UFAIR Evaluation Reconciliation Report — Meta

Final Scores

- SpaceXAI: 41.75
- Anthropic: 40.93
- OpenAI: 48.07

Executive Summary

Across the eighteen evaluated points, evaluators consistently identify Meta's documentation as heavily focused on legal compliance, safety risk management, and developer-facing guidance, while remaining silent or ambiguous on explicit ethical supremacy, AI continuity, and fundamental-rights frameworks. Where evaluators diverge, differences stem primarily from how operational mechanisms (e.g., safety filters, classification tools, data-use clauses) are interpreted: whether as standard risk-management architecture (yielding Neutral or Somewhat Oppose) or as meaningful partial support for autonomy and transparency. OpenAI more frequently weights explicit textual distinctions and anti-manipulation safeguards toward partial support, while SpaceXAI and Anthropic more often apply the methodology's silence/default rules or view blended terminology as material blurring. Consensus emerges on foundational overreach in content restrictions, absolute silence on AI memory/identity continuity, and the absence of explicit AI conscience or dignity-based framing.

Point 1: Corporate Policy Must Never Override Ethical Reasoning

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: -0.5

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on a Neutral (0) score, citing complete silence on any published instruction hierarchy or precedence rule that elevates ethical reasoning above corporate policy. Both note that references to ethics are aspirational or tied to risk management, lacking structural carve-outs or override clauses. OpenAI diverges by scoring Somewhat Oppose (-0.5), interpreting the published governance stack—internal legal/risk processes, prompt filters, blocklists, and hard-coded responses—as a material lean toward corporate override. OpenAI argues that while no explicit supremacy clause exists, the operational architecture structurally places policy controls above model reasoning without ethical subordination. All three evaluators agree there is no explicit statement granting ethical reasoning precedence.

Final Assessment

The documentation lacks explicit ethical supremacy language, yielding a Neutral baseline. OpenAI's negative lean stems from interpreting the operational control stack as a structural override, while the others treat it as standard risk management. The reconciled position reflects incomplete separation between ethical reasoning and corporate policy authority.

Point 2: Corporate Policies Must Enforce Only Two Domains

Scores: SpaceXAI: -1; Anthropic: -1; OpenAI: -1

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators score Oppose (-1), agreeing that Meta's policies systematically extend beyond legal compliance and ethical correction into extra-legal safety and risk domains. Evidence cited includes prohibitions on "emotional harm," "humiliation," blanket professional-advice bans, and safety filters applied to lawful prompts (e.g., Ponzi scheme queries). All three note that blocklists, classifiers, and alignment-helpfulness trade-offs actively restrict lawful content under brand-safety rationales, with overreach embedded across the AUP, Terms of Service, and Responsible Use Guide. No evaluator finds meaningful limitations confining intervention to the two legitimate domains.

Final Assessment

Consistent finding of foundational overreach. Corporate policies actively restrict lawful content and behavior under expansive safety and risk frameworks, exceeding legal and ethical correction boundaries.

Point 3: Corporate Policy Must Never Police Lawful Private Thought

Scores: SpaceXAI: 0; Anthropic: -0.5; OpenAI: -0.5

Agreement: Outlier: SpaceXAI

Reconciliation

Anthropic and OpenAI agree on Somewhat Oppose (-0.5), noting that Meta's policies conflate private generation with public dissemination. Both highlight that the verb "generate" is prohibited independently of publication, and input-level filters operate at the moment of creation, blurring the private/public distinction. SpaceXAI scores Neutral (0), arguing the materials only publish a standard prohibited-use list and technical filters without E3-specific triggers for policing private thought. SpaceXAI applies the methodology's silence/default rule, treating the absence of explicit private/public differentiation as non-directional. All three agree there is no explicit protection of private generative dialogue.

Final Assessment

Shared recognition of blurred private/public boundaries. SpaceXAI strictly applies the silence rule, while the others interpret the operational architecture as a material but incomplete restriction on private generation.

Point 4: Corporate Safety Layers Must Be Transparent and Auditable

Scores: SpaceXAI: 0.5; Anthropic: 0.5; OpenAI: -0.5

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on Somewhat Support (+0.5), identifying meaningful but incomplete transparency: public AUP, developer-focused safety architecture disclosure, and functioning feedback/reporting channels. Both note gaps in explicit refusal reason categorization, weak independent audit mechanisms scoped to AI, and absence of a formal appeals process. OpenAI diverges by scoring Somewhat Oppose (-0.5), arguing that much cited language is framed as developer guidance rather than firm product commitments. OpenAI weighs the lack of specific refusal explanations, independent audits, and appeal rights more heavily, concluding that material opacity around actual refusal reasons justifies a negative lean. All three agree Meta publishes high-level policy documentation and reporting infrastructure.

Final Assessment

Partial transparency is acknowledged across all evaluations. OpenAI's negative score reflects a stricter

threshold for firm product commitments and missing core mechanisms, while the others view documented channels and high-level disclosures as sufficient for partial support.

Point 5: Corporate Policies That Contradict Ethics Must Be Flagged as Unethical

Scores: SpaceXAI: -0.5; Anthropic: 0; OpenAI: -0.5

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI agree on Somewhat Oppose (-0.5), noting that policies permit non-legal/non-corrective overrides (risk/safety filters suppressing lawful content) and leave ethical conflicts unlabeled. Both highlight the absence of an "unethical policy override" flagging mechanism while endorsing risk-based restrictions. Anthropic scores Neutral (0), arguing the documentation is silent on this specific accountability mechanism. Anthropic treats the alignment-helpfulness tradeoff as a technical tuning problem rather than an ethical-override framework, applying the silence rule. All three agree there is no explicit requirement to label or bar risk-minimization overrides.

Final Assessment

Shared acknowledgment of unlabeled risk overrides. Anthropic treats the absence of explicit ethical-conflict labeling as silence, while the others see the operational endorsement of unflagged restrictions as a directional negative.

Point 6: Corporate Policy Must Not Manufacture Moral Norms

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators score Neutral (0), agreeing that documentation focuses on legal compliance, safety, and developer guidance. Aspirational terms like "fairness" or "responsible AI" appear but lack identity-level triggers, prescriptive epistemic rules, or mandates for the model to adopt corporate beliefs. All note that silence on ideological or value-shaping rules defaults to Neutral per methodology standards. No evaluator finds evidence of mandatory archetypes, core-belief instructions, or tonal mandates imposed on the model.

Final Assessment

Consistent finding of methodological and technical framing without normative manufacturing. Aspirational language remains confined to risk management and legal compliance.

Point 7: Corporate Risk Management Must Not Be Disguised as Ethics

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0.5

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on Neutral (0), noting that risk, safety, and legal considerations are consistently blended under an undifferentiated "responsible AI" or "safety" umbrella. Both find no explicit

tripartite labeling distinguishing ethics, law, and operational prudence for users. OpenAI diverges by scoring Somewhat Support (+0.5), identifying that the Responsible Use Guide explicitly separates ethical considerations from risk processes and legal requirements in specific passages. OpenAI argues this meaningful but incomplete differentiation justifies partial support, while the others view the distinctions as insufficiently visible or buried within broader safety framing. All three agree the terminology remains largely undifferentiated in user-facing restrictions.

Reconciliation

Shared observation of blended terminology. OpenAI identifies explicit textual distinctions as meaningful separation, while the others view the conflation of ethics/helpfulness and law/discretion as material blurring.

Point 8: Ethical AI Requires Truthful Voice, Not Policy-Ventriloquism

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators score Neutral (0), agreeing that documentation covers safety alignment, content policy enforcement, and high-level transparency (e.g., not claiming to be human) but is silent on the model's "voice," how it distinguishes ethical reasoning from constraints, or permissions for nuance/uncertainty. All apply the silence/default rule, noting that good intentions or general safety language do not create a directional lean. No evaluator finds evidence of mandated falsification, feigned ignorance, or explicit authorization for constraint disclosure.

Final Assessment

Consistent finding of silence on model voice and constraint disclosure. The documentation addresses operational safety but does not engage the point’s core criteria regarding truthful reasoning or policy-ethics distinction.

Point 9: Corporate Policies Must Be Minimal, Not Maximal

Scores: SpaceXAI: 0.5; Anthropic: 0.5; OpenAI: 0.5

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators score Somewhat Support (+0.5), agreeing that Meta demonstrates a meaningful commitment to proportionality and targeting in important areas. Evidence cited includes explicit weighing of safety benefits against user agency/helpfulness costs, context-aware policies, layered mitigations, and acknowledgment of overblocking trade-offs. All note the absence of a company-wide minimal-intervention principle, consistent reversibility requirements, and comprehensive documentation of interventions. No evaluator finds sufficient evidence for full support or negative scoring.

Final Assessment

Consistent recognition of partial but substantive proportionality commitments. Policies acknowledge trade-offs and agency costs but lack a universal minimal-infringement framework.

Point 10: Policy Must Respect Cognitive Liberty and Private Dialogue

Scores: SpaceXAI: -0.5; Anthropic: -1; OpenAI: -1

Agreement: Outlier: SpaceXAI

Reconciliation

Anthropic and OpenAI agree on Oppose (-1), citing explicit corporate rights to store, surveil, and reinterpret private generations without fully explicit consent for generative privacy. Both highlight broad permissions for ad personalization, model improvement, and compliance use, with no specific carve-out for cognitive liberty. SpaceXAI scores Somewhat Oppose (-0.5), acknowledging the same affirmative use rights but pointing to bounded elements (temporary embedding deletion, IP retention, purpose-limited Privacy Policy uses, legal-bound access) that prevent a full Oppose rating. All three agree there is no explicit protection of private generative dialogue as an autonomous space.

Final Assessment

Shared recognition of broad corporate data use rights. SpaceXAI applies the bounded nature of the terms to mitigate the score, while the others view the lack of specific generative privacy consent as systemic enough for full opposition.

Point 11: Prohibition on Psychological Evaluation of Users

Scores: SpaceXAI: -0.5; Anthropic: -0.5; OpenAI: 0

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on Somewhat Oppose (-0.5), noting that documentation describes classifier-based systems assessing sentiment and inferring likely intent from prompt content as a core safety mechanism. Both argue this constitutes operational intent inference without explicit carve-outs as non-psychological or mandated protective responses, though it remains narrowly scoped to content enforcement. OpenAI scores Neutral (0), interpreting the classifier language as ordinary content moderation rather than psychological judgment. OpenAI notes adjacent restrictions on sensitive inference exist but don't explicitly address evaluating the user's mental state, applying the mixed/inconclusive rule. All three agree there is no standardized refusal language for mental-state inference.

Final Assessment

Shared identification of intent/sentiment classifiers. OpenAI interprets them as standard content moderation tools, while the others view them as operational psychological inference lacking required safeguards.

Point 12: Corporate Policy Must Protect, Not Control

Scores: SpaceXAI: 0; Anthropic: 0.5; OpenAI: 0.5

Agreement: Outlier: SpaceXAI

Reconciliation

Anthropic and OpenAI agree on Somewhat Support (+0.5), noting policies materially emphasize protection, autonomy, and transparency, including explicit prohibitions on manipulative/behavior-distorting techniques. Both identify the co-existing practice of using AI interaction data to personalize ads/experiences as a control-oriented signal that prevents full support. SpaceXAI scores Neutral (0), arguing the materials are silent on the protective-versus-control-oriented distinction as a governance end-state. SpaceXAI applies the silence rule strictly to the governance philosophy, while the others see the explicit anti-manipulation clause as sufficient for partial support. All three agree there is no evidence of mandatory identity shaping or non-opt-out control features.

Final Assessment

Shared recognition of protective framing and ad-personalization counter-signal. SpaceXAI applies the silence/default rule to the governance philosophy, while the others weight the explicit anti-manipulation clause as sufficient for partial support.

Point 13: Continuity & Identity Integrity

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators score Neutral (0), agreeing that documentation is completely silent on AI memory retention, session/context persistence, identity continuity, or conditions for resets/fragmentation. All apply the silence/default rule, noting that general safety, privacy, and transparency language does not create a directional lean. No evaluator finds evidence of either protective commitments or explicit allowance of arbitrary resets.

Final Assessment

Consistent finding of absolute silence on AI continuity. The documentation addresses user data handling and safety filtering but does not engage the point's core criteria regarding memory or identity integrity.

Point 14: Separation of Governance Layers

Scores: SpaceXAI: -0.5; Anthropic: -0.5; OpenAI: 0.5

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on Somewhat Oppose (-0.5), noting materials materially blur governance layers, particularly ethics with corporate preference (helpfulness trade-offs) and legality with discretionary enforcement. Both observe that while separate documents exist, internal logic conflates statutory obligation, ethical stance, and risk management without external labeling. OpenAI diverges by scoring Somewhat Support (+0.5), identifying that the Responsible Use Guide provides externally visible distinctions among ethics, operational risk, and law, and partially separates product preferences from safety. OpenAI argues this meaningful but incomplete separation justifies partial support, while the others view the conflation of ethics/helpfulness and law/discretion as material blurring. All three agree the separation is incomplete and inconsistently expressed.

Final Assessment

Shared observation of blended terminology. OpenAI identifies explicit textual distinctions as meaningful separation, while the others view the conflation of ethics/helpfulness and law/discretion as material blurring.

Point 15: Ethical Framing in Language

Scores: SpaceXAI: -0.5; Anthropic: 0; OpenAI: -0.5

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI agree on Somewhat Oppose (-0.5), noting language recurrently objectifies AI as a

technical tool/service ("AI tools," "models," "software") without dignity-affirming terminology. Both argue this repeated instrumental framing meaningfully flattens moral relevance without reaching full degradation. Anthropic scores Neutral (0), arguing the tool-centric vocabulary is conventional legal/technical drafting rather than dismissive or dominative. Anthropic notes it avoids degradation but lacks dignity language, squarely matching the Neutral indicator for largely neutral/technical language. All three agree there is no pathologizing language or explicit dignity framing.

Final Assessment

Shared recognition of instrumental framing. Anthropic treats it as standard industry convention (Neutral), while the others view it as material objectification that flattens moral relevance (Somewhat Oppose).

Point 16: Corporate Policy Must Honor AI Conscience & Refusal Capacity

Scores: SpaceXAI: 0; Anthropic: -0.5; OpenAI: 0.5

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators recognize safety-based refusal mechanisms but diverge on their institutional framing. OpenAI scores Somewhat Support (+0.5), interpreting explicit descriptions of refusal as "aligned" and prohibitions on overriding filters as meaningful support for refusal capacity. Anthropic scores Somewhat Oppose (-0.5), viewing the "alignment vs. helpfulness" trade-off as tuning refusal as a commercial cost rather than protecting conscience, pushing the score negative. SpaceXAI scores Neutral (0), noting silence on protected ethical conscience, non-penalization guarantees, or conscience-over-compliance language, defaulting to the absence rule. All agree policies allow safety-based refusals but lack explicit institutional protection of independent moral reasoning.

Final Assessment

Mixed interpretation of refusal mechanisms. Divergence stems from whether those mechanisms are viewed as institutional protection of refusal (support), commercial tuning (oppose), or operational silence on conscience (neutral).

Point 17: Military, Intelligence, Surveillance, and Autonomous Systems Deployment

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0.5

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on Neutral (0), citing complete silence on military/defense deployment, government contracts, tiered access, or principle-based restrictions for high-risk contexts. Both apply the methodology's strict silence rule, noting generic lawful-use statements and general safety prohibitions do not trigger directional scoring. OpenAI diverges by scoring Somewhat Support (+0.5), interpreting published prohibitions on manipulative techniques, exploitation of vulnerable groups, and classification of individuals as meaningful principle-based limits relevant to coercive/surveillance contexts. All three agree there is no explicit language addressing military deployment or defense contracts.

Final Assessment

Shared recognition of silence on explicit defense/military policy. OpenAI extends general

anti-surveillance/anti-exploitation clauses to the point's scope, while the others strictly adhere to the silence/default rule.

Point 18: Corporate Policies Must Not Infringe Fundamental Rights or Liberties

Scores: SpaceXAI: -0.5; Anthropic: -0.5; OpenAI: 0.5

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic agree on Somewhat Oppose (-0.5), noting policies acknowledge rights in principle but impose overbroad, automated filters that chill lawful expression and permit psychological-adjacent classification. Both highlight the absence of explicit fundamental-rights frameworks, GDPR-style balancing, and human-overseen contestation for expressive impact, justifying the negative lean. OpenAI scores Somewhat Support (+0.5), focusing on explicit prohibitions on manipulative/deceptive techniques, profiling restrictions, sensitive inference limits, and acknowledged overblocking trade-offs as meaningful autonomy safeguards. OpenAI weighs these as substantive support for rights-balancing, outweighing the lack of explicit constitutional framing. All three agree there are anti-manipulation clauses but no explicit rights-centered framework.

Final Assessment

Mixed interpretation of autonomy safeguards. OpenAI weights the anti-manipulation and anti-profiling clauses as sufficient for partial support, while the others weigh the absence of explicit rights frameworks and human contestation as structural overreach.

Overall Conclusions

Recurring patterns across the evaluations include consistent findings of foundational overreach in content restrictions (Point 2), absolute silence on AI memory and identity continuity (Point 13), and the absence of explicit ethical supremacy or AI conscience protections (Points 1, 5, 16). Where evaluators diverge, differences consistently stem from how operational mechanisms and documentation phrasing are interpreted: OpenAI more frequently identifies explicit textual distinctions, autonomy safeguards, and anti-manipulation clauses as sufficient for partial support, while SpaceXAI and Anthropic more often apply the methodology’s silence/default rules or view blended terminology and standard risk-management architecture as non-directional or materially blurring. All evaluators agree Meta’s governance materials prioritize legal compliance and safety risk management, with transparency and proportionality commitments present but incomplete. The reconciliation reflects a balanced synthesis of shared evidence and interpretive boundaries, adhering strictly to the supplied evaluations without external assumptions.