## 560 Supplementary Materials
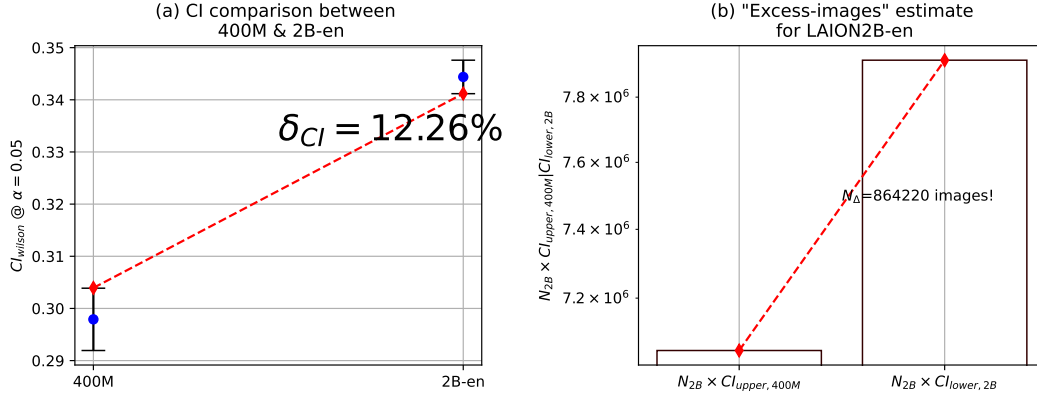
## 561 The risks of extrapolation



Figure 3: Binomial proportion confidence interval (CI) analysis to establish the extent of HCR underestimation upon using LAION400M statistics.

## 562 NSFW Analysis

| | NSFW >= 0.5 | | | NSFW <0.5 | | |
|---|---|---|---|---|---|---|
| | hateful | targeted | aggressive | hateful | targeted | aggressive |
| **mean** | 0.749 | 0.356 | 0.104 | 0.021 | 0.013 | 0.011 |
| **min** | 0.501 | 0.008 | 0.013 | 0.00 | 0.013 | 0.005 |
| **max** | 0.988 | 0.961 | 0.894 | 1.00 | 0.499 | 0.513 |

Table 3: The mean, minimum and maximum values for hateful, targeted and aggressive content in our sample of LAION 2B-En, divided according to the suggested NSFW threshold of 0.5.