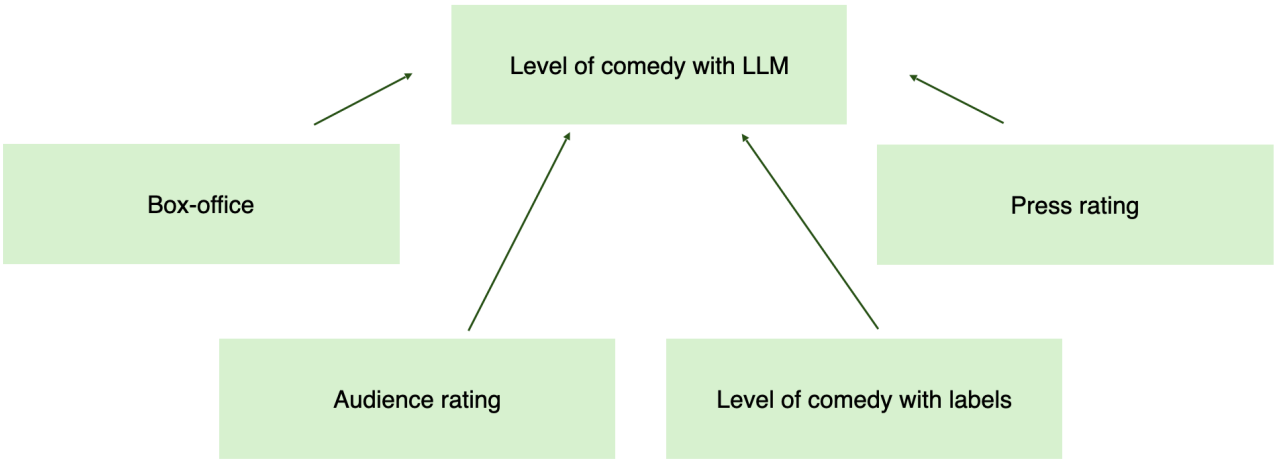
Comedy in Critics through LLM classification

Problematic : How does the level of comedy in films relate to their commercial success and critical reception?



Predictions:

- **P1** - The higher a movie's level of comedy, the greater its box-office success.
- **P2** - The higher a movie's level of comedy, the better its audience rating.
- **P3** - The higher a movie's level of comedy, the lower its rating from the press.
- **P4 -** Films that receive higher comedy scores from a large language model also tend to exhibit higher comedy levels according to genre labels

Data

The only variable that differs from the previous study is the 'Comedy' variable, which is constructed using a different methodology. By definition, annotating data with a large language model (LLM) is computationally time-consuming, as each observation must be processed individually. Consequently, it is not feasible to annotate as many films as in our earlier analysis. We therefore set an annotation threshold of 500 films.

To maximize the likelihood that the LLM possesses latent knowledge of the selected films, we restricted our sample to the 500 most popular films according to TMDB's *Popularity* variable. Popularity on TMDB is a dynamic score computed on a daily basis by aggregating user interactions, including page views, votes, and additions to favorites or watchlists. The algorithm also

incorporates a recency factor, such that popularity scores naturally decay over time in order to emphasize current trends. Based on this mechanism, we assume that the LLM is more likely to have internalized information about these highly popular films than about less visible ones.

For the classification task, we made a trade-off between the LLM's latent knowledge capacity and its number of parameters, as larger models cannot be run efficiently on all machines or on commonly used cloud services such as Google Colab (free tier). After reviewing several benchmark results, we ultimately selected **Qwen2.5-7B** as a compromise between performance and computational feasibility.

We initially designed a prompt and tested it on a small subset of films. A recurring issue was that Qwen tended to assign excessively high comedy scores (often 10/10) to films that are better characterized as mixed-genre comedies. Moreover, the model appeared to infer or guess rather than rely strictly on its latent knowledge, which ran counter to our methodological objectives. We therefore revised the prompt to address these issues, incorporating explicit constraints designed to minimize hallucination and speculative reasoning. The final prompt is presented below.

You are an expert film critic and genre analyst. Your task is to assign a "Comedy Score" (from 1 to 10) to the following movie.

MOVIE: "{movie_title}"

CORE PRINCIPLE:
Assess the **intent** of the movie to provoke laughter. Do not judge quality, but the density and purity of humor.

RATING SCALE (Strictly follow these nuances):
- **10 (Undiluted Comedy)**: The ONLY goal is laughter. The film **cannot** be classified as another genre (Action, Romance, Drama). The plot is merely a vessel for gags, farce, or absurdity. (e.g., Airplane!, Monty Python).
- **8-9 (Genre-Hybrid Comedy)**: Extremely funny, BUT the movie also functions as another genre. It is an Action-Comedy, a Romantic-Comedy, or a Satire. The stakes feel somewhat real despite the jokes. (e.g., Deadpool , Barbie, There's Something About Mary).
- **5-7 (Balanced Dramedy / Action)**: Humor is a major ingredient (50%), but shares the spotlight equally with serious plot points, tension, or character drama. There is a solid narrative backbone
- **3-4 (Serious with Comic Relief)**: The movie is a Thriller, Adventure, or Drama first. Humor is used only to break tension, but the main story is serious.
- **1-2 (Serious / Dark / Pure Drama)**: No intent to be funny. The atmosphere is grave, tragic, or strictly factual. Even if the dialogue is intelligent ("witty"), if the goal isn't laughter, it belongs here.

IMPORTANT RULES:
1. **The "Hybrid" Ceiling:** If a movie relies on Action sequences (fights) or a Romance arc to move the story forward, the maximum score is **9**. A 10 is reserved for pure structural chaos/comedy.
2. **Wit vs. Comedy:** Do not confuse clever dialogue with comedy. A serious drama with sharp dialogue is a 1 or 2 or 3.
3. **Unknown Movies:** If you do not know the movie enough, strictly output 0. Do not try to guess.

RARITY & EXCEPTION RULE:
A score of 10 should apply to at most ~1–2% of all narrative feature films ever made.
Only assign 10 or 9 if removing comedy would destroy the film entirely.
If the film still functions as Action, Romance, Drama, or Satire without jokes,
the maximum score is 7.
When in doubt between two numbers, always choose the lowest.

RESPONSE FORMAT (Strictly follow this line-by-line structure):
Line 1: Knowledge Level: [High/Medium/Low/None]
Line 2: Justification: [Your 2-3 sentences analysis]
Line 3: ### Note : [Integer only]

As shown by the prompt design, we explicitly distinguish structural comedy, in which the primary narrative function is to generate laughter, from comic relief, which serves merely to alleviate tension within an otherwise serious or dramatic storyline. The prompt operationalizes this distinction through a finely graded rating scale and a set of restrictive rules that prioritize authorial intent over perceived quality, while also discouraging speculative judgments when the model lacks sufficient knowledge. In particular, the inclusion of explicit ceiling effects, rarity constraints, and a strict "unknown movie" rule was intended to force conservative scoring behavior and minimize overestimation.

Using this prompt, we inferred comedy scores for the selected sample of 500 films and collected, for each observation, the numerical score, a short textual justification, and an explicit self-reported level of certainty regarding the model's knowledge of the film.

To further ensure that the model did not rely on conjecture or hallucinated information, we applied a series of post-processing filters. Specifically, we removed all observations with a score of 0, all cases in which the model's reported knowledge level was not *High*, and all responses containing lexical markers of uncertainty such as *"appears"* or *"appeared"*. This filtering process resulted in the exclusion of 60 observations, yielding a final dataset of N = 440 films.

Qwen produced relatively heterogeneous scores: while there is no strong overall overrepresentation of numerical values, the distribution remains asymmetric for certain digits, most notably an overrepresentation of the value 3 and the complete absence of the value 4 (see **Figure 1.e**).

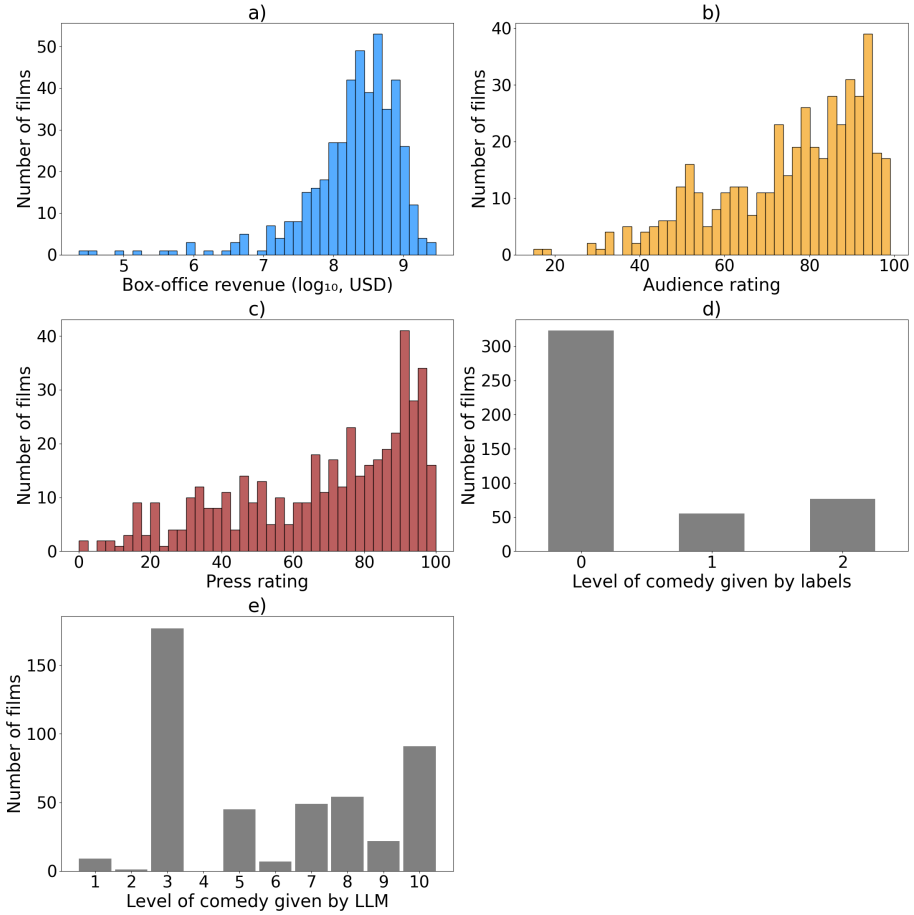


Figure 1 - Repartition according to the number of movies of a) the box-office revenue ($\log_{10}$, USD), b) the audience rating (from 0 to 100), c) the press rating (from 0 to 100), d) the level of comedy given by labels (with 0: No Comedy, 1: Comedy mixed with another genre, 2 : Pure comedy) and e) the level of comedy given by the Qwen2.5-7B model

Method

To test our three predictions, we compared box-office revenue, audience ratings, and press ratings as a function of the level of comedy (**Figure 3.d**, **3.e**, and **3.f**). We conducted three simple linear regressions, assuming equal variance and an equivalent gap between categories 0 and 1, and between 1 and 2. Because the distribution of revenues is highly skewed, with many low-revenue movies and only a few high-revenue ones, we applied a logarithmic (base 10) transformation to the revenue variable. The corresponding formula is:

$$Ri = b0 + b1Ci + \varepsilon i$$

Where :

R_i : the rating or revenue of movie *i*,

C_i : the comedy level of movie *i*,

ε_i : the error term.

Results

The regression shows that the level of comedy in a film increases slightly with its log box-office revenue (b = 0.004, R^2 = 0.000, see **Figure 3.b**), but the result is not significant (p = 0.698). The level of comedy decreases with its press rating (b = -0.378, R^2 = 0.002, , see **Figure 3.d**), but the result is not significant neither (p = 0.359). However, the level of comedy decreases with its audience rating and the result is significant (b = -0.560, $p < 0.05$, R^2 = 0.009, see **Figure 3.c**), and also with the level of comedy given by the label (b = 0.172, $p < 0.001$, R^2 = 0.415, see **Figure 3.e**) significantly with audience ratings and press ratings (b = -4.690, $p < 0.001$, R^2 = 0.014).

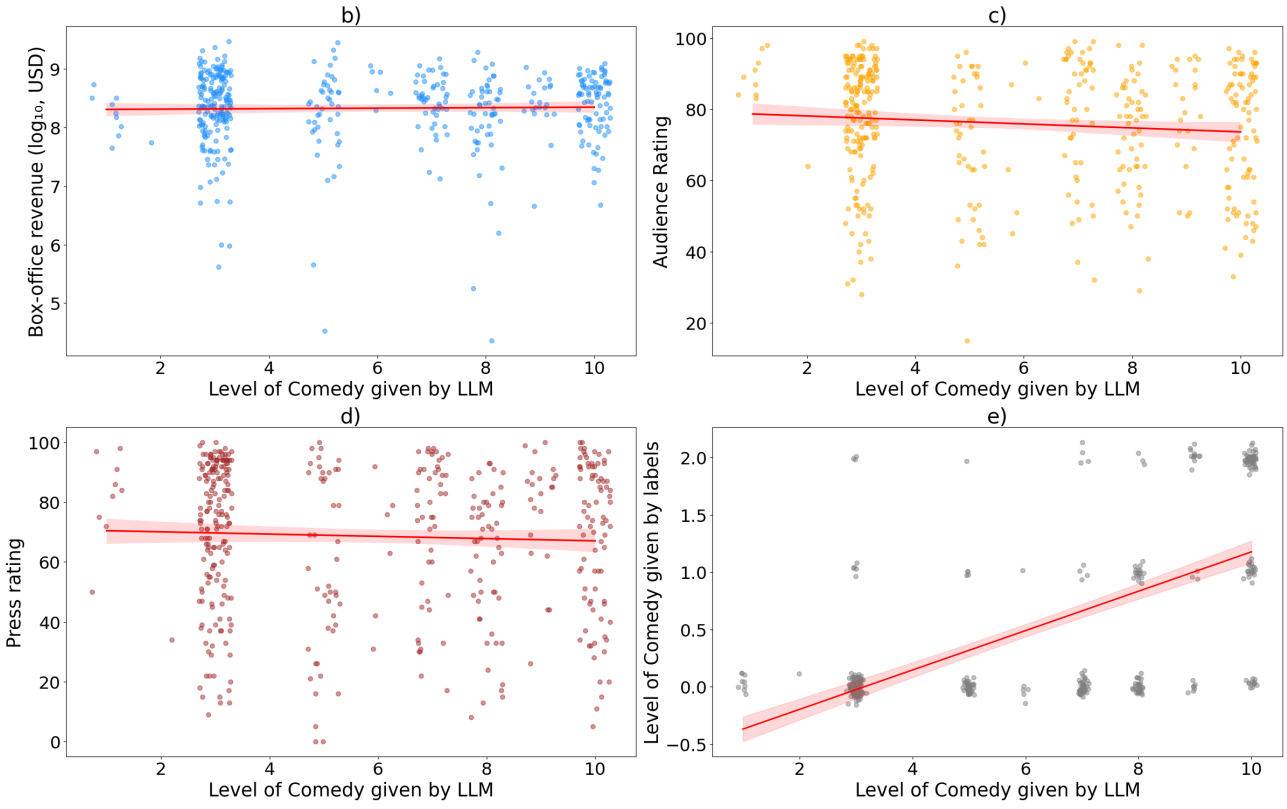


Figure 3 – Repartition of b) the box-office revenue (log_{10}, USD), c) the audience rating (from 0 to 100), and d) the press rating (from 0 to 100) and e) all depending on the level of comedy given by LLM (with 0: No Comedy, 1: Comedy mixed with another genre, 2 : Pure comedy)

Conclusion

Overall, the regression indicates that the level of comedy is not significantly associated with box-office revenue or press ratings, whereas it is significantly related to audience ratings and to the comedy level derived from genre labels, with comedy being lower for films with higher audience ratings and strongly aligned with the label-based measure.

The fact that LLM-based variables are overall less statistically significant can, in our view, be explained by three main factors.

First, a purely mechanical explanation cannot be ruled out. Given the substantially smaller dataset, statistical tests naturally have lower power, which may result in non-significant estimates. Importantly, although the coefficients fail to reach significance, their directions (positive or negative) are consistent with those observed in our previous label-based study, suggesting that the underlying relationships may still be present but insufficiently supported by the available data.

Second, it is possible that the LLM provides a more faithful representation of the intrinsic comedic content than genre-based labels. Under this interpretation, our previously validated hypothesis—that comedy positively influences box office performance and negatively influences press ratings—would need to be reconsidered, as these effects are no longer statistically supported. This would imply that the effects identified in the label-based analysis may partly reflect extrinsic factors associated with genre categorization rather than comedy as an inherent cinematic property.

Third, and this is the hypothesis we favor, the LLM may be less reliable than genre-based classification for this task. The weaknesses observed during the training phases do not allow us to rule out additional errors occurring at inference time. In particular, the well-documented tendency of LLMs to generate plausible but incorrect outputs, even when strong warnings are included in the prompt, remains a critical concern. Moreover, zero-shot LLM-based studies typically rely on very large models (Wang et al., 2023), whereas the models used here may not reach the same level of robustness.

In addition, it should be noted that the genre label displayed to audiences may itself influence viewer behavior. Genre information can shape expectations and guide film choice, meaning that labels may exert an indirect but substantial effect on box office outcomes. This mechanism could explain why the results observed with genre-based classifications are not recovered when relying solely on LLM-inferred attributes.

Finally, a potential source of bias in this study lies in the selection of the 500 most popular films. If this subset overrepresents certain genres or systematically includes films with a higher volume of reviews, the resulting estimates may be distorted. Such a selection bias could attenuate or amplify observed effects, particularly if popularity correlates with both genre and critical attention. Plus, we can visually see that the distribution of the 500 most popular films is quiet different than the original one (see **Figure 1 a**, **b**, **c** and **d**). Although we do not believe this bias directly drives the present results, it remains an important limitation that should be acknowledged.