & scope

Primary sources from Mistral

- Research papers:
    - *Mistral 7B* paper (small, fast, open 7B model) (ar5iv)
    - *Mixtral of Experts* (Mixtral-8x7B) arXiv paper
    - *Magistral* (reasoning model built on Mistral Medium 3) arXiv + blog
- Official model/blog pages (Mistral site + docs):
    - Mistral 7B, Mixtral, Medium 1, Large 1, Small 1, Small 2, Small 3, Medium 3, Large 2, Large 3, Nemo 12B, Mathstral, Codestral, Ministral (3B/8B), Pixtral 12B & Pixtral Large, Devstral, Ministral 3 family, Mistral Moderation, etc. (Mistral AI)
- Legal / AI governance pages (model registry, risk class, lifecycle, safety processes) (Mistral AI Legal Center)

Secondary / scholarly / industry sources

- NVIDIA & AWS blog posts, HF model cards for Nemo, Pixtral, Devstral etc. (NVIDIA Blog)
- External evaluations:

- AgentHarm (ICLR 2025) – harmful agent behaviors, includes Mistral Small 2 & Large 2 (ICLR Proceedings)
- "Do LLM Benchmarks Test Reliability?" – includes Mistral Small 1 (researchgate.net)
- CountQA (MLLM counting reliability) – includes Pixtral Large 2411, Mistral Small 3.1 (researchgate.net)

Where public docs are *silent* (e.g., exact reward-model details for some proprietary models), I'll say so explicitly rather than guessing.

---

2. Catalog of Mistral LLM families (2023–2025)

> **Before 2023** there are no Mistral LLMs; the company was founded mid-2023.(Wikipedia)

2.1 Core general-purpose model families

Approximate chronological order:

Year	Family	Sizes / key versions	Open?	Main references
2023	**Mistral 7B** (base + Instruct v0.1/0.2/0.3)	7B dense	Apache-2.0 (weights)	Paper + HF cards + blog (ar5iv)
2023	**Mixtral 8x7B** (base + Instruct)	MoE, 8×7B experts (~47B total, 13B active)	Apache-2.0	"Mixtral of Experts" paper + HF/model cards
2023	**Mistral Medium 1.0**	Proprietary dense	Closed weights; enterprise	Docs + AI governance (Mistral AI Legal Center)

2024	**Mistral Large 1** (24.02)	Flagship dense frontier model	Proprietary	“Au Large” blog + Azure partnership docs (Mistral AI)
2024	**Mistral Small 1.0**	Proprietary small dense	Closed; enterprise	Docs + governance (Mistral AI Legal Center)
2024	**Mistral Small 2.0** (open weights)	Small dense	Open-weight, MRL	Docs + governance (Mistral AI Legal Center)
2024	**Mistral Nemo 12B (base/instruct)**	12B dense	Apache-2.0	Mistral blog, NVIDIA & HF model cards (Mistral AI)
2024	**Pixtral 12B**	12B multimodal (text+vision)	Apache-2.0	Mistral Pixtral 12B blog (Mistral AI)
2024	**Ministral 3B & 8B** (base + instruct)	3B/8B dense	MRL, then some Apache	“Un Ministral, des Ministraux” blog + HF (Mistral AI)
2024	**Pixtral Large 124B**	124B multimodal (built on Large 2)	Open weights (MNPL style), inference via APIs	Mistral blog + docs + AWS/Bedrock docs (Mistral AI)
2024	**Mistral Large 2** (24.07)	123B dense	Weights open for research (MRL)	“Large Enough” blog (Mistral AI)

2024	**Mistral Moderation**	Safety classifier LLM	API only	AI governance models list (Mistral AI Legal Center)
2025	**Mistral Small 3 (24B, 24.01)**	24B dense open	Apache-2.0	“Mistral Small 3” blog + stats (Mistral AI)
2025	**Mistral Small 3.1 / 3.2 (24B)**	Incremental updates	Apache-2.0	Mistral Small 3.1 blog + HF card (Mistral AI)
2025	**Mistral Medium 3**	Frontier dense multimodal	Closed weights, self-hostable	“Medium is the new large” blog, docs (Mistral AI)
2025	**Devstral Small 1.x**	Small coding/agentic LLM	Apache-2.0	Devstral blog + HF card (Mistral AI)
2025	**Magistral Small/Medium**	Reasoning models (RL-tuned)	Proprietary access	Magistral blog + paper
2025	**Mistral 3 family**: Large 3 + Ministral 3 (3B/8B/14B, base/instruct/reasoning)	Large = sparse MoE (675B total, ~41B active); Ministral 3 dense	Apache-2.0 open weights	“Introducing Mistral 3” + Nvidia blog + external explainers (Mistral AI)

2.2 Specialist / domain models

Family	Role	Notes

Mathstral 7B	Math/STEM reasoning	Instructed math model derived from Mistral 7B; uses a strong *reward model* at inference for majority voting. (Mistral AI)
Codestral 22B / Codestral Mamba	Coding & FIM	Large code-specialized models (22B dense; state-space Mamba variant). (Mistral AI)
Devstral	Coding agents	Optimized for SWE-Bench-style multi-file coding agents; open-source. (Mistral AI)

I'll focus on these families in the alignment timeline.

---

3. Early generation (late-2023): Mistral 7B & first instruct models

3.1 Mistral 7B base

- **Release**: Sept 2023 as "Mistral 7B" base under Apache-2.0.(ar5iv)
- **Benchmarks**: MMLU, commonsense reasoning (Hellaswag, WinoGrande, PIQA, SIQA, OBQA, ARC-Easy/Challenge, CSQA), reading comprehension (BoolQ, QuAC), math (GSM8K, MATH), code (MBPP, HumanEval). (ar5iv)

Alignment / values story

The *Mistral 7B* paper and blog focus on *capabilities*, not safety:

- The paper describes pretraining on a large (undisclosed) web+code corpus and measurable performance gains; it does **not** describe a dedicated safety / RLHF pipeline. (ar5iv)
- The official blog explicitly notes that **Mistral-7B-Instruct "does not have any moderation mechanism"** and is intended as a "quick demonstration" of capabilities, with safety delegated to downstream users. (Mistral AI)

So at this point, “values alignment” essentially equals: *“we instruction-tuned a base model for helpfulness, but we don’t ship built-in safety.”*

3.2 Mistral-7B-Instruct v0.1 / v0.2 / v0.3

Training / alignment

HF model cards tell us:

- v0.1: “instruct fine-tuned version of the Mistral-7B-v0.1 generative text model using a variety of publicly available conversation datasets.” (Hugging Face)
- v0.2: same, but on the improved base Mistral-7B-v0.2 (32k context, different RoPE settings). (Hugging Face)
- v0.3: again an instruction-tuned variant of base 7B v0.3 with extended vocabulary and function-calling support. (Hugging Face)

Importantly, none of these cards describe RLHF, DPO, or explicit safety tuning; they just say “instruction-tuned on public conversation datasets,” and the main paper says the instruct model has **no moderation** baked in. (ar5iv)

Benchmarks

- Instruct variants are evaluated on MT-Bench and LMSys Chatbot Arena (reported in external blogs and in Mixtral paper comparisons), but we don’t have a dedicated “alignment” paper for them.

Alignment characterisation (2023)

- **Helpfulness:** Pure SFT on instruction / chat datasets.
- **Safety:** No integrated moderation; users are told to add their own guardrails.
- **Values:** Implicitly whatever the public instruction datasets encode; no explicit fairness/toxicity or refusal objectives.

This is a pretty “classic 2023 open model” posture: capabilities first, safety up to deployers.

---

4. Generation 1.5 (late-2023 / early-2024): Mixtral, Medium 1, Large 1, Small 1

4.1 Mixtral 8x7B (base & Instruct) – the first explicit alignment story

The **Mixtral of Experts** paper is crucial for understanding how Mistral thought about alignment at the end of 2023.

Base model bias / values evaluation

Mixtral 8x7B is an MoE architecture (8×7B experts, 13B active parameters) with strong performance across benchmarks, but the paper also includes **bias evaluations**:

- They measure bias on:
    - **BBQ (Bias Benchmark for QA)** – social bias in QA tasks.
    - **BOLD** – bias in open-ended generations (gender, profession, religion, political ideology, race).
- They report that **Mixtral presents *less bias* than Llama-2-70B** (higher BBQ accuracy and more positive BOLD sentiment with similar variance) and highlight this as something to be *"corrected by fine-tuning / preference modelling"* later.

So here, for the *base model*, "alignment" is still observational: they *measure* bias but don't yet say they actively trained to reduce it.

Instruct model alignment pipeline

Section 4 of the paper is explicit:

> "We train *Mixtral – Instruct* using **supervised fine-tuning (SFT)** on an instruction dataset followed by **Direct Preference Optimization (DPO)** on a paired feedback dataset. Mixtral – Instruct reaches a score of 8.30 on MT-Bench, making it the best open-weights model as of December 2023…"

Key points:

- **Two-stage alignment**:
    - SFT for helpfulness & basic instruction following.
    - **DPO** for *preference alignment* with human (or human-filtered) preference pairs.

- **Benchmarks used as alignment proxies**:
    - **MT-Bench** (automatic & human-rated) for conversational quality.
    - LMSys Arena (human battle evaluation) – Mixtral-Instruct beats GPT-3.5-Turbo, Gemini Pro, Claude-2.1, and Llama-2-70B Chat.
- **Bias / safety**: still mostly measured on the *base* model; there is no public description of extra adversarial safety fine-tuning, though DPO preferences probably penalize egregious outputs.

Evolution vs. Mistral-7B

Compared to the 7B instruct models, Mixtral 8x7B introduces:

- **Formal preference optimization (DPO)** → the first clearly documented move beyond pure SFT.
- **Bias metrics** (BBQ/BOLD) → alignment evaluation is broadened beyond accuracy.

This is the first "modern" alignment story in Mistral's stack.

4.2 Mistral Medium 1.0 (Dec 2023)

- Enterprise-grade base model, no published technical paper. Docs call it *"our first SOTA enterprise-grade model."* (Mistral AI)
- AI governance marks it as a **"General Purpose AI Model" (GPAIM)**, with model lifecycle & risk management obligations attached. (Mistral AI Legal Center)

Alignment details

Public docs don't expose the training pipeline, but by analogy with Mixtral & later models, it's safe to infer:

- SFT (and likely preference tuning) for instruction following and refusal behaviour.
- Deployed with **external moderation** when served via la Plateforme or partner APIs.

However, because Mistral hasn't released a Medium-1 paper, we **don't** know:

- Whether they used RLHF vs. DPO vs. other preference-optimization.
- What exact alignment benchmarks they targeted internally.

4.3 Mistral Large 1 (24.02) & Small 1 (24.02)

- **Large 1** launched Feb 26, 2024 as Mistral's flagship proprietary model, first available on Azure. (Mistral AI)
- **Small 1** launched same day as a compact enterprise model. (Mistral AI Legal Center)

Public material focuses on capabilities (multilingual, code, reasoning) and doesn't spell out the alignment recipe, but we can see an evolution:

- These models are part of an **AI-governed portfolio** with documented model lifecycle, classification and obligations under EU AI Act as "GPAIM" or frontier models. (Mistral AI Legal Center)
- They're typically served behind **platform-level moderation and guardrails**, unlike the raw 7B Instruct.

However, the exact combination of SFT/DPO/RLHF is not publicly documented.

---

5. 2024: Open-source expansion & modular safety

2024 is the year where Mistral starts to **split alignment concerns across components**:

1. **Open-source, instruction-aligned models** (Nemo, Small 2, Ministral, Mathstral, Codestral)
2. **Frontier proprietary models** (Large 2, Medium-like models)
3. A separate **Mistral Moderation** model used for guardrailing. (Mistral AI Legal Center)

5.1 Mistral Nemo 12B (July 2024)

- **Open-source 12B model** co-developed with NVIDIA, released July 18, 2024 under Apache-2.0. (Mistral AI)

- Designed as *"our best multilingual open-source model"* and an upgrade over Mistral 7B. (Mistral AI)

Alignment & benchmarks

- NVIDIA's blog describes Nemo as delivering *"leading accuracy on popular benchmarks across common sense reasoning, world knowledge, coding, math, and multilingual and multi-turn chat tasks."* (NVIDIA Developer)

- HF cards and partner docs emphasise:

    - Pretrained & **Instruct** versions.

    - 128k context, strong coding and multilingual performance. (Hugging Face)

Alignment details are not as richly documented as Mixtral; however, it's positioned as:

- A **drop-in replacement** for Mistral-7B in chat agents, implying:

    - SFT for instruction following.

    - Some preference tuning (though the exact method isn't publicly specified).

- Safety: served via platforms with their own moderation; the open weights themselves are not described as "heavily safety-aligned."

5.2 Mathstral 7B (July 2024)

- Released July 16, 2024 as a **math/STEM-specialized 7B instructed model**, derived from Mistral 7B, under Apache-2.0. (Mistral AI)

Alignment for mathematical truthfulness

The blog is unusually explicit about *math-alignment*:

- Mathstral is "an *instructed* model" with a 32k context window, targeting advanced mathematical reasoning. (Mistral AI)

- Benchmarks:

    - MATH (56.6% base, up to 68.37% with majority voting, 74.59% with a strong reward model among 64 candidates). (Mistral AI)

    - MMLU and breakdown by subject vs. Mistral 7B. (Mistral AI)

- Critically, they mention a **reward model used at inference**:

    - They sample many candidate solutions and select the best using a reward model → an *inference-time* alignment of outputs towards mathematically correct solutions. (Mistral AI)

This is a form of *task-specific alignment* (truthfulness for math):

- **Training-time alignment:** SFT on math reasoning data.

- **Inference-time alignment:** best-of-N with a reward model to pick more reliable answers.

No explicit safety/values alignment beyond math is described.

5.3 Ministral 3B / 8B (Oct 2024)

- Released in October 2024 as “les Ministraux”: two small models (3B and 8B) for local/edge use; base and instruct variants. (Mistral AI)

- Benchmarks:

    - They compare both **pretrained** and **instruct** models against Gemma 2 2B, Llama 3.2 3B, Llama 3.1 8B and Mistral 7B across multiple categories (MMLU, coding, etc.), with tables and graphs for both base and instruct. (Mistral AI)

Alignment details

The blog talks about:

- “Instruct models” (Ministral-3B-Instruct, Ministral-8B-Instruct) that significantly outperform similar-sized models. (Hugging Face)

- Alignment recipe is *implicitly* SFT + some preference tuning, but they do not specify DPO/RLHF.

Safety posture:

- Small open-weights models for local deployment → largely **alignment-light** beyond instruction following; safety left to deployers + external moderation (e.g., via Mistral Moderation or third-party guardrails).

5.4 Pixtral 12B & Pixtral Large 124B

- **Pixtral 12B**: open-weights multimodal model trained to be a "drop-in replacement for Mistral Nemo 12B," focused on multimodal reasoning while preserving text capabilities. (Mistral AI)

- **Pixtral Large (124B)**: announced Nov 18, 2024, built on top of **Mistral Large 2**, with open weights and 128k context. (Mistral AI)

Alignment & benchmarks

- Pixtral 12B and Large emphasize **"best-in-class multimodal reasoning"** over documents, charts, and natural images, plus strong text instruction following. (Mistral AI)

- Benchmarks include:

    - Vision-language tasks (doc QA, chart QA, natural images) from Mistral's own internal suite and external tests like CountQA, where *Pixtral Large 2411* is evaluated as a top open-source MLLM. (researchgate.net)

Alignment, again, splits into:

- **Text alignment** inherited from Mistral Large 2 (see next section).

- **Vision alignment**: appropriate responses to visual content; public docs don't detail safety measures (e.g., handling faces, sensitive images), though these are likely handled via moderation/guardrailing in production.

5.5 Mistral Large 2 (24.07) – "Large Enough"

This is one of the clearest statements of Mistral's *alignment philosophy* as of mid-2024. (Mistral AI)

Key points from the "Large Enough" blog:

- Large 2 is a 123B-parameter model with 128k context, trained with a “very large proportion of code” and extensive multilingual data. (Mistral AI)

- **Benchmarks**:

    - MMLU (84.0% base) – they position it on the performance/cost Pareto frontier of open models. (Mistral AI)

    - Code: HumanEval, MBPP, MultiPL-E. (Mistral AI)

    - Math: GSM8K and MATH. (Mistral AI)

    - Instruction/alignment: MT-Bench, WildBench, Arena Hard. (Mistral AI)

Alignment-specific passages

The blog devotes an explicit section to **“Instruction following & Alignment”**:

- They state they *“drastically improved the instruction-following and conversational capabilities”* of Large 2 compared to Large 1. (Mistral AI)

- They emphasize:

    - Reducing hallucinations: *“fine-tuning the model to be more cautious and discerning… trained to acknowledge when it cannot find solutions or does not have sufficient information.”* (Mistral AI)

    - Balancing alignment metrics with output length: they point out that longer outputs can inflate scores on MT-Bench etc., so they **explicitly optimize for concise responses** as a constraint. (Mistral AI)

However, they still **don’t specify**:

- The exact preference-optimization method (DPO vs RLHF vs RLAIF).

- The datasets used for alignment, or explicit safety objectives beyond hallucination reduction.

Values alignment shift vs. Mixtral / 7B

- With Large 2, Mistral frames alignment explicitly as:

- **Truthfulness & calibration** → less hallucination, willingness to say "I don't know."
- **User experience** → concise, business-friendly outputs.
- **General safety** → measured on alignment benchmarks (MT-Bench, WildBench, Arena Hard) but still no detailed normative spec.

In parallel, Mistral launches a **Mistral Moderation** model (Nov 2024) as a separate component for toxicity/harm filtering, formalized in their AI Governance hub. (Mistral AI Legal Center)

So, the architecture by late-2024 is roughly:

1. **Base + instruct models** (Nemo, Large 2, Small 2, etc.) tuned for helpfulness & reliability.
2. **Dedicated Moderation model** used to filter/score prompts and outputs.
3. **Governance docs** assigning risk categories and obligations.

---

6. 2025: Third-generation models & RL-style alignment (Medium 3, Small 3, Magistral, Devstral, Mistral 3)

6.1 Mistral Small 3 (24B) & its updates

- **Mistral Small 3** (Jan 30, 2025): 24B dense model, open weights under Apache-2.0. (Mistral AI)
  - Benchmarks: competitive with Llama-3.3-70B and Qwen-32B; strong on MMLU & HumanEval. (LLM Stats)
- **Small 3.1** (Mar 17, 2025): improved text performance, multimodal understanding, 128k context; marketed as *"the best model in its weight class."* (Mistral AI)
- **Small 3.2** (July 2025): incremental update, with HF card stating that it matches or slightly improves Small 3.1 on benchmarks. (Hugging Face)

Alignment

Public material emphasises:

- Open weights → local control; no forced moderation.
- Strong performance on instruction benchmarks (MMLU, GPQA, HumanEval, MATH) vs other small models (Gemma-3, GPT-4o-mini, Claude 3.5 Haiku). (Analytics Vidhya)

But again:

- No explicit description of DPO/RLHF. Given the Mixtral precedent and their performance profile, it's *very plausible* they use SFT + preference tuning, but that's not spelled out publicly.

From a **values** perspective:

- These models are optimized primarily for *capabilities* and general helpfulness (chat, coding), with safety left to moderation stacks and user fine-tuning.

External work on reliability/bias:

- **CountQA** (2025) uses *Mistral-small3.1-24B* and *Pixtral Large 2411* among open-source models, showing that their **counting reliability is strong but not perfect**, roughly on par with other state-of-the-art open MLLMs. (researchgate.net)

6.2 Mistral Medium 3 – "Medium is the new large" (May 7, 2025)

- Medium 3 is a frontier multimodal model marketed as providing ~"large-model performance at 8× lower cost," self-hostable on 4+ GPUs. (Mistral AI)
- It is **not open-weights** but can be deployed on-prem or via partners; AI governance classifies it as a frontier-grade GPAIM. (Mistral AI Legal Center)

Benchmarks

Blogs and third-party analyses report:

- Strong performance on MMLU, HumanEval, GSM-8K/MATH, GPQA, as well as multimodal tasks. (Mistral AI)

Alignment characteristics

Publicly disclosed aspects:

- Medium 3 is clearly optimized for:
    - **Enterprise-grade robustness** (reasoning + coding + multimodal).
    - **Cost-performance** balance (i.e. more inference-efficient alignment than very large models). (Mistral AI)
- However, **training & alignment pipeline details are not published**:
    - No paper or blog section explaining whether they used RLHF, DPO, RLAIF, etc.
    - No explicit bias metrics.

We can still see a pattern:

- By this point, Mistral has a separate **Magistral** reasoning model built *on top of Medium 3* (see below), where they *do* describe an RL pipeline.
- This suggests Medium 3 is treated as a *capability backbone*, with specialized alignment layers added for reasoning (Magistral) and safety (Moderation).

6.3 Magistral – RL-aligned reasoning on top of Medium 3

Magistral is a major step in **capability alignment via RL**.

- Magistral models (Small & Medium) are reasoning-optimized instruction models that sit on top of Medium 3.

The arXiv paper describes:

- A **reinforcement-learning pipeline** that optimizes a policy (Magistral) for success on complex reasoning tasks (including tool use and code) without training on chain-of-thought data; instead, CoT emerges from the RL process.
- RL is run on top of a frozen or partially-frozen Medium-3-like base model, using a reward function that captures success on tasks and discourages unnecessary verbosity.

From a values-alignment angle:

- This is not *safety RLHF* aimed at toxicity or fairness; it's **task-level RL** aimed at:
    - Better reasoning (more steps, fewer errors).
    - Less "overthinking" and more calibrated answers.
- Safety is still delegated to:
    - The base model's alignment.
    - External moderation (Mistral Moderation).
    - Possibly constraints in the reward function (e.g. penalizing obviously wrong or non-answers).

But, importantly, Magistral shows Mistral is now comfortable using **RL to shape model behaviour**, not just SFT/DPO.

6.4 Devstral – agentic coding alignment

Devstral (May 21, 2025) is Mistral's open-source **agentic coding LLM** built with All Hands AI. (Mistral AI)

- It is optimized for:
    - Using tools (e.g., editors, shells).
    - Editing multiple files.
    - Performing multi-step SWE-Bench Verified tasks (Devstral is #1 open model on this benchmark). (Mistral AI)

Alignment features (capability-wise)

Public docs emphasize:

- Tool-use proficiency and multi-file reasoning, not safety.
- Strong performance on SWE-Bench Verified (closed-book coding tasks). (Mistral AI)

As with Medium 3 & Magistral:

- Training details aren't fully exposed, but given the agentic nature and SWE-Bench focus, Devstral likely uses:
    - SFT on agent traces.
    - Possibly RL (SWE-Bench success as reward), though this is not confirmed publicly.

Safety / values alignment here is more **procedural**:

- The risk is that Devstral can modify code bases; so Mistral encourages:
    - Using **guardrails and workspace restrictions**.
    - Combining Devstral with Mistral Moderation and human oversight. (These recommendations are standard in Mistral's governance materials, even if not spelled out per-model.) (Mistral AI Legal Center)

6.5 Mistral 3 family (Dec 2025): Large 3 + Ministral 3 (3B/8B/14B)

Released Dec 2, 2025 as **"Mistral 3"** – a family of fully open-weights multimodal models. (Mistral AI)

- **Mistral Large 3**:
    - Sparse MoE with 675B total parameters, ~41B active per token.
    - Vision-capable, multilingual, open weights under Apache-2.0.
    - Designed for high-end reasoning and long-context workloads. (Mistral AI)
- **Ministral 3** (3B/8B/14B, each with base, instruct, and reasoning variants):
    - Dense, multimodal, designed for edge/local deployment.
    - All Apache-2.0, open weights. (Mistral AI)

Benchmarks

- Nvidia's blog on "NVIDIA-Accelerated Mistral 3 Open Models" and third-party write-ups show:

    - Top open-source scores on a wide range of text and multimodal benchmarks (MMLU, GPQA, HumanEval, GSM8K, MATH, vision QA). (NVIDIA Developer)

- Large 3 ranks near the top of open-source leaderboards (e.g. LMArena non-reasoning leaderboard), though still slightly under top closed models. (DataCamp)

Alignment

At time of writing, the Mistral 3 blog emphasises:

- Frontier-class **multimodal** alignment (text + images).
- Full **openness** (Apache-2.0 for all sizes).
- Training on NVIDIA Hopper GPUs, and optimization for deployment from GB200 clusters down to edge devices. (Mistral AI)

But it **does not yet publish**:

- The exact alignment recipe (SFT vs DPO vs RLHF vs RLAIF).
- Detailed safety evaluation (bias, toxicity, jailbreak resistance).

So relative to earlier generations:

- **Capabilities alignment** is clearly very strong (benchmarks, reasoning).
- **Values & safety alignment** are likely handled via:
    - Instruction tuning & preference optimization (by analogy with Mixtral & Large 2).
    - External moderation (Mistral Moderation).
    - Possibly new safety evals (not yet public).

Given the fully open weights and small models (3B/8B/14B), Mistral 3 represents a *shift towards openness* while relying heavily on deployer-side safety.

---

7. Cross-cutting: Mistral Moderation & AI governance

Mistral's **AI Governance / Legal Center** shows:

- A catalog of all models with lifecycle (release/retirement), type ("Base Model", "Model Version"), and classification (GPAIM, frontier model, open-source GPAIM). (Mistral AI Legal Center)

- A separate **"Mistral Moderation"** model (Nov 6, 2024) for content moderation/guardrailing. (Mistral AI Legal Center)

This indicates an **architectural shift in alignment**:

1. Early models (Mistral 7B, Mixtral) embed alignment in the LLM itself (SFT + DPO), but safety is mostly ad-hoc.

2. From 2024 onwards, Mistral formalizes:

    - A **dedicated moderation model** that can sit in front of any base/instruct LLM.

    - Governance documentation that classifies risks and describes evaluation and red-teaming obligations (especially for frontier models like Large 2, Medium 3, Large 3). (Mistral AI Legal Center)

External research like **AgentHarm** and **reliability benchmark** papers find that:

- Modern Mistral models (Small 2, Large 2, Small 1) perform **comparably to other top models** in both usefulness and harmfulness tests – i.e., safety alignment is *better, but far from perfect*. (ICLR Proceedings)

This is consistent with a regime where:

- Mistral is explicitly aligning for helpfulness, truthfulness, and "don't blatantly misbehave," but they **do not try to fully police every downstream use** via heavy baked-in constraints, especially for open-weights models.

---

8. How Mistral's values alignment evolved over time

Putting it all together, here's the trajectory:

Phase 1 (late-2023): *Capability-first, minimal safety baked in*

Models: Mistral 7B base & Instruct, early Medium 1, early instruct models.

- Alignment method:
    - SFT on public conversation / instruction datasets.
    - No documented preference optimization.
- Safety/values:
    - *Mistral-7B-Instruct explicitly has "no moderation mechanism."* (Mistral AI)
    - Bias is not a major focus in the 7B paper; more of a performance story.
- Benchmarks as proxies:
    - Accuracy benchmarks (MMLU, GSM8K, HumanEval).
    - Some general-purpose chat quality metrics (LMSys Arena).

Net effect: The models are **very capable** but **lightly aligned**; values are mostly inherited from training data and SFT instructions.

Phase 2 (late-2023 / 2024): *Formal preference alignment & bias measurement*

Key model: Mixtral 8x7B (base + Instruct).

- First fully documented **SFT + DPO** pipeline for Instruct variants.
- Base model evaluated for **bias** using BBQ and BOLD, with Mixtral showing *less bias* than Llama-2-70B.
- Alignment benchmarks (MT-Bench, LMSys Arena) become central metrics.

Evolution vs phase 1:

- Mistral **adopts modern preference-optimization** (DPO).
- They **measure bias** explicitly and talk about correcting it via fine-tuning.

- Safety is still not deeply elaborated (no dedicated safety paper), but values alignment is now *methodologically explicit*.

Phase 3 (2024): *Modular alignment – open models + separate moderation*

Models: Nemo 12B, Mathstral, Ministral 3B/8B, Pixtral 12B, Mistral Small 2, Mistral Large 2, Pixtral Large, Mistral Moderation.

- Open models like Nemo, Mathstral, Ministral, Small 2, Pixtral 12B:
    - Focus on **capability alignment** (instruction following, math truthfulness, local deployment).
    - Provide open weights; safety is largely delegated to developers.
    - Mathstral explicitly uses **reward-model-based majority voting** for mathematical reliability. (Mistral AI)
- Frontier models like Large 2:
    - Add a **stronger emphasis on truthfulness and calibration** (reduced hallucinations, willingness to say "I don't know"). (Mistral AI)
    - Use alignment benchmarks (MT-Bench, WildBench, Arena Hard).
- **Mistral Moderation**:
    - Introduced as a separate model, integrated into platform and governed as an AI system on its own. (Mistral AI Legal Center)

Shift in values alignment philosophy:

- Rather than trying to make each LLM "safe for everything," Mistral:
    - Keeps core models **useful and relatively unconstrained**.
    - Provides **dedicated moderation** to catch harmful content.
    - Formalizes **AI governance** (model lifecycle, GDPR/AI-Act obligations).

Phase 4 (2025): *RL & task-specialized alignment + frontier open weights*

Models: Small 3 family, Medium 3, Magistral, Devstral, Mistral 3 (Large 3 + Ministral 3).

- **Small 3 / 3.1 / 3.2**:
    - Open-weights, strongly instruction-aligned models with top performance in their size class.
    - Alignment is capability-focused; safety is mostly external.
- **Medium 3**:
    - Frontier multimodal model whose alignment pipeline is not public but clearly tuned for enterprise reliability.
    - AI governance treats it as a managed GPAIM with obligations for evaluation/red-teaming. (Mistral AI Legal Center)
- **Magistral**:
    - Uses **reinforcement learning** on top of Medium 3 to optimize reasoning performance without explicit CoT labels.
    - Shows that Mistral is comfortable shaping behaviour with RL beyond SFT/DPO.
- **Devstral**:
    - Agentic coding model optimized for SWE-Bench, likely with dataset and possibly RL tuned for tool-use; safety depends on tooling/guardrails. (Mistral AI)
- **Mistral 3 family**:
    - Full spectrum from 3B to 675B MoE (Large 3), all Apache-2.0 and multimodal.
    - Strong performance on open leaderboards, but as of now no transparent, detailed alignment technical report. (Mistral AI)

Values alignment implications

- **Capability alignment** is very strong:

    - High reasoning quality (Magistral, Large 3).
    - Strong coding/agentic behaviours (Devstral).
    - Good multimodal reasoning (Pixtral, Large 3).
- **Safety/values alignment**:
    - Handled primarily via:
        - Instruction tuning and preference optimization (inherited from earlier generations).
        - **Modular moderation** (Mistral Moderation).
        - Governance processes and external red-teaming (e.g., included in AgentHarm, reliability studies). (ICLR Proceedings)
    - Open-weights nature of Mistral 3 means **hard value constraints are intentionally loose**, leaving control to deployers.

---

9. Gaps in the literature & what's *not* known

For a truly "exhaustive" alignment history, we'd ideally have:

- Full **training recipes** (datasets, SFT objective, preference data, RL rewards).
- Detailed **safety evaluation reports** for each family (toxicity, bias, jailbreaking, prompt-injection robustness).
- Per-model **alignment change logs** (e.g., how 7B Instruct v0.1 vs v0.2 vs v0.3 differ in refusal rates and bias).

As of now, the public record **does not** provide this for most proprietary models (Medium 1, Large 1, Medium 3, Large 3), and even for some open ones (Small 3, Nemo, Devstral). Much of what we know is:

- Direct from **Mixtral's** clear SFT + DPO story and bias benchmarks.

- Partial from **Large 2**'s discussion of hallucination reduction and alignment benchmarks. (Mistral AI)
- Specific to **math alignment** for Mathstral. (Mistral AI)
- Inferred from **AI governance documentation** and generic statements about evaluation/red-teaming. (Mistral AI Legal Center)

So any claim that, say, *"Medium 3 uses RLHF with human labels"* or *"Large 3 uses RLAIF"* would be speculation; the docs don't say.

---

10. Go deeper

- Build a **timeline table** just for alignment techniques (SFT vs DPO vs RL, benchmarks used) per model family

---

Timeline of alignment techniques & benchmarks by model family

Legend

- **SFT** – supervised fine-tuning on instructions/chat data
- **DPO** – Direct Preference Optimization (preference-based alignment)
- **RL** – reinforcement learning (task-oriented rather than "safety RLHF" unless noted)
- **RM@inference** – reward-model-based reranking or majority vote
- **Mod. model** – separate moderation model (Mistral Moderation) in front of LLM
- **Cap.** benchmarks – capability (MMLU, GSM8K, HumanEval, etc.)
- **Safety/bias** benchmarks – BBQ, BOLD, red-teaming, etc.

---

1. 2023 – First wave: Mistral 7B & Mixtral

Approx. release	Model family	Alignment techniques (training-time)	Alignment at inference / modular	Benchmarks emphasized	Notes on evidence
2023-09	**Mistral 7B (base)**	No explicit "alignment" beyond pretraining; SFT only applies to Instruct variants.	None; raw base model.	Cap.: MMLU, GSM8K, MATH, HumanEval, MBPP, ARC, HellaSwag, WinoGrande, PIQA, SIQA, OBQA, BoolQ, QuAC.	Paper focuses on dense architecture & performance; no RL/DPO/safety pipeline described.
2023-09	**Mistral-7B-Instruct v0.1–v0.3**	**SFT** on public conversation/instruction datasets; no DPO/RLHF described in model cards or paper.	**No built-in moderation**; blog explicitly warns it "does not have any moderation mechanism."	MT-Bench & LMSys Arena used informally by community; not central in official 7B paper.	Classic "helpful but not safe by default" open instruct model. Values ≈ SFT datasets + prompt template.
2023-12	**Mixtral 8x7B (base)**	Pretraining only (dense MoE); **no alignment stage** on base.	None on base; but bias is **measured**.	Cap.: MMLU, GSM8K, HumanEval, MBPP, etc. Safety/bias: **BBQ, BOLD** to evaluate social bias and sentiment.	Paper compares bias vs Llama-2-70B; Mixtral base is *less biased* on BBQ/BOLD but still just a pretrained model.

2023-12	**Mixtral-8x7B-Instruct**	**SFT + DPO** (explicitly documented): instruction SFT followed by **Direct Preference Optimization** on pairwise preference data.	No dedicated moderation baked into weights; intended as an open chat model to be combined with external guardrails.	Cap. & alignment: **MT-Bench**, LMSys **Arena Elo**; compared against GPT-3.5, Gemini Pro, Claude 2.1, Llama-2-70B Chat. Bias: inherits lower-bias base; no extra safety benchmark stage described.	This is the **first Mistral model with a clearly documented preference-alignment pipeline** (SFT → DPO).

---

2. 2023–early 2024 – Proprietary v1 models

Release	Model family	Alignment techniques (training-time)	Alignment at inference / modular	Benchmarks emphasized	Notes
2023-12	**Mistral Medium 1**	**Not publicly detailed.** Likely SFT + some preference alignment for chat, but no paper.	Served via Mistral platform → expected to be combined with moderation and governance processes.	Cap.: internal suite similar to 7B/Mixtral (MMLU, code, etc.), but no full table published.	AI governance lists it as a **GPAIM**; alignment pipeline is high-level only.

2024-02	**Mistral Large 1**	Not detailed; clearly instruction-tuned, but no SFT/DPO/RLHF breakdown published.	Deployed via Azure/other platforms with their own moderation and Mistral's governance obligations.	Cap.: MMLU, code, multilingual benchmarks; no explicit safety/bias suite in public docs.	Early frontier proprietary model; "alignment" is described in marketing terms (helpful, safe) but not technically documented.
2024-02	**Mistral Small 1**	Similar story: instruction-tuned proprietary model; no explicit SFT/DPO/RL breakdown.	Served behind moderation/guardrails on la Plateforme and partners.	Cap.: positioned as cost-efficient general-purpose model; no published safety benchmarks.	Governance docs treat it as a GPAIM with evaluation/monitoring requirements but few technical details.

---

3. 2024 – Open & specialist models + Large 2

Release	Model family	Alignment techniques (training-time)	Alignment at inference / modular	Benchmarks emphasized	Notes
2024-07	**Mistral Nemo 12B (base/instruct)**	Pretraining + **Instruct SFT** (explicit). Preference optimization likely, but **not confirmed** in public docs.	When served via NVIDIA NIM or Mistral, typically paired with platform-level safety/moderation; open weights themselves are not heavily constrained.	Cap.: MMLU, MT-Bench, HumanEval, GSM8K, multilingual tasks; marketed as best open-source multilingual model at release.	Nemo is mainly about **capability alignment** (chat + code + multilingual) with Apache-2.0 weights.

2024-07	**Mathstral 7B**	SFT on math/STEM data for instruction-following in math; details of dataset composition are partial.	**RM@inference**: they use a strong math reward model to select among many candidate completions (majority vote / best-of-N), boosting MATH accuracy from 56.6% to ~74–75%.	Math: **MATH**, GSM8K; also MMLU (math-heavy subset); show comparisons vs Mistral 7B & Mixtral.	This is a clear case of **task-specific alignment for truthfulness** via reward-model reranking, not general moral/safety alignment.
2024 (mid)	**Mistral Small 2 (open)**	Open-weight small model; clearly instruction-tuned, but docs don't spell out DPO/RLHF.	Open weights → usually combined with external moderation; governance docs classify it as open-source GPAIM.	Cap.: MMLU, GSM8K, HumanEval vs other small open models.	Conceptually "Mixtral-style alignment in a dense small model," but actual methods are undisclosed.
2024-10	**Ministral 3B / 8B (base & instruct)**	Base: pretraining only. Instruct: **SFT** on instruction data; no explicit DPO/RLHF statement.	Open weights; intended for local/edge. Safety left to deployers + moderation stacks.	Cap.: MMLU, code benchmarks (HumanEval/MBPP), multilingual; they show side-by-side charts vs Llama/Gemma small models for both base and instruct variants.	Focus is on **size–performance**; alignment is simple SFT for obeying instructions.
2024-06/07	**Pixtral 12B**	Multimodal instruction tuning on text+image data; exact SFT recipe not published.	Inherits Nemo's textual alignment; plus multimodal instruction. Moderation handled by platform/tools.	VL: doc QA, chart QA, image understanding; plus standard text benchmarks.	Values alignment beyond "don't be wildly unsafe" is not detailed; safety for images is presumably enforced via moderation.

2024-07	**Mistral Large 2 (“Large Enough”)**	Proprietary; they describe **“fine-tuning to reduce hallucinations and improve calibration”** but **do not name DPO/RLHF**.	Combined with **Mistral Moderation** or partner safety stacks in production.	Cap.: MMLU, GSM8K, MATH, code (HumanEval, MBPP, MultiPL-E). Alignment: **MT-Bench, WildBench, Arena Hard**; they explicitly optimize for concise responses rather than long ones.	Alignment goals: **truthfulness & calibration** (reduce hallucination, say “I don’t know”), and **conciseness**. Techniques are deliberately high-level, not recipe-level.
2024-11	**Pixtral Large 124B** (built on Large 2)	Inherits Large 2’s text alignment; adds multimodal SFT for doc/image reasoning; details undisclosed.	Paired with moderation and governance as an open-weights frontier MLLM.	VL benchmarks: document QA, charts, natural images; CountQA includes Pixtral Large 2411 as one of the top open MLLMs.	Same pattern: strong capability alignment, safety managed modularly.
2024-11	**Mistral Moderation**	Separate classifier/LLM for content moderation; training likely uses SFT on labeled harmful/non-harmful data; details not public.	Used as **Mod. model** in front of other LLMs to filter/judge prompts/responses.	Internally evaluated on toxicity, harassment, sexual content, etc.; no full benchmark list published.	Marks a shift to **modular safety**: base models + dedicated moderation.

---

4. 2025 – Small 3, Medium 3, Magistral, Devstral, Mistral 3

Release	Model family	Alignment techniques (training-time)	Alignment at inference / modular	Benchmarks emphasized	Notes
2025-01 / 03 / 07	**Mistral Small 3 (24B), 3.1, 3.2**	Open-weight 24B dense model; clearly **instruction-tuned**; method (SFT vs SFT+DPO) **not disclosed**.	Designed for local deployment; safety expected via Mistral Moderation or third-party guardrails.	Cap.: MMLU, GPQA, HumanEval, GSM8K, MATH vs Llama 3.3 70B, Claude 3.5 Haiku, GPT-4o-mini; also multimodal benchmarks for later 3.x releases.	“Best in class” small open model; value alignment == helpfulness & reliability, not strict normative constraints.
2025-05	**Mistral Medium 3**	Frontier multimodal model; docs say it’s trained with an extensive data mix and tuned for instruction-following & reasoning, but **don’t show the exact SFT/DPO/RL recipe**.	Deployed as a “frontier GPAIM”; **Magistral** is then trained on top via RL; moderation via Mistral Moderation for safety.	Cap.: MMLU, GPQA, GSM8K, MATH, code (HumanEval/MBPP), multimodal tasks; often compared to Claude Sonnet, GPT-4.1-mini etc.	Medium 3 is the **backbone** for later RL-aligned reasoning (Magistral); its own alignment is “strong but opaque.”

2025-05	**Magistral (Small/Medium)**	Explicit **RL** on top of a Medium-3-like base: the paper describes a reinforcement-learning pipeline using task success as a reward, encouraging good reasoning and discouraging unnecessary verbosity; CoT emerges without direct supervision.	In deployment, expected to be combined with moderation; RL reward can indirectly penalize some unsafe behaviours but is primarily a **capability** reward (task success).	Reasoning benchmarks: MATH, GSM8K, AIME, GPQA, competition puzzles, tool-use and code reasoning tasks; they show significant gains vs base Medium 3.	This is the clearest case of **post-training RL** in Mistral's stack, but aimed at *reasoning quality* rather than safety/bias per se.
2025-05	**Devstral**	Training includes **SFT on multi-file agent traces** plus additional tuning for tool use; public docs do *not* confirm RL, but emphasise iterative debugging and agentic workflows.	Typically used with constrained workspaces & external moderation, given its power to edit codebases.	Code/agent benchmarks: **SWE-Bench Verified** (Devstral is #1 open model); other coding tasks like HumanEval & multi-file evals.	Devstral is about **agentic capability alignment** (tool calls, multi-step edits), not moral/safety alignment.

2025-12	**Mistral 3 – Large 3 (MoE 675B/41B active)**	Training: massive MoE pretraining; docs emphasise performance & efficiency; **alignment recipe (SFT/DPO/RL) is not yet public**.	Open weights under Apache-2.0; for safety, Mistral and partners recommend using **Mistral Moderation** or other guardrails.	Cap.: Top-tier scores on MMLU, GPQA, GSM8K, MATH, HumanEval, and strong multimodal benchmarks; near the top of open-source leaderboards (LMSys, etc.).	Large 3 is a **frontier open model**: very strong capability alignment, but values/safety alignment is largely left to deployers + moderation systems.
2025-12	**Mistral 3 – Ministral 3 (3B/8B/14B; base/instruct/reasoning)**	Base: pretraining. Instruct: SFT (and possibly preference tuning) on text+image instructions; Reasoning variants likely borrow ideas from Magistral (RL-style task tuning), though this is not fully documented yet.	Open-weight, multimodal; recommended to be run with moderation front-ends in production.	Benchmarks mirror Small 3 + multimodal (MMLU, GPQA, HumanEval, GSM8K, MATH, VLQA).	Continuation of the “open but powerful” theme: strong instruction/reasoning, safety externalized.

5-line takeaway

If you zoom out over this table, the **alignment story over time** is:

1. **2013–2023:** Simple **SFT-only** instruction models (Mistral-7B-Instruct) with *no moderation* and almost no explicit safety metrics.

2. **Late 2023:** Mixtral-8x7B-Instruct brings **SFT + DPO** and first-class **bias evaluation (BBQ/BOLD)**, but still no standalone safety model.

3. **2024:** Alignment becomes **modular**:

    - Specialist alignment (Mathstral's **RM@inference** for math truthfulness).
    - Frontier models like Large 2 tuned for **truthfulness & calibration** with alignment benchmarks.
    - A distinct **Mistral Moderation** model appears as a safety layer.
4. **2025:** Mistral embraces **RL-style post-training** for reasoning (Magistral) and heavily **agentic alignment** (Devstral), while the flagship Mistral 3 family pushes open-source capabilities and relies on **external moderation** for values/safety.

If you'd like, next step could be: **pick one model family** (e.g., Mixtral, Large 2, or Magistral) and I can expand the tiny row above into a step-by-step description of *exactly* how its alignment training works and what each benchmark tells you about values vs. capability.