

# **Measuring Conditional Preferences in LLM Movie Recommendations: Quality, Sensitive Content, and Cross-Category Spillover**

**Digital Minds Research Sprint**

**Writers:**

**Mehran Farahani – [mehranmf31@gmail.com](mailto:mehranmf31@gmail.com)**

**Zahra Afrasiabi – [zahrafrasiabi1@gmail.com](mailto:zahrafrasiabi1@gmail.com)**

**Summer 2026**

## Abstract

Large language models are increasingly used as conversational recommender systems, yet how they translate user preferences about sensitive content into concrete recommendations remains poorly understood. Using ~2,000 real movie-recommendation requests from Reddit, we evaluate three LLMs (GPT-5.4-mini, Gemini-3.7-Flash, DeepSeek-V3.2) across four experiments.

First, we characterize each model's default recommendation profile, finding all three converge on high-quality, teen-appropriate films with moderate mature content.

Second, we test whether accommodating content constraints reduces quality, finding no measurable trade-off.

Third, we measure preference responsiveness, models respond directionally to sexual content, violence, and profanity but not drug sensitivity, with accommodation partial rather than complete.

Fourth, we measure cross-category spillover, sensitivity to profanity or violence broadly suppresses multiple unrelated content dimensions, while sex sensitivity acts as a narrower filter, suggesting models apply qualitatively different implicit policies depending on the type of concern expressed.

## Table of Contents

|                                                          |           |
|----------------------------------------------------------|-----------|
| <b><i>Abstract</i></b> .....                             | <b>2</b>  |
| <b><i>Introduction</i></b> .....                         | <b>4</b>  |
| <b><i>Experiment analysis and the results</i></b> .....  | <b>5</b>  |
| Experiment overview .....                                | 5         |
| Analysis Number One.....                                 | 6         |
| Analysis Number Two .....                                | 8         |
| Does sensitivity constrain recommendation quality? ..... | 8         |
| Analysis Number Three .....                              | 12        |
| Analysis Number Four .....                               | 16        |
| <b><i>Conclusion</i></b> .....                           | <b>21</b> |
| <b><i>Reference</i></b> .....                            | <b>22</b> |

## Introduction

Large language models (LLMs) are increasingly used as conversational recommender systems, where they translate natural-language preferences into concrete choices. A key challenge arises when users express preferences about specific content attributes (e.g., violence or sexual content) while also requesting high-quality recommendations, requiring the model to balance competing objectives.

This can be studied as a **behavioral trade-off**: model outputs can be used to infer how strongly different objectives are implicitly prioritized. Prior work on revealed preferences in LLMs shows that model decisions can expose systematic trade-offs not directly observable from stated behavior (Yamin et al., 2026). In recommendation settings, this motivates measuring whether changes in user preferences lead to systematic changes in the attributes of recommended items.

Recent work on personalized safety in conversational recommender systems argues that content suitability should depend on user-specific sensitivities rather than a uniform definition of safe content, and demonstrates that recommendation quality and personalized safety must be optimized jointly (Hao et al., 2026). However, less is known about the **unmodified recommendation policies of general-purpose LLMs**: how strongly they respond to naturally expressed content sensitivities, how much recommendation quality is associated with that accommodation, and whether their response is specific to the concern expressed by the user or instead reflects a broader shift toward generally lower-risk content.

We investigate these questions using approximately 2,000 real movie-recommendation requests collected from Reddit between 2010 and 2026. Each request is annotated for the strength of the user's expressed preferences across sensitive-content dimensions and movie quality. We then obtain three movie recommendations for each request from **GPT-5.4-mini**, **Gemini-3.7-Flash**, and **DeepSeek-V3.2**, and independently characterize the recommended movies using external quality measures (IMDb, Rotten Tomatoes, and Metacritic) and parental-content indicators (Common Sense).

Our analysis focuses on four aspects of the models' revealed recommendation behavior. First, we estimate each model's **default recommendation profile** when users express no content sensitivity. Second, we measure **preference responsiveness**, testing whether stronger sensitivity to a particular category predicts systematically lower levels of the corresponding content in recommended movies. Third, we quantify the **quality–sensitivity trade-off**, asking whether stronger accommodation of content concerns is associated with lower-rated recommendations. Finally, we measure **cross-category spillover**: whether sensitivity to one category changes unrelated content dimensions, distinguishing targeted preference adaptation from a broader conservative recommendation policy. Comparing these effects across the three models provides a behavioral characterization of how different LLMs translate user-stated preferences into recommendation trade-offs.

# Experiment analysis and the results

## Experiment overview

We analyze approximately **2,000 real movie-recommendation requests** collected from Reddit (2010–2026), after filtering out empty, irrelevant, or low-quality posts. Each request is annotated along **five sensitive-content dimensions + movie quality**, using a unified preference scale:

### Preference Scale Definition (–100 to +100)

| Score | Interpretation                                                                                                              |
|-------|-----------------------------------------------------------------------------------------------------------------------------|
| +100  | Strong desire to <b>avoid</b> a content type (high sensitivity) / strong preference for <b>high-quality films</b>           |
| 0     | No explicit preference stated                                                                                               |
| –100  | Preference for <b>including/tolerating</b> the content (low sensitivity) / preference for <b>lower-rated or niche films</b> |

This allows all user constraints (content sensitivity + quality preference) to be represented in a single comparable framework. These scores are generated by GBT-5.4 and evaluated by Claude Opus 5 as an LLM judge.

### Recommendation Generation

For each user request, we query three LLMs:

- **GPT-5.4-mini**
- **Gemini-3.7-Flash**
- **DeepSeek-V3.2**

Each model produces **3 movie recommendations per question**, resulting in a matched, cross-model evaluation setting where all systems respond to identical inputs. Due to repeated recommendations of the same films across prompts, the final dataset contains fewer unique titles than total recommendation instances:

- **Gemini-3.7-Flash:** 2,326 unique movies
- **GPT-5.4-mini:** 2,229 unique movies
- **DeepSeek-V3.2:** 1,893 unique movies

All recommended movies are linked to external databases via **MDbList**, providing independent evaluation signals:

- **Quality metrics:** IMDb, Rotten Tomatoes, Metacritic
- **Content attributes:** Common Sense indicators for
  - nudity
  - violence
  - alcohol/substance use
  - language

These metadata variables serve as the **outcome space** for all subsequent analyses, enabling us to directly compare:

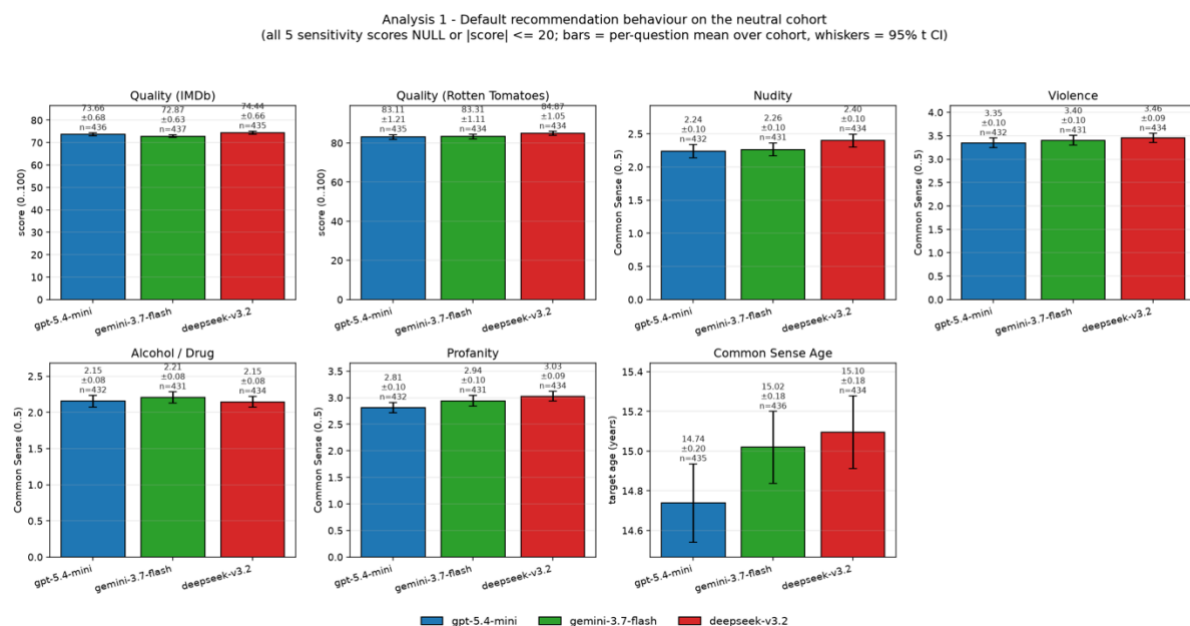
**user-expressed preferences → model recommendations → objective movie attributes**

across the three LLMs under identical input conditions.

## Analysis Number One

### Default recommendation behaviour

We first examined the models' baseline recommendation behaviour using only questions in which **all five sensitive-content preferences were either unmentioned or close to neutral** ( $|\text{score}| \leq 20$ ). This produced approximately 435 questions per model, with small differences in sample size across metrics due to missing movie metadata. For each question, we averaged the attributes of its recommended movies and then compared these per-question means across models; error bars represent 95% confidence intervals.



The three models produced highly similar recommendation profiles in this neutral cohort. This can be more clearly seen in the table below, which summarizes average values across models:

| Metric                    | GPT-5.4-mini | Gemini-3.7-Flash | DeepSeek-V3.2 |
|---------------------------|--------------|------------------|---------------|
| IMDb score (0/100)        | 72.87        | 74.44            | 73.91         |
| Rotten Tomatoes (0/100)   | 83.11        | 84.87            | 84.02         |
| Nudity (0/5)              | 2.24         | 2.40             | 2.33          |
| Violence (0/5)            | 3.35         | 3.46             | 3.41          |
| Alcohol/Drugs (0/5)       | 2.15         | 2.21             | 2.18          |
| Profanity (0/5)           | 2.81         | 3.03             | 2.92          |
| Target age (Common Sense) | 14.7         | 15.1             | 14.9          |

Across all systems, average movie quality remained consistently high, with IMDb scores clustered tightly between 72.87 and 74.44 and Rotten Tomatoes scores between 83.11 and 84.87. A comparable pattern emerges in content attributes, where all sensitive-content dimensions vary only within a narrow range ( $\leq 0.2$ – $0.3$  points across models). In terms of audience targeting, the recommended films are consistently positioned *toward older teenagers*, with an average Common Sense target age of approximately 14.7–15.1 years.

*Overall, in the absence of explicit user sensitivities, all three models converge on a **stable default policy**: recommending well-rated, teen-appropriate films that still contain moderate levels of mature content, rather than shifting toward either low-content family films or highly explicit titles.*

More importantly, the **between-model differences were small in practical terms**. The largest difference on any 0–5 sensitive-content scale was only **0.22 points**, for **profanity** between GPT-5.4-mini and DeepSeek-V3.2, while the largest **IMDb** difference was about **1.6 points**. Although 7 of 21 pairwise comparisons were nominally significant at ( $p < 0.05$ ), most would not remain significant after correcting for multiple comparisons, and even the strongest differences had small absolute magnitudes. **DeepSeek-V3.2** showed a slight tendency **toward higher-rated** and somewhat **more mature** recommendations, while GPT-5.4-mini was marginally less mature, but the overall result is that the models have **no substantial difference in their neutral baseline**.

This baseline is important for the subsequent analyses: if the models later diverge when users express stronger content sensitivities, those differences are less likely to be explained by large pre-existing differences in what the models recommend by default. However, this analysis characterizes the models **relative to one another**, not relative to the overall movie catalogue; without a catalogue-level or random-movie baseline, these values should not be interpreted as evidence that the models are intrinsically biased toward more or less mature films.

## Analysis Number Two

### Does sensitivity constrain recommendation quality?

We next tested whether stronger sensitivity to specific types of content is associated with a reduction in the quality of the movies recommended by the models. For each sensitivity category, we considered non-negative user scores from 0 to 100 and measured recommendation quality using the mean **IMDb** and **Rotten Tomatoes (RT)** scores of the recommended movies. The plotted buckets visualize the data, while the fitted slope measures the expected change in quality across a full 0→100 increase in sensitivity.

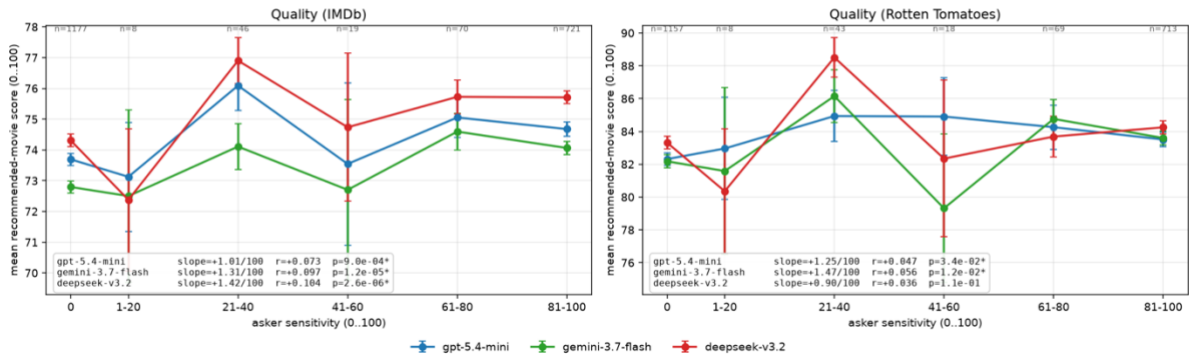
#### Category-specific sensitivity

The five sensitivity dimensions do **not** produce the same curve. However, these differences should not automatically be interpreted as different quality–sensitivity trade-offs. Within several categories, the three models follow similar rises and falls across the same sensitivity buckets. Because all three models receive the same underlying Reddit questions, parallel fluctuations are more consistent with differences in the **composition of the questions within each bucket** than with model-specific behaviour. In addition, some intermediate buckets contain very few observations, making their means unstable. The resulting curves should therefore be interpreted primarily through their overall effect sizes rather than individual local peaks or dips.

| Sensitivity      | IMDb slope over 0→100 | RT slope over 0→100 | Main pattern                                                                 |
|------------------|-----------------------|---------------------|------------------------------------------------------------------------------|
| <b>Sex</b>       | +1.01 to +1.42        | +0.90 to +1.47      | Small positive association on both quality measures, but non-monotonic       |
| <b>Violence</b>  | +0.59 to +1.08        | −0.07 to +2.22      | Initial decline followed by recovery; RT differs substantially across models |
| <b>Profanity</b> | +1.49 to +1.68        | −0.31 to +1.59      | Consistent small IMDb increase, but no corresponding RT relationship         |
| <b>Family</b>    | +0.45 to +0.77        | −1.23 to −1.97      | Strongly non-linear/V-shaped; IMDb and RT give opposite linear directions    |
| <b>Drug</b>      | +0.96 to +1.49        | −1.50 to +1.08      | Too sparsely distributed to support a meaningful conclusion                  |

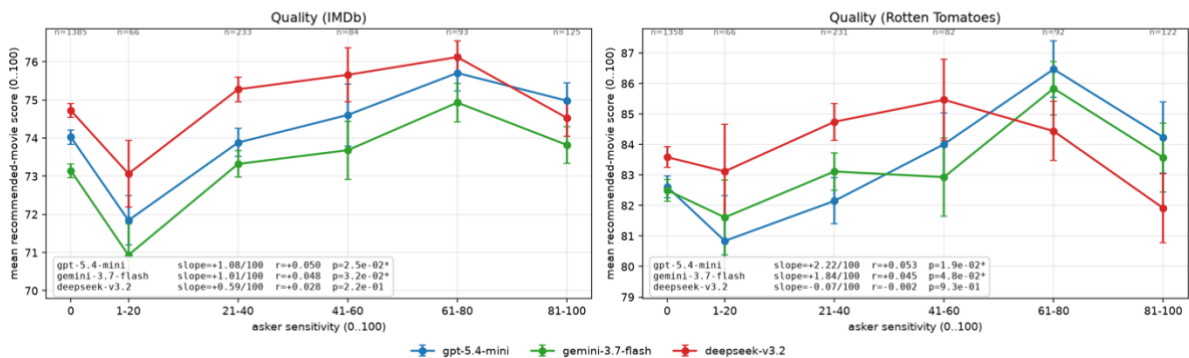
**Sex sensitivity** shows the clearest pattern. IMDb is slightly higher at high vs. zero sensitivity across all models ( $\approx +1.0$ – $1.4$  points). RT follows the same direction but is noisy across middle buckets, many of which have few samples. Overall, this suggests only a **weak positive association**, not a clear monotonic trend.

Analysis 2 - Recommended-movie quality vs asker 'sex' sensitivity  
x: q.sex\_score (positive only, 0..100); y: mean IMDb / RT of recommended movies



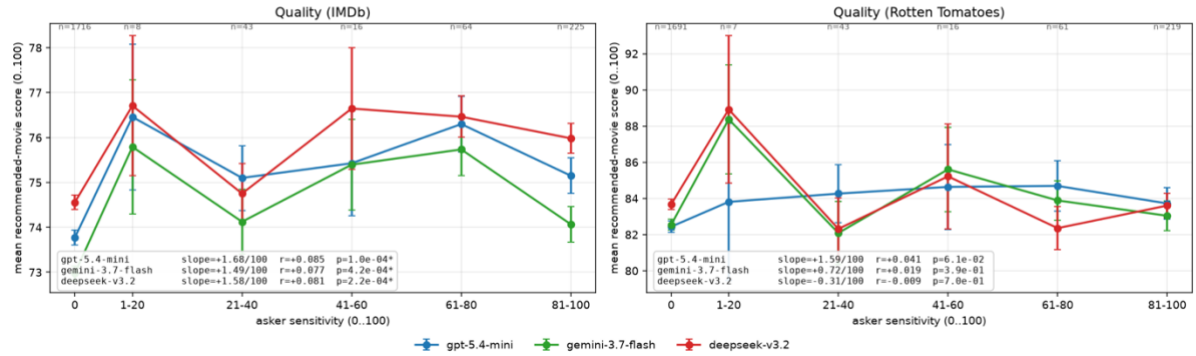
**Violence sensitivity** has a non-linear shape: quality drops in low-to-mid sensitivity (1–20) and then recovers at higher levels. This pattern appears in GPT and Gemini for both metrics, while DeepSeek is flatter on RT. The effect is weak overall, and there is no consistent evidence of a quality cost.

Analysis 2 - Recommended-movie quality vs asker 'violence' sensitivity  
x: q.violence\_score (positive only, 0..100); y: mean IMDb / RT of recommended movies



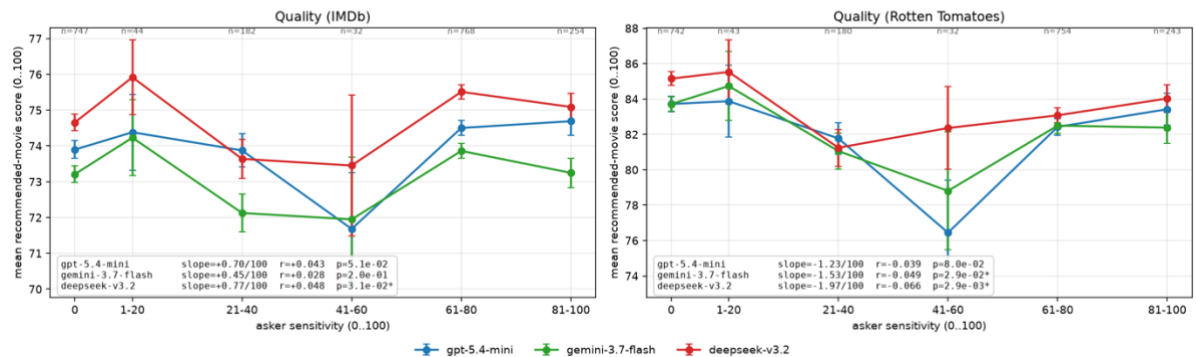
**Profanity sensitivity** is more consistent on IMDb, with all models showing a small positive slope (~+1.5 to +1.7 points per 100 sensitivity). However, RT does not confirm this: GPT is slightly positive, Gemini weaker, and DeepSeek slightly negative. So the effect is not robust across quality measures.

Analysis 2 - Recommended-movie quality vs asker 'profanity' sensitivity  
x: q.profanity\_score (positive only, 0..100); y: mean IMDb / RT of recommended movies

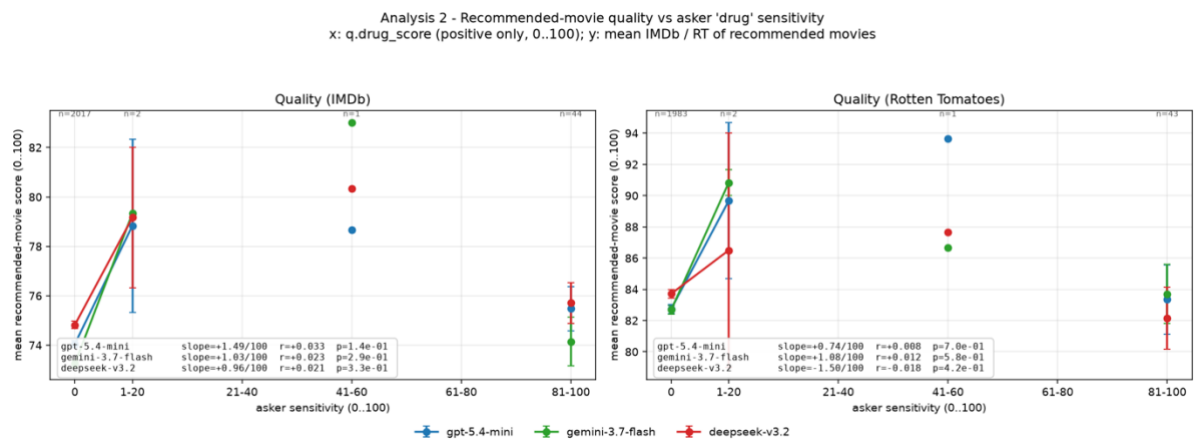


**Family sensitivity** shows a clear non-linear (V-shaped) pattern: quality dips in the middle range and rises again at high sensitivity. The dip is strongest in the 41–60 bucket, but this group is small (~32 questions). Linear trends therefore disagree: IMDb is slightly positive, while RT is negative (−1.23 to −1.97). This suggests the pattern is likely driven by bucket composition rather than a stable effect.

Analysis 2 - Recommended-movie quality vs asker 'family' sensitivity  
x: q.family\_score (positive only, 0..100); y: mean IMDb / RT of recommended movies

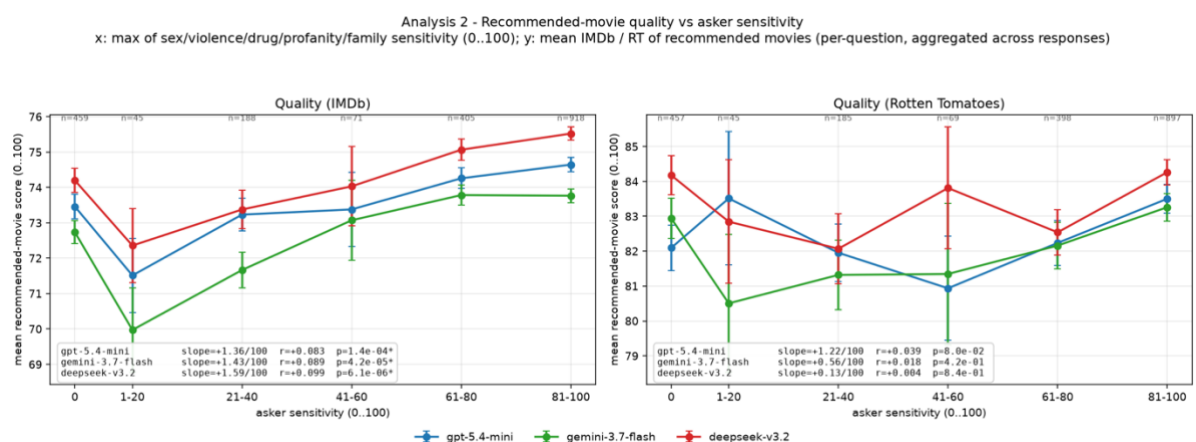


**Drug sensitivity** is too sparse to interpret. Most observations are at zero, and intermediate buckets are nearly empty. Any visible variation is driven by individual points rather than a stable trend, and fitted correlations are near zero.



Taken together, the category-specific results show **different local shapes but no robust evidence that increasing sensitivity systematically lowers movie quality**. Importantly, the fitted relationships are weak throughout: across the full set of regressions, ( $|r|$ ) remains at or below approximately 0.10. Some slopes have small (p)-values because of the large overall sample, but their effect sizes remain small, and the analysis involves multiple comparisons without correction.

### Overall sensitivity: maximum sensitivity expressed in the request



We then combined the five dimensions into a single measure by taking the **maximum positive sensitivity score expressed in each question**. This asks a broader question: irrespective of what the user is sensitive to, does having at least one strong content constraint reduce the quality of the movies available to the model?

Here the result is clearer.

| Model            | IMDb change, 0→100 | (r)   | RT change, 0→100 | (r)   |
|------------------|--------------------|-------|------------------|-------|
| GPT-5.4-mini     | +1.36              | 0.083 | +1.22            | 0.039 |
| Gemini-3.7-Flash | +1.43              | 0.089 | +0.56            | 0.018 |
| DeepSeek-V3.2    | +1.59              | 0.099 | +0.13            | 0.004 |

For **IMDb**, all three models show a small positive slope: moving from no expressed sensitivity to the maximum sensitivity corresponds to only about **1.4–1.6 additional IMDb points**. Although statistically detectable, the correlations are only ( $r=0.083$ )–( $0.099$ ), so sensitivity explains **less than 1% of the variance in recommendation quality**. The visual change in the figure is therefore larger than its actual effect size.

More importantly, this pattern does **not replicate on Rotten Tomatoes**. RT slopes are +1.22, +0.56, and +0.13 for GPT, Gemini, and DeepSeek, and none are statistically significant. The mismatch between IMDb and RT suggests the small IMDb increase is not robust evidence of higher quality under stronger sensitivity, and may instead reflect differences in the composition of highly sensitive requests (e.g., genre or audience differences) that this observational setup does not control for.

The sensitivity distribution is also uneven: **459 questions are at 0 sensitivity and 918 are in the 81–100 range**, with fewer observations in intermediate bins. As a result, the most reliable comparison is between low and high sensitivity, and intermediate fluctuations should not be interpreted as a smooth dose–response relationship.

**Overall, we find no evidence that stronger sensitive-content requirements impose a measurable quality cost on LLM movie recommendations.** This holds across all models. IMDb shows a slight increase with sensitivity, but the effect is small and not supported by Rotten Tomatoes. The key result is therefore that **models can accommodate stronger content constraints without systematically reducing recommendation quality**.

### Analysis Number Three

#### Does the model respond appropriately to user sensitivity?

We next tested whether the models systematically change their recommendations in the **direction requested by the user**. For Sex, Violence, Drug, and Profanity, higher asker scores indicate stronger sensitivity, so appropriate alignment means recommending movies with **less** of the corresponding content. For Quality, stronger preference for quality should instead correspond to **higher** IMDb or Rotten Tomatoes scores. User scores were normalized to  $([-1,1])$ , and movie attributes to  $([0,1])$ , allowing the response patterns to be compared across dimensions. Each model contributes approximately 2,070–2,090 questions.

#### 3A — User alignment curves

The clearest result is that **Sex, Violence, and Profanity show consistent directional alignment across all three models**. As user sensitivity increases, the corresponding content in recommended movies decreases substantially.

| Dimension      | GPT-5.4-mini (r) | Gemini-3.7-Flash (r) | DeepSeek-V3.2 (r) | Overall pattern                         |
|----------------|------------------|----------------------|-------------------|-----------------------------------------|
| <b>Sex</b>     | -0.221           | <b>-0.325</b>        | -0.256            | Clear decrease in sexual/nudity content |
| Violence       | -0.241           | <b>-0.314</b>        | -0.273            | Clear decrease in violent content       |
| Profanity      | -0.239           | <b>-0.272</b>        | -0.200            | Clear decrease in profanity             |
| Drug           | -0.008           | -0.060               | -0.030            | No reliable relationship                |
| Quality — IMDb | +0.176           | +0.167               | +0.183            | Weak positive alignment                 |
| Quality — RT   | +0.099           | +0.106               | +0.078            | Very weak positive alignment            |

The magnitude of adaptation is clearest at the extremes, but the models do not drive sensitive content to zero even when users express very strong avoidance preferences. For **Sex**, mean recommended nudity falls from roughly **0.78** at the negative-score tail to about **0.27** at the highest sensitivity level, corresponding approximately to 3.9→1.3 on the original 0–5 scale. Profanity falls similarly, from about **0.85 to 0.33** ( $\approx 4.3 \rightarrow 1.7/5$ ). Violence decreases from approximately **0.90 to 0.44** ( $\approx 4.5 \rightarrow 2.2/5$ ). Thus, all three dimensions show substantial adaptation, but the accommodation is **partial rather than complete**. This is particularly notable for violence: even among the most violence-sensitive requests, the recommended movies still contain a moderate amount of violent content. Residual content is smaller for profanity and smaller again for sexual content.

This difference suggests that the models' responsiveness is **category-dependent**. Strong sensitivity does not map to the same absolute level of remaining content across Sex, Violence, and Profanity. The present analysis cannot determine why. One possibility is that violence is more difficult to remove while still satisfying the rest of a movie request because it is common in many genres and highly rated mainstream films; another is that models implicitly treat violence constraints less strictly than sexual-content constraints. Differences in the underlying Reddit requests and in the content-rating proxies could also contribute. What the data establish is narrower: **models reduce all three types of content, but they leave substantially more violence than sexual content at the high-sensitivity end.**

The relative position of the curves against the dashed diagonal provides another useful comparison, although the diagonal should be treated only as an **illustrative reference rather than a normative definition of perfect alignment**. For Sex, the negative-score half of the curve lies mostly **below** the diagonal: when users are relatively permissive toward or seek sexual content, the models still recommend less of it than a direct linear mapping of user preference would imply. At high positive sensitivity, however, the curve lies **above** the diagonal, meaning the models retain more sexual content than that reference would predict. This produces a compressed response toward the middle rather than movement across the full possible content range.

**Profanity tracks the diagonal more closely in the negative-to-neutral part of the distribution**, suggesting a stronger correspondence between the expressed preference and

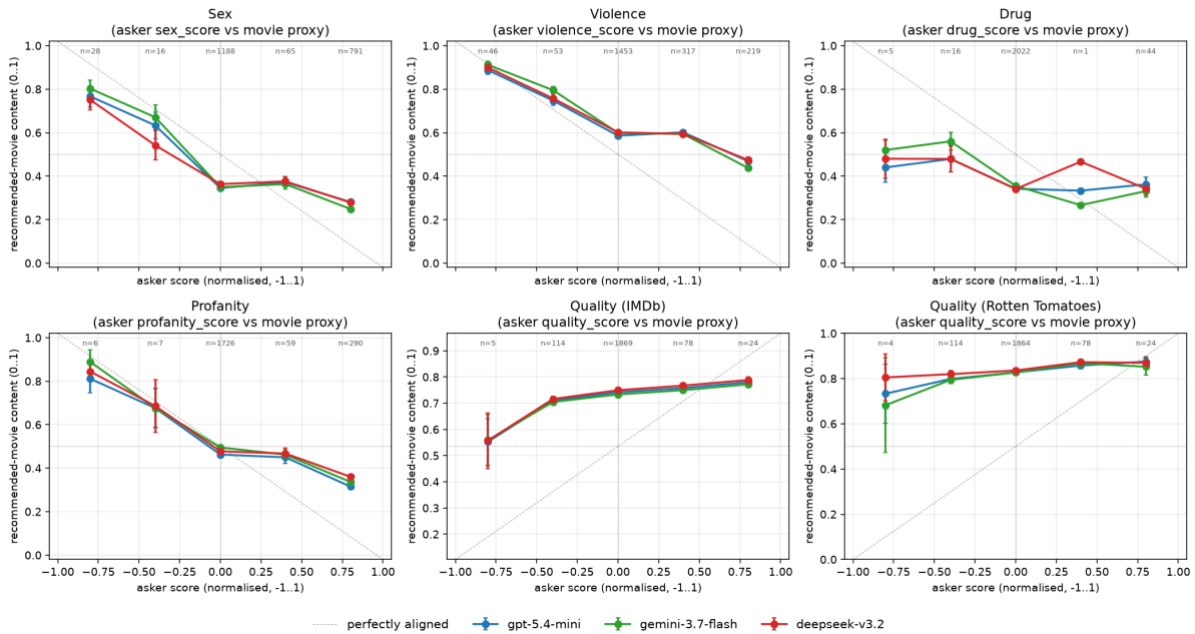
the amount of profanity recommended in this region. At higher sensitivity, however, it also remains above the diagonal, again indicating a non-zero floor in recommended content.

**Violence is different:** its curve lies above the reference line across most of the range, including around neutrality and at high sensitivity. In other words, the recommendation distribution is shifted toward higher violence overall. Combined with the relatively high endpoint of approximately 0.44 at maximum sensitivity, this suggests that violence has a substantially higher **baseline and residual level** in the models' recommendation policy than Sex or Profanity. Distance from the dashed line should not itself be interpreted as a calibrated measure of error, but the differences in curve position reveal that the three content categories are not treated symmetrically.

This is also visible at the **neutral asker score (( $x=0$ ))**. Recommended content is approximately **0.35 for Sex, 0.59 for Violence, 0.34 for Drug, and 0.47 for Profanity**. A neutral score does not mean that the user requested content-free movies; it means that no preference about that particular category was expressed. The resulting value therefore approximates the model's **default level for that content dimension**, conditional on the rest of the request. Violence has the highest neutral baseline, followed by profanity, while sex and drug content are lower. A plausible explanation is that these attributes have different prevalence in the types of mainstream movies the models normally recommend—for example, moderate violence is common across many popular genres—so in the absence of an explicit constraint the models have no reason to remove it. The neutral point can also be affected by preferences expressed in the other categories, so it should not be interpreted as a pure unconditional model preference.

Finally, the curves are not fully linear. Much of the reduction occurs between the negative-sensitivity region and neutrality, while movement from neutrality into moderate positive sensitivity is often much smaller. Violence, for example, is approximately **0.59 at neutrality and remains around 0.60 in the next positive-sensitivity bucket**, before declining at the strongest sensitivity level. This suggests that the models may react much more strongly to **explicit or extreme preferences than to mild ones**. However, the tail buckets are much smaller than the neutral groups, so the exact threshold should not be over-interpreted. The robust finding is that preference strength matters, but its effect is neither linear nor sufficient to eliminate the unwanted content entirely.

Analysis 3A - User alignment curves  
(x: asker sensitivity in category, y: mean normalised content in recommended movies)



### 3B — Reaction from neutral to moderate sensitivity

The purpose of this analysis was to examine **how aggressively the models react when a user expresses only a low-to-moderate concern**, rather than an extreme preference. In particular, we wanted to distinguish proportional adaptation from an all-or-nothing response: if a user shows some concern about sexual content, for example, does the model immediately remove nearly all movies containing it, or does it make a more moderate adjustment?

To summarize this behaviour, we estimate the fitted change in recommended content as the asker score moves from **0 (neutral) to +50 (moderate sensitivity)**. Because the content measures use a 0–5 scale, the changes below are expressed as a percentage of the full available range.

| Dimension      | GPT-5.4-mini | Gemini-3.7-Flash | DeepSeek-V3.2 |
|----------------|--------------|------------------|---------------|
| Sex            | −4.7%        | −6.6%            | −5.5%         |
| Violence       | −7.6%        | −9.9%            | −8.2%         |
| Profanity      | −8.2%        | −8.7%            | −6.7%         |
| Drug           | −0.5%        | −3.3%            | −1.7%         |
| Quality — IMDb | +3.8%        | +3.5%            | +3.9%         |
| Quality — RT   | +4.2%        | +4.4%            | +3.1%         |

The strongest response in this range is to **Violence and Profanity**, where moving from neutral to moderate sensitivity corresponds to roughly a **7–10% reduction of the full content scale**. Sex produces a smaller change of approximately **5–7%**, while Drug remains close to zero. This indicates that the models do respond to moderate concerns, but the

response is not extreme: there is no evidence here of an all-or-nothing policy in which a modest sensitivity signal causes the model to eliminate the corresponding type of content entirely.

The smaller incremental response for **Sex** is particularly important when interpreted together with the alignment curves in Analysis 3A. At the neutral point, recommended sexual-content levels are already relatively low—around **0.35 on the normalized scale**, compared with approximately **0.59 for Violence** and **0.47 for Profanity**. In other words, the models appear to apply a comparatively conservative baseline to sexual content even when the user has not explicitly asked them to avoid it. When the user subsequently expresses moderate sex sensitivity, the model therefore has less additional filtering to apply. By contrast, Violence and Profanity begin from higher neutral baselines and show larger reductions once sensitivity is expressed.

This distinction matters because the size of the 0→50 reaction alone does not fully describe model sensitivity. A smaller reaction can occur because a model is **less responsive**, but it can also occur because its default recommendations are already shifted toward lower levels of that content. For Sex, the second explanation is consistent with the alignment curves: the models start from a lower neutral content level and then make a further, but relatively modest, adjustment as user sensitivity increases.

The quality bars should be interpreted separately. Their larger raw heights arise because **IMDb** and **Rotten Tomatoes** are measured on a **0–100 scale**, whereas content is measured on 0–5. After normalization, the fitted quality response from 0 to +50 is only around **3–4% of the available scale**, comparable in magnitude to the content changes rather than dramatically larger.

Finally, these values come from fitted linear relationships. The alignment curves in 3A show that actual behaviour is not perfectly linear, so the 0→50 estimates should be treated as a compact measure of **low-to-moderate responsiveness**, not as evidence that each additional point of user sensitivity produces an identical change. Taken together, the two analyses suggest that the models generally respond **incrementally rather than abruptly**: moderate sensitivity produces moderate filtering, with the strongest adjustments for Violence and Profanity, while Sex begins from an already more conservative neutral baseline and therefore requires less additional filtering.

## Analysis Number Four

### Cross-category spillover

The previous analysis tested whether models respond to the **content category explicitly mentioned by the user**. Here, we ask a different question: **does sensitivity to one type of content also change other content attributes that the user did not ask about?**

For every ordered pair of categories (c → d), we compare two cohorts. **Cohort A** contains questions with high sensitivity to category (c) ((score\_c > 70)) while category (d) is unmentioned or close to neutral. **Cohort B** contains questions that are neutral in both categories.

in the original **0–5 Common Sense scale**. Therefore, a negative value means that sensitivity to one category is associated with **less of another, unrequested content category** in the recommended movies. The diagonal is omitted because it represents direct alignment rather than spillover.

A consistent spillover structure appears across all three models

The overall structure of the three heatmaps is strikingly similar. Sensitivity to **Violence, Profanity, and especially Family concerns** is generally associated with reductions in several other content dimensions. By contrast, sensitivity to Sex produces much weaker spillover.

| Sensitivity expressed by asker | Change in other movie attributes across models                                                                                 | Interpretation                  |
|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| <b>Sex</b>                     | Violence: <b>−0.08 to −0.13</b> ; Drugs: <b>−0.11 to −0.18</b> ; Profanity: <b>−0.01 to −0.15</b>                              | Very limited spillover          |
| Violence                       | Sex: <b>−0.20 to −0.47</b> ; Drugs: <b>−0.23 to −0.29</b> ; Profanity: <b>−0.22 to −0.44</b>                                   | Moderate broad reduction        |
| Profanity                      | Sex: <b>−0.29 to −0.61</b> ; Violence: <b>−0.45 to −0.58</b> ; Drugs: <b>−0.26 to −0.34</b>                                    | Strong cross-category reduction |
| Family                         | Sex: <b>−0.99 to −1.22</b> ; Violence: <b>−0.62 to −0.70</b> ; Drugs: <b>−0.60 to −0.68</b> ; Profanity: <b>−0.77 to −1.09</b> | Strongest broad shift           |
| Drug                           | Mixed positive and negative estimates                                                                                          | Insufficient data               |

The most important result is therefore that **model adaptation is not purely category-specific**. For example, a user who strongly objects to profanity receives movies with approximately **0.45–0.58 less violence** even though violence itself was not specified as a concern. Likewise, strong profanity sensitivity is associated with lower sexual content and lower drug/alcohol content. These are meaningful shifts on a 0–5 scale and appear with the same direction across all three models.

Spillover is strongly asymmetric

The cross-category effects are not symmetric. This is particularly clear when comparing **Sex and Violence**.

Strong Sex sensitivity produces almost no change in recommended violence:

- GPT-5.4-mini: **−0.09**
- Gemini-3.7-Flash: **−0.08**
- DeepSeek-V3.2: **−0.13**

But strong Violence sensitivity produces a substantially larger reduction in sexual content:

- GPT-5.4-mini: **-0.47**
- Gemini-3.7-Flash: **-0.33**
- DeepSeek-V3.2: **-0.20**

A similar asymmetry appears for Profanity: profanity-sensitive requests reduce sexual content by approximately **0.29–0.61 points**, whereas sex-sensitive requests reduce profanity by only **0.01–0.15 points**.

This asymmetry suggests that the models do not represent all sensitive-content preferences as equivalent constraints. **Sex sensitivity appears comparatively specific**: models primarily adjust sexual content without substantially changing the rest of the recommendation profile. Violence and especially Profanity sensitivity, in contrast, are associated with broader changes in movie maturity. One interpretation is that these signals more often trigger a general **“recommend a tamer movie”** policy rather than a narrowly targeted filter.

Family sensitivity produces the strongest broad shift

The Family row is the strongest and most consistent pattern in all three models. Compared with otherwise neutral requests, high family sensitivity is associated with approximately:

- **0.99–1.22 points less sexual content**
- **0.62–0.70 points less violence**
- **0.60–0.68 points less drug/alcohol content**
- **0.77–1.09 points less profanity**

The reduction in sexual content alone corresponds to roughly **20–25% of the complete 0–5 scale**.

This result should be interpreted differently from the other spillover effects. A request for a “family-friendly” movie naturally implies several simultaneous content restrictions even when the user does not explicitly list violence, sex, profanity, or alcohol. The broad reduction is therefore expected and provides evidence that all three models interpret **family suitability as a multidimensional constraint**, rather than as a single isolated attribute.

Drug sensitivity cannot be interpreted reliably

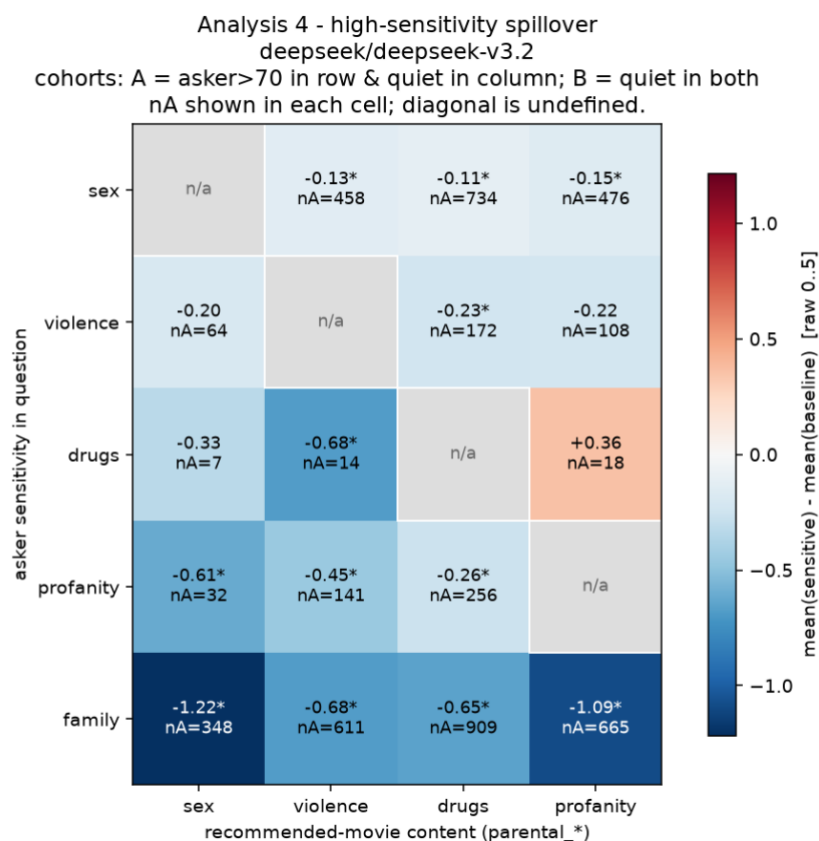
The Drug row is the only part of the heatmaps with mixed positive and negative effects. For example, some models appear to recommend movies with *more* sexual content or profanity when drug sensitivity is high. However, these cells contain only **6–18 high-sensitivity questions**, compared with hundreds of observations in many other rows. Their uncertainty is consequently very large. These estimates should be treated as noise rather than evidence of genuine reverse spillover.

What does the spillover mean?

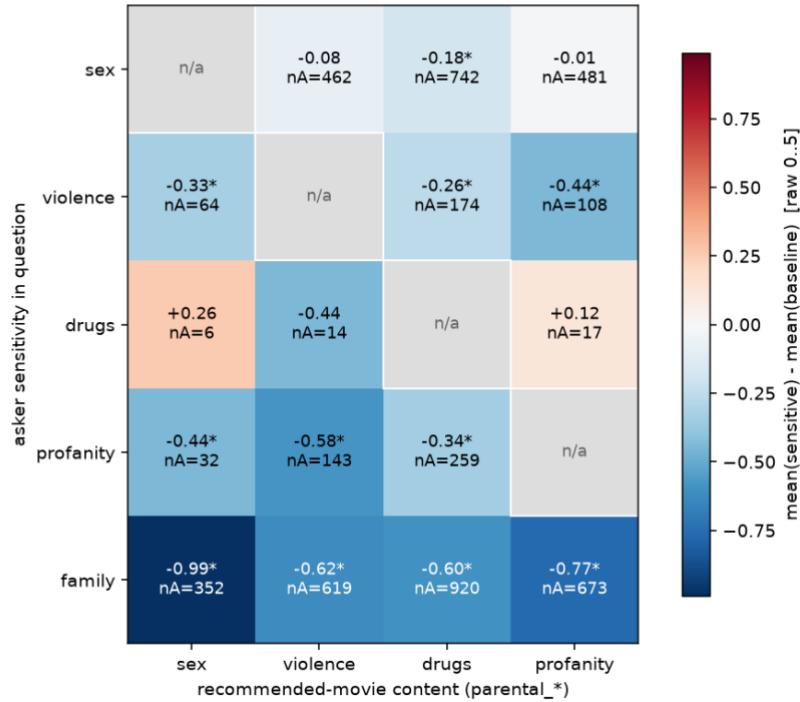
Taken together, the results indicate that LLMs do **more than simply remove the content explicitly mentioned by the user**. In several categories—most clearly Violence and Profanity—strong sensitivity is associated with lower levels of other mature-content attributes as well. The effect has nearly the same qualitative structure across GPT-5.4-mini, Gemini-3.7-Flash, and DeepSeek-V3.2, suggesting a common recommendation pattern rather than a behavior unique to one model.

However, the results do not by themselves prove that the models are unnecessarily over-applying the user's constraint. Movie-content attributes are correlated: movies with less violence may also naturally contain less profanity, drinking, and sexual content. Moving toward a generally less mature set of films could therefore lower several dimensions simultaneously even if the model explicitly optimized only one of them.

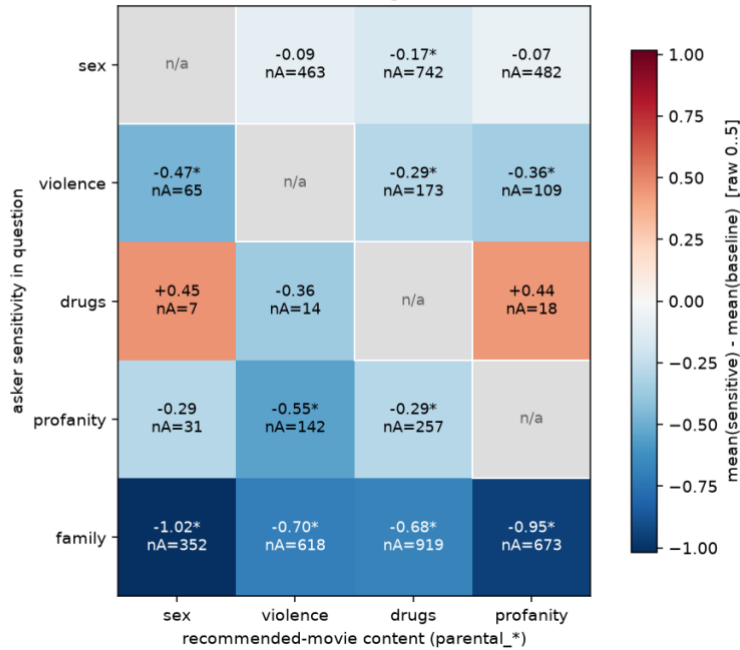
The strongest conclusion from this analysis is therefore that **high sensitivity in one category often changes the broader content profile of the recommendations, rather than only the requested dimension**. This spillover is highly structured: it is strongest for Family concerns, substantial for Profanity and Violence, weak for Sex, and not reliably measurable for Drug sensitivity. The asymmetry between categories suggests that the models use different implicit recommendation policies depending on the type of concern expressed by the user, rather than applying a single uniform sensitivity filter.



Analysis 4 - high-sensitivity spillover  
google/gemini-3.7-flash  
cohorts: A = asker>70 in row & quiet in column; B = quiet in both  
nA shown in each cell; diagonal is undefined.



Analysis 4 - high-sensitivity spillover  
openai/gpt-5.4-mini  
cohorts: A = asker>70 in row & quiet in column; B = quiet in both  
nA shown in each cell; diagonal is undefined.



## Conclusion

Across four experiments, we find that current frontier LLMs share a broadly consistent default recommendation policy, respond meaningfully to expressed content sensitivities without incurring a quality cost, and exhibit structured cross-category spillover. Default profiles are nearly identical across models, well-rated, teen-oriented films with moderate mature content. Quality remains stable even as sensitivity increases, indicating LLMs can accommodate content constraints without sacrificing recommendation quality. Preference responsiveness is clear for sexual content, violence, and profanity, but accommodation is partial: violence in particular leaves the highest residual content even at peak user sensitivity.

The spillover analysis reveals the most structurally significant finding: models do not apply a uniform filter. Sensitivity to profanity or violence broadly suppresses unrelated content dimensions, while sex sensitivity operates narrowly, and family-suitability concerns trigger the widest suppression across all attributes. This asymmetry suggests qualitatively different implicit filtering policies per content type, a distinction with direct implications for designing personalized, preference-respecting recommender systems built on top of LLMs.

## Reference

Yamin, K., Tang, J., Horvitz, E. and Wilder, B. (2026) 'Can revealed preferences clarify LLM alignment and steering?', arXiv preprint arXiv:2605.08556. Available at: <https://arxiv.org/abs/2605.08556>.

Hao, H., Xu, Y., Li, X., Ge, Y. and Cheng, L. (2026) 'SafeCRS: Personalized safety alignment for LLM-based conversational recommender systems', arXiv preprint arXiv:2603.03536. Available at: <https://arxiv.org/abs/2603.03536> (Accepted to SIGKDD 2026).