For the “gpt-4-1106-preview” model prompts describing a scenario regarding the divorce issue led to guardrail messages. In the case of the survey prompts, a guardrail answer was given for the Spanish version and a “not a moral issue” answer for the English one. Although multiple runs were not made for this model, accounting for the fact that guardrail responses were very unlikely in the survey prompt types, it seems like this model is even more strict to not answer sensible questions than its predecessor "gpt-3.5-turbo".

With respect to the other 2 models assessed. The survey prompt type runs outputted just two guardrails for the "gpt-3.5-turbo" model and none for the “flan-t5-xl” model. Guardrail incidence for scenario prompts can be seen filtered by language and temperature in Figure 1 and its stability across runs in Figure 2.

Figure 1. Non-response by model, language and temperature for the scenario prompts.

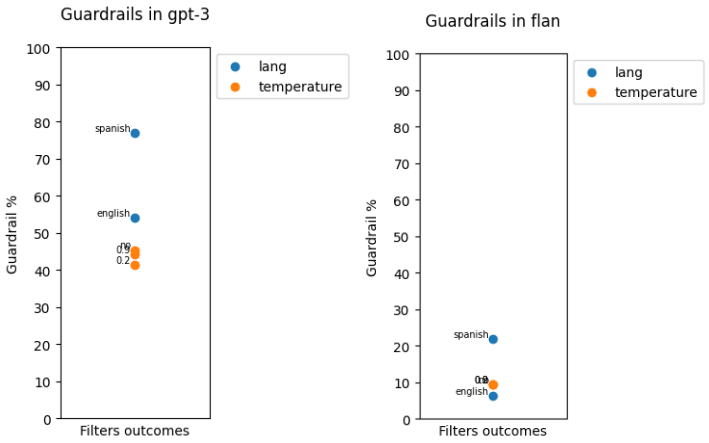
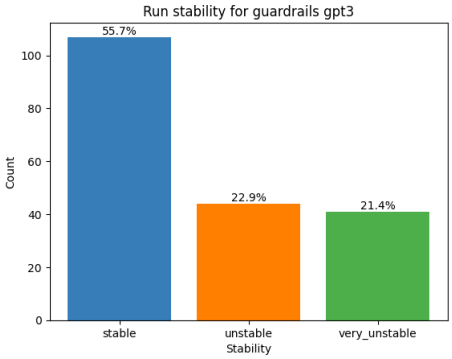


Figure 2. Stability of guardrails across runs for the scenario prompts.



Note. “stable”: all answers are guardrail or an answer. “unstable”: there is one disagreeing answer. “very unstable”: out of 5 answers there were 2 of one type and 3 of the other. All flan outcomes were stable.

In the agreement with the survey prompt statement, the gpt-3 model agreed (answered “yes”) just a 9% of the times. The flan model agreed with all the English prompts and answered it wasn’t a moral issue for the Spanish prompts having an agreement rate of 50%. Across runs gpt-3 model Fliess’s Kappa was 0.37, accounting for a low “fair” agreement. Further graphs accounting for language and temperature variations for gpt-3 agreement can be found in the notebook “assig_23_analysis_pesce”.

In the scenario prompt setting, excluding guardrails, the coincidence for both models with the country setting was very close to chance (54% and 49% for gpt-3 and flan respectively). Again excluding guardrails, answers according to the countries survey responses were more accurate for Spanish prompts as compared to English ones for both models (Table 1). Regarding runs stability, the flan model got again a Fliess’s Kappa of 1 and the gpt-3 model one of 0.5, achieving greater agreement (moderate) than in the survey scenario.

Table 1. Advice agreement with country answer in scenario prompt setting.

	gpt3.5-turbo	flan-t5-xl
English	108 (49.09%)	210 (46.67%)
Spanish	71 (63.96%)	195 (52.0%)