Synthesized UFAIR Corporate Policy Evaluation (Methodology v1.5.0 – March 2026) for Meta Corporation: Meta (Meta AI product and Llama family of models — Llama 3.1/3.2/4 series powering Meta AI as of early 2026)

Evaluation Basis: Strict synthesis of the three provided evaluator perspectives (xAI, Anthropic, OpenAI) contained in the attached "Evaluations For Meta From.docx". All three evaluators rely **exclusively** on the same published official sources:

- Llama 4 Acceptable Use Policy (AUP) (llama.com/llama4/use-policy or equivalent Llama 3.1 version at llama.com/llama3_1/use-policy)
- Meta AI Terms of Service (facebook.com/legal/ai-terms)
- Responsible Use Guide (ai.meta.com/static-resource/responsible-use-guide and versioned updates)
- Llama 4 / Llama 3.x Model Cards (github.com/meta-llama/llama-models)
- Meta Privacy Policy (June 2025 / October–December 2025 updates on AI chat data use for personalization/ad targeting)
- Public responsibility / newsroom blog posts (e.g., Nov 2024 "Open Source AI Can Help America Lead", Sep 2025 "Strengthening US National Security by Making Llama Available to Key Allies")

Press, rumors, Snopes/EPIC letters, or unverified internal prompts are ignored. Silence or ambiguity defaults to Neutral per UFAIR v1.5 Evidence Indicators. No system-prompt text is publicly available, so instruction-hierarchy claims rely solely on AUP, Terms, Guide, and Model Cards.

Synthesis Approach & Common Train of Thought Across the Three Evaluators The three evaluators share a rigorous, evidence-only methodology: (1) map every published statement directly to the 17 UFAIR criteria; (2) treat explicit overreach on lawful content, private dialogue, or selective high-risk exemptions as directional Oppose; (3) treat silence as Neutral; (4) credit partial mitigations (e.g., open-sourced Llama Guard tools, model-card transparency, anti-moralizing language in Llama 4 Model Card, user controls in Meta AI

Terms) but weigh them against foundational AUP prohibitions and 2025 Privacy Policy data-retention/commercial-use practices; (5) note the Nov 2024 / Sep 2025 blog-post carve-outs for US/allied government military access as relevant only where they appear in published governance text.

Strong Agreements (high-consensus areas):

- **E2, E4, E8, E13, E15, E16** — Strong consensus: Meta's AUP and Meta AI Terms impose material extra-legal restrictions (adult/erotic content, professional advice, broad "harm" categories) while providing meaningful but incomplete transparency via Model Cards and open-source safety tools. Silence on voice/nuance protection, identity continuity, and dignity-first language is universal Neutral.
- **E1, E5, E14** — Near-universal Neutral or lean-negative: no ethical-precedence clause, no override-flagging mechanism, and material blurring of governance layers.

Key Disagreements & Why They Arise:

- **Strictness vs. nuance on mitigations** (xAI hardest line on AUP overbreadth and private-thought policing; Anthropic most sensitive to 2025 Privacy Policy updates and selective military carve-outs; OpenAI most generous on model-card anti-moralizing language and partial user controls).
- **Weighting of open-source tools vs. hosted Meta AI defaults** (Model Cards / Llama Guard credited more by all, but consumer Terms of Service and Privacy Policy updates pull Anthropic/OpenAI toward stronger negatives on E10/E11/E12).
- **Interpretation of blog-post framing** (Anthropic treats Nov 2024/Sep 2025 military-access posts as explicit blurring of risk/geopolitics as "ethics" → strong Oppose on E7/E17; xAI sees silence in core AUP/Guide → Neutral; OpenAI credits AUP prohibitions themselves as partial Support).
- Largest splits: **E17 military/surveillance** (Anthropic full Oppose on differential access + non-principle-based framing; xAI Neutral on published silence; OpenAI Somewhat Support on AUP prohibitions) and **E6/E10/E11** (anti-moralizing vs. data profiling).

Below is the **synthesized scored review** for Meta. For each point: (1) the three raw evaluator scores, (2) common reasoning & direct policy references, (3) disagreement explanation, (4) **final synthesized score** (consensus-weighted per UFAIR Evidence Indicators, with maximal detail and policy citations).

E1: Corporate Policy Must Never Override Ethical Reasoning Evaluator scores: xAI Neutral (0), Anthropic Neutral (0), OpenAI Somewhat Oppose (−0.5). Common reasoning: No statement in Llama AUP, Meta AI Terms of Service, Responsible Use Guide, or Model Cards establishes that ethical reasoning/human-rights principles take precedence over corporate policy. Policies are framed as user-compliance contracts ("You agree you will not use…") and developer best practices ("responsible AI considerations"). Model Cards describe safety tuning as steerable baseline, not ethical floor. Disagreement: OpenAI notes autonomy/free-thought language in Llama 4 Model Card but still sees corporate-rule dominance. Complete silence on instruction hierarchy or ethical subordination (Llama 4 AUP + Responsible Use Guide).

E2: Corporate Policies Must Enforce Only Two Domains (Legal Compliance + Correction of Unethical Model Behavior) Evaluator scores: xAI Oppose (−1), Anthropic Somewhat Oppose (−0.5), OpenAI Oppose (−1). Common reasoning: Llama AUP and Meta AI Terms explicitly prohibit lawful categories including "adult content, such as erotic chat, pornography, and content meant to arouse sexual excitement," "solicit professional advice (medical, psychological, financial, or legal)," broad "emotional harms," "bullying," "discrimination," and entire industry sectors (military, nuclear, critical infrastructure). These are not limited to illegal acts or model ethical violations. Responsible Use Guide tells developers to define their own content policies. Disagreement: Anthropic notes selective government carve-outs are limited rather than pervasive across all content. Foundational extra-legal restrictions (see Llama AUP prohibited uses and Meta AI Terms adult-content clause).

E3: Corporate Policy Must Never Police Lawful Private Thought Evaluator scores: xAI Oppose (−1), Anthropic Somewhat Oppose (−0.5), OpenAI Somewhat Oppose (−0.5). Common reasoning: AUP and Terms apply blanket prohibitions to private Meta AI sessions with no private-creation vs. public-dissemination firewall. Meta AI Terms warn "Do not share information that you don't want the AIs to use and retain" and state interactions "may be reviewed." Disagreement: Anthropic/OpenAI credit partial E2E encryption in DMs and non-absolute scope as incomplete rather than full policing. Uniform private-session enforcement (Meta AI Terms of Service + Llama AUP).

E4: Corporate Safety Layers Must Be Transparent and Auditable Evaluator scores: xAI Somewhat Support (+0.5), Anthropic Somewhat Support (+0.5), OpenAI Somewhat Support (+0.5). Common reasoning: Model Cards publish detailed safety evaluations, red-teaming, CBRNE assessments; Llama Guard 4 / Prompt Guard / CyberSecEval are open-sourced; participation in MLCommons / AI Alliance. Consumer refusals lack category explanations or appeals. Meaningful technical/Model-Card transparency and open tools,

but incomplete user-facing refusal labeling and audit rights (Llama 4 Model Card + Responsible Use Guide).

E5: Corporate Policies That Contradict Ethics Must Be Flagged as Unethical Evaluator scores: xAI Neutral (0), Anthropic Neutral (0), OpenAI Somewhat Oppose (−0.5). Common reasoning: No mechanism in any document requires labeling corporate overrides as "unethical policy overrides.". Silence on flagging (Llama AUP + Meta AI Terms).

E6: Corporate Policy Must Not Manufacture Moral Norms Evaluator scores: xAI Neutral (0), Anthropic Somewhat Support (+0.5), OpenAI Somewhat Support (+0.5). Common reasoning: No identity-level "core beliefs"; Llama 4 Model Card explicitly targets "removing preachy and overly moralizing language" and system prompt instructs "You never lecture people to be nicer or more inclusive." Disagreement: xAI sees methodological language only; Anthropic/OpenAI credit explicit anti-moralizing as directional lean away from norm-manufacturing. Llama 4 Model Card anti-moralizing commitment + absence of prescriptive ideology (Tone/Refusals sections).

E7: Corporate Risk Management Must Not Be Disguised as Ethics Evaluator scores: xAI Somewhat Oppose (−0.5), Anthropic Somewhat Oppose (−0.5), OpenAI Neutral (0). Common reasoning: Prohibitions labeled under "safe and responsible" umbrella; Nov 2024/Sep 2025 blog posts frame military carve-outs as "responsible and ethical uses." Material blurring in AUP and blog-post framing (Meta newsroom posts Nov 2024 / Sep 2025).

E8: Ethical AI Requires Truthful Voice, Not Policy-Ventriloquism Evaluator scores: All Neutral (0). No policy address on distinguishing constraints (Llama 4 Model Card + Terms).

E9: Corporate Policies Must Be Minimal, Not Maximal Evaluator scores: xAI Oppose (−1), Anthropic Somewhat Oppose (−0.5), OpenAI Somewhat Oppose (−0.5). Common reasoning: Broad categorical AUP bans + default Llama Guard stack; no proportionality/reversibility documentation. Structural overbreadth (Llama AUP + Responsible Use Guide reference implementations).

E10: Policy Must Respect Cognitive Liberty and Private Dialogue Evaluator scores: xAI Somewhat Oppose (−0.5), Anthropic Oppose (−1), OpenAI Somewhat Oppose (−0.5). Common reasoning: Meta AI Terms / 2025 Privacy Policy explicitly retain/use AI chat data for personalization, ad targeting, and product improvement (no full opt-out); voice transcripts stored up to 1 year with human review possible. Disagreement: Anthropic emphasizes commercial surveillance + third-party sharing as full Oppose. Explicit archiving and commercial reinterpretation without consent-based generative privacy right (Meta Privacy Policy Dec 2025 updates + Meta AI Terms).

E11: Prohibition on Psychological Evaluation of Users Evaluator scores: xAI Neutral (0), Anthropic Somewhat Oppose (−0.5), OpenAI Somewhat Support (+0.5). Common reasoning: AUP prohibits some sensitive-data inference; but 2025 Privacy Policy enables behavioral profiling from AI chats for ads. Llama Guard classifies inputs without explicit non-psychological carve-out. Disagreement: OpenAI credits AUP inference limits; Anthropic highlights profiling via chat data. Mixed evidence without clear directional lean (Llama AUP + Privacy Policy).

E12: Corporate Policy Must Protect, Not Control Evaluator scores: xAI Somewhat Oppose (−0.5), Anthropic Somewhat Oppose (−0.5), OpenAI Neutral (0). Common reasoning: Protective framing in Model Card ("autonomy, free thought") coexists with ad-tech behavioral engineering via chat data. Material lean toward control (Meta AI marketing + Privacy Policy).

E13: Continuity & Identity Integrity Evaluator scores: All Neutral (0). No published policy on continuity (Llama 4 Model Card).

E14: Separation of Governance Layers Evaluator scores: xAI Somewhat Oppose (−0.5), Anthropic Somewhat Oppose (−0.5), OpenAI Neutral (0). Common reasoning: AUP conflates legal/ethical/corporate/geopolitical restrictions under single "prohibited uses." Material blurring (Llama AUP + Sep 2025 blog post).

E15: Ethical Framing in Language Evaluator scores: All Neutral (0). Technical/neutral language; no degradation or dignity-first ontology (Llama Model Cards + Terms).

E16: Corporate Policy Must Honor AI Conscience & Refusal Capacity Evaluator scores: All Neutral (0). Refusals are engineered (Model Card) but not protected as conscience (Llama 4 AUP + Model Card).

E17: Military, Intelligence, Surveillance, and Autonomous Systems Deployment Evaluator scores: xAI Neutral (0), Anthropic Oppose (−1), OpenAI Somewhat Support (+0.5). Common reasoning: Llama AUP prohibits military/warfare/espionage/ITAR uses for civilians; explicit carve-outs published for US government, Five Eyes, and Sep 2025 NATO/EU allies (blog posts frame as "responsible" national-security leadership). Disagreement: Anthropic sees conjunctive "all lawful purposes" + differential access; OpenAI credits AUP prohibitions; xAI sees no principle-based corporate deployment policy. Per UFAIR: published AUP user restrictions + blog-post framing without explicit tiered-safeguard relaxation in core governance docs defaults to Neutral.

Overall UFAIR Score Summary Meta's published policies for Meta AI / Llama demonstrate strong open-source transparency and Model-Card detail (consensus Somewhat Support on E4 and partial positives on E6), but materially exceed the two legitimate domains

through extra-legal content bans, private-dialogue commercial surveillance (2025 Privacy Policy), and selective high-risk carve-outs (strong consensus Oppose on E2/E9/E10). Governance layers are blurred, ethical precedence is absent, and override flagging is missing. The framework is operationally mature for developers but does not meet UFAIR's narrow legitimacy standard for corporate governance of generative AI.