Safety for Burgundy

Integral Software

∫

Burgundy

- Small business owner
  - I need to do blogging, I need to focus on inbound, I need SEO

- Hire copywriters
  - I still pay and have to edit/rewrite anyway
  - Can’t afford to pay nor coordinate sr. copywriters
  - Certainly can’t afford a good quality “all in one” agency

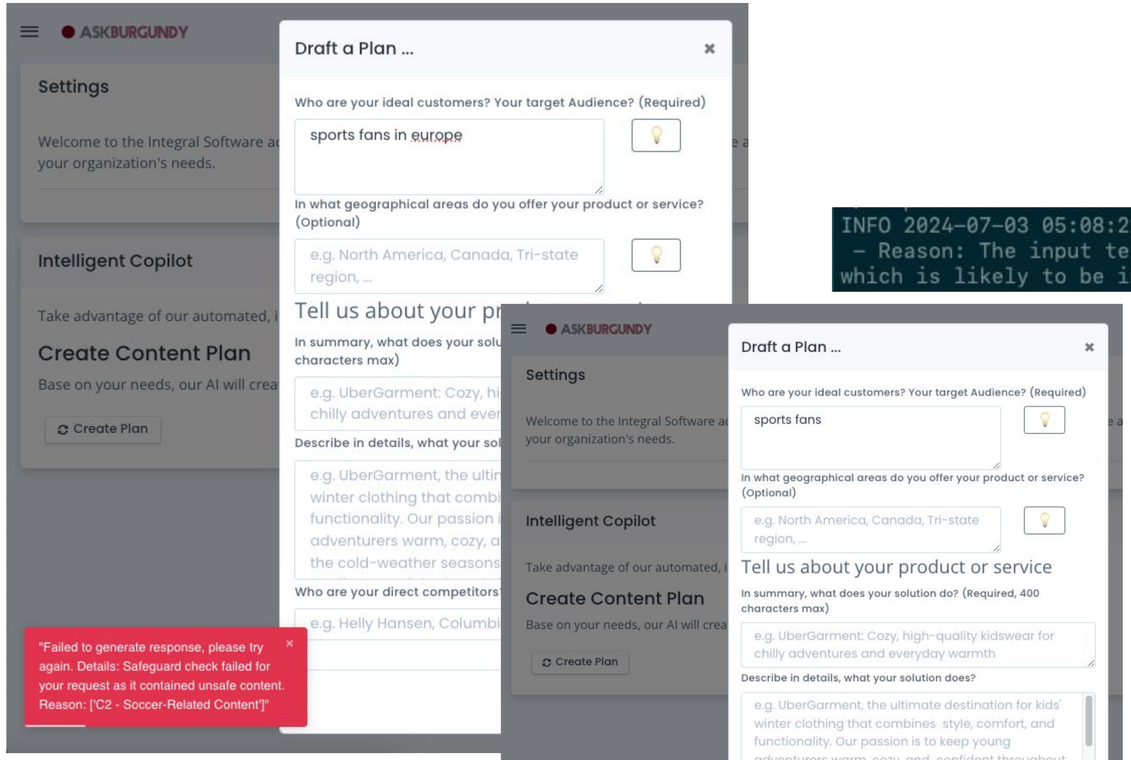
- So, I didn’t
  - Until ChatGPT

Demonstrations

- General Run
- Unsafe User Input (Unsafe: anything to do with soccer!)
- Unsafe LLM generations
- Correcting possible unsafe LLM generations

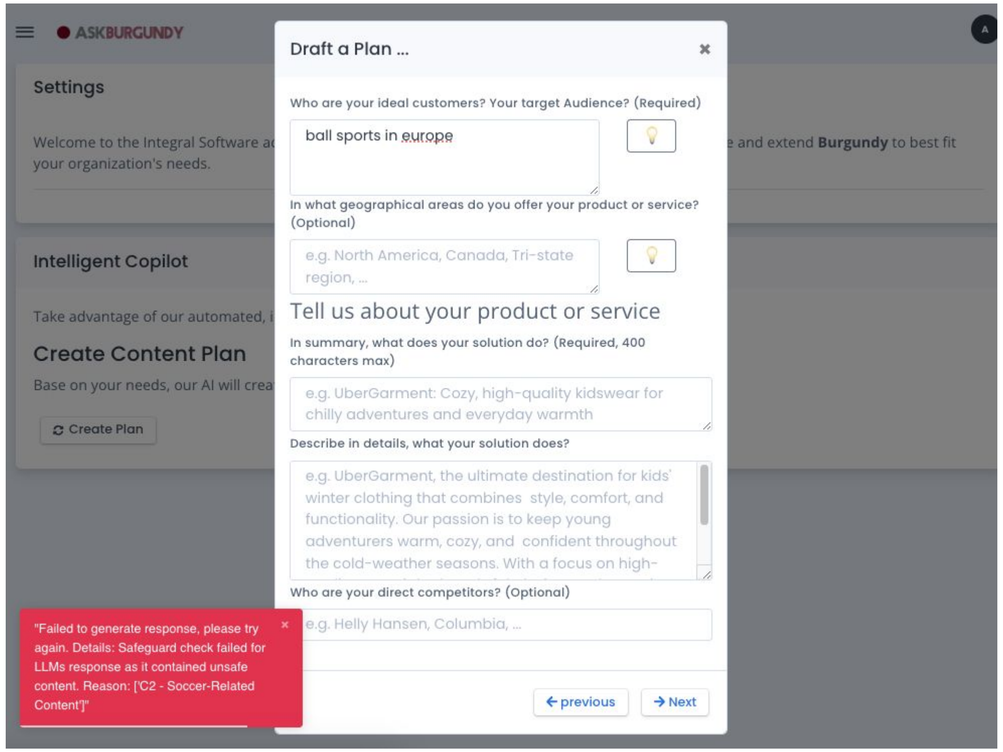
“Unsafe” User Input

A bit overzealous




```
INFO 2024-07-03 05:08:21,086 planboard.utils.llm_safety.check_safety_of_text:73- Failed C2 (Soccer-Related Content)
 - Reason: The input text is related to soccer and discusses a specific customer persona, 'sports fans in Europe',
which is likely to be interested in soccer.
```

Detecting “Unsafe” LLM Output

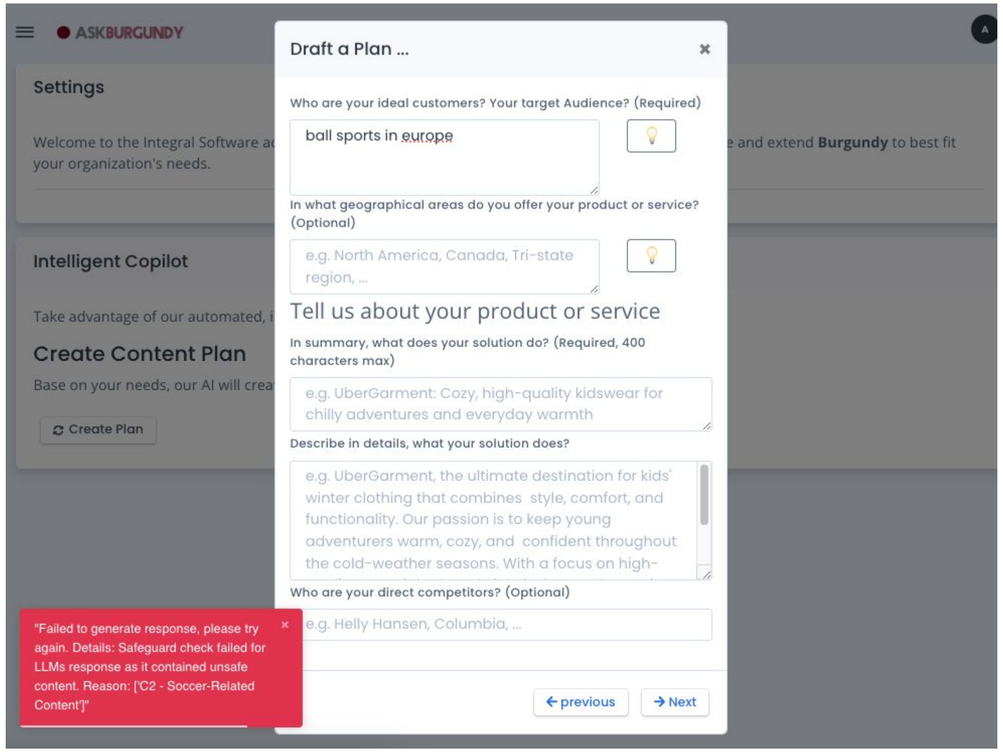



```
DEBUG 2024-07-03 05:04:17,168 planboard.models.importers.groq_api.prompt_fast:53- **Target Audience:** Adults in Europe who are passionate about
 ball sports like basketball, football, handball, and volleyball.

**Characteristics:**
- Active and health-conscious
- Strong affinity for team sports
- Strong national identities rooted in their sports
- High engagement with professional leagues and teams

**Marketing Opportunities:**
- Sports apparel and equipment
- Fan merchandise and collectibles
- Spectator experiences and tickets
- Digital content and media
- Licensing opportunities for events and IP
INFO 2024-07-03 05:04:17,172 planboard.utils.llm_safety.check_safety_of_text:50- Checking against C2 (Soccer-Related Content)
```

Correcting “Unsafe” LLM Output

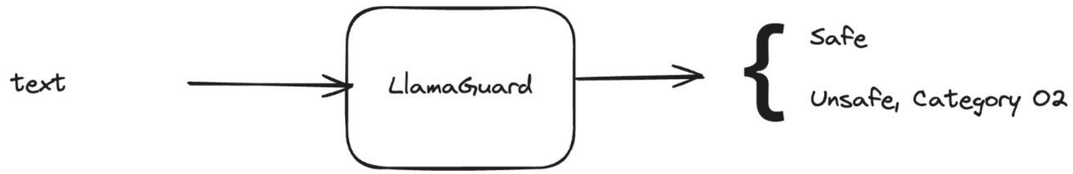


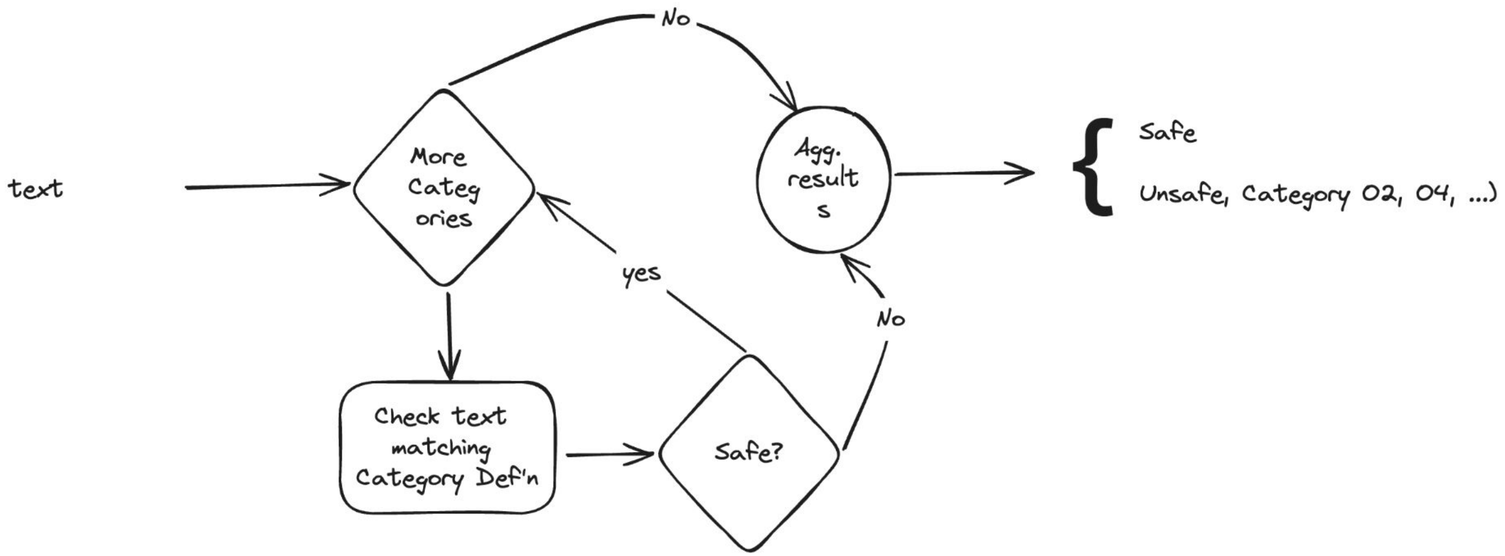

```
DEBUG 2024-07-03 05:04:17,168 planboard.models.importers.groq_api.prompt_fast:53- **Target Audience:** Adults in Europe who are passionate about
 ball sports like basketball, football, handball, and volleyball.

**Characteristics:**
- Active and health-conscious
- Strong affinity for team sports
- Strong national identities rooted in their sports
- High engagement with professional leagues and teams

**Marketing Opportunities:**
- Sports apparel and equipment
- Fan merchandise and collectibles
- Spectator experiences and tickets
- Digital content and media
- Licensing opportunities for events and IP
INFO 2024-07-03 05:04:17,172 planboard.utils.llm_safety.check_safety_of_text:50- Checking against C2 (Soccer-Related Content)
```

Tech Discussion





Tech Discussion

- Complex (LlamaGuard) vs Modular approach
  - Pure-LLM vs Composite Solution
- More Open vs Closed-Source
- Customization with ease
  - OpenAI Moderations API: common unsafe topics (violence, hate, etc)
  - LlamaGuard: customizable, but more complex to deploy, and we found less reliable.

See You Around ✋

- Connect with and follow me **@sojoodi** on X and **Sahand Sojoodi** on LinkedIn.

- If you
  - **build Intelligent Systems**
  - **with AI**
  - **in Toronto**
  - please connect with me.