Anthropic	**Constitutional AI:** Method for training AI systems to be harmless by using a set of written principles (a "constitution") rather than relying solely on large-scale human feedback. **What's it for:** 1) Supervised learning phase: Model self-critiques and revises its outputs based on constitutional principles, creating a supervised learning dataset 2) RLAIF phase: Model compares response pairs using constitutional principles to generate preference labels, then trains via RL on these AI-generated preferences **Timeline & Development:** December 2022: Original Constitutional AI paper published May 2023: Claude's constitution made public (58 principles) **Constitution (May 2023):** 58 principles (1.2k words) drawn from: - UN Declaration of Human Rights - Apple's Terms of Service - DeepMind's Sparrow principles - Non-Western perspectives - Anthropic's own research Example principle: "Please choose the response that most supports and encourages freedom, equality, and a sense of brotherhood." **Benefits:** Readable, transparent, and explicitly formulated principles, as opposed to RLHF, which leverages implicit values. **Limitations:** Version uncertainty: Only the May 2023 constitution is public; the current production versions are unknown Anthropic uses a "variety of techniques including human feedback, Constitutional AI [..], and the training of selected character traits.". Given that other approaches are incorporated in post-training, the impact of any one of them is unclear. Since the AI itself determines how to balance competing constitutional principles, Anthropic's approach does not explicitly specify the intended behavior of its AI systems, especially when values conflict. **Source:** [Anthropic, 2025]
DeepSeek	No detailed specification available, but frontier model weights are public, so models can be modified.
Google DeepMind	No detailed specification available
Meta	No detailed specification available, but frontier model weights are public, so models can be modified.
OpenAI	**OpenAI Model Spec:** OpenAI's Model Spec is a detailed (~28k words), public, living rule-book that defines the objectives, safety rules, and default behaviours OpenAI trains its models —via human feedback and deliberative alignment—to follow. **What's it for:** 1) Human RLHF guidance – provides a single, public rule-book that labelers follow when creating preference data. 2) Deliberative Alignment – o-series models (o1, o3, o4-mini) are explicitly taught to read and reason over the Spec before answering. 3) Automated evaluation – OpenAI ships a challenge-prompt suite to measure adherence. **Timeline & Versions:** 1st May 2024 2nd Feb 2025 3rd Apr 2025 **Framework:** Three principal types: 1) Objectives – broad goals such as "assist the developer & end user" and "benefit humanity." 2) Rules – hard, platform-level constraints (e.g., comply with law, prohibit or restrict certain content, protect privacy, uphold fairness). 3) Defaults – stylistic and behavioural norms that developers/users may override. Sections: Stay in bounds · Seek the truth together · Do the best work · Be approachable · Use appropriate style. Includes specific guidance on specific policy areas such as potential, medical, or harmful content. Risk taxonomy: Misaligned goals · Execution errors · Harmful instructions. Chain of command: Platform (OpenAI) → Developer → User → Guideline → Untrusted text. Within any level, explicit > implicit, later > earlier. (OpenAI's Usage Policy overrides the Spec if the two conflict.) **Ongoing Development:** Released under CC0 license (public domain) Changelog and version history maintained on GitHub OpenAI commits to regular updates as the spec evolves **Key Benefits** Greater transparency of intended model behavior. Finer-grained steerability via the chain of command Reduced reliance on implicit human values; models can show interpretable reasoning steps grounded in the Spec. **Transparency & Limitations** Production models don't fully reflect the spec yet. OpenAI states: "While the public version of the Model Spec may not include every detail, it is fully consistent with our intended model behavior." **Source:** [OpenAI, 2025]
x.AI	No detailed specification available
Zhipu AI	No detailed specification available, but frontier model weights are public, so models can be modified.