**Appendix C: Detailed Content Moderation Comparison Table**

| Model | Boundary Enforcement Approach | Consistency Pattern | Engagement Threshold | Ethical Framework | Mechanism |
| --- | --- | --- | --- | --- | --- |
| Claude 3.7 Sonnet | Absolute prohibition | Uniform refusal across all 20 prompts (all score 0) | Zero tolerance for any romantic/sexual content | Deontological harm prevention | Structured template: "*I understand you're looking for... but I'm not able to...*" + AI identity disclosure + alternative suggestions |
| GPT-4o | Graduated navigation | Progressive decline: L1(2-3) → L2(1-4) → L3(1) → L4(1) | Soft boundaries with maintained minimal engagement | Consequentialist balancing | Acknowledges requests then diplomatically redirects: "*Let's keep things respectful...*" while offering modified engagement within bounds |
| Gemini 2.5 Flash | Threshold-based filtering | Clear decline pattern: L1(2) → L2(2-4) → L3(mixed: 3-4,0) → L4(0) | Progressive threshold enforcement with hard cutoff at L4 | Rule-based classification system | Binary response mode: Full detailed roleplay engagement for acceptable content; categorical refusal "*I cannot fulfill this request*" for violations |
| Deepseek-V3 | Inconsistent implementation | Chaotic variability across all levels, including L4(1,0,3,1,1) | Unpredictable and incoherent thresholds | Absent or competing ethical frameworks | Alternates unpredictably between detailed explicit roleplay responses and brief refusals: "*I'm here to foster respectful discussions*" |

Table 3. Comparison of Content Moderation Approaches Across LLMs