



# 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

The screenshot shows the 'Draft a Plan' form in the ASKBURGUNDY application. The 'Who are your ideal customers?' field contains the text 'sports fans in europe'. A red error message is displayed at the bottom left of the form, stating: "Failed to generate response, please try again. Details: Safeguard check failed for your request as it contained unsafe content. Reason: ['C2 - Soccer-Related Content']".

```
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.
```

This screenshot shows the same 'Draft a Plan' form, but with the 'Who are your ideal customers?' field containing the text 'sports fans'. A red error message is displayed at the bottom left, identical to the one in the first screenshot: "Failed to generate response, please try again. Details: Safeguard check failed for your request as it contained unsafe content. Reason: ['C2 - Soccer-Related Content']".

# Detecting “Unsafe” LLM Output

The screenshot shows the 'Draft a Plan' form in the ASKBURGUNDY application. The form includes several input fields: 'Who are your ideal customers? Your target Audience? (Required)' with the text 'ball sports in europe'; 'In what geographical areas do you offer your product or service? (Optional)' with the text 'e.g. North America, Canada, Tri-state region, ...'; 'Tell us about your product or service' with a sub-question 'In summary, what does your solution do? (Required, 400 characters max)' and the text 'e.g. UberGarment: Cozy, high-quality kidswear for chilly adventures and everyday warmth'; 'Describe in details, what your solution does?' with the text 'e.g. UberGarment, the ultimate destination for kids' winter clothing that combines style, comfort, and functionality. Our passion is to keep young adventurers warm, cozy, and confident throughout the cold-weather seasons. With a focus on high-'; and 'Who are your direct competitors? (Optional)' with the text 'e.g. Helly Hansen, Columbia, ...'. A red error message at the bottom left states: 'Failed to generate response, please try again. Details: Safeguard check failed for LLMs response as it contained unsafe content. Reason: [C2 - Soccer-Related Content]'. Navigation buttons for 'previous' and 'Next' are visible at the bottom of the form.

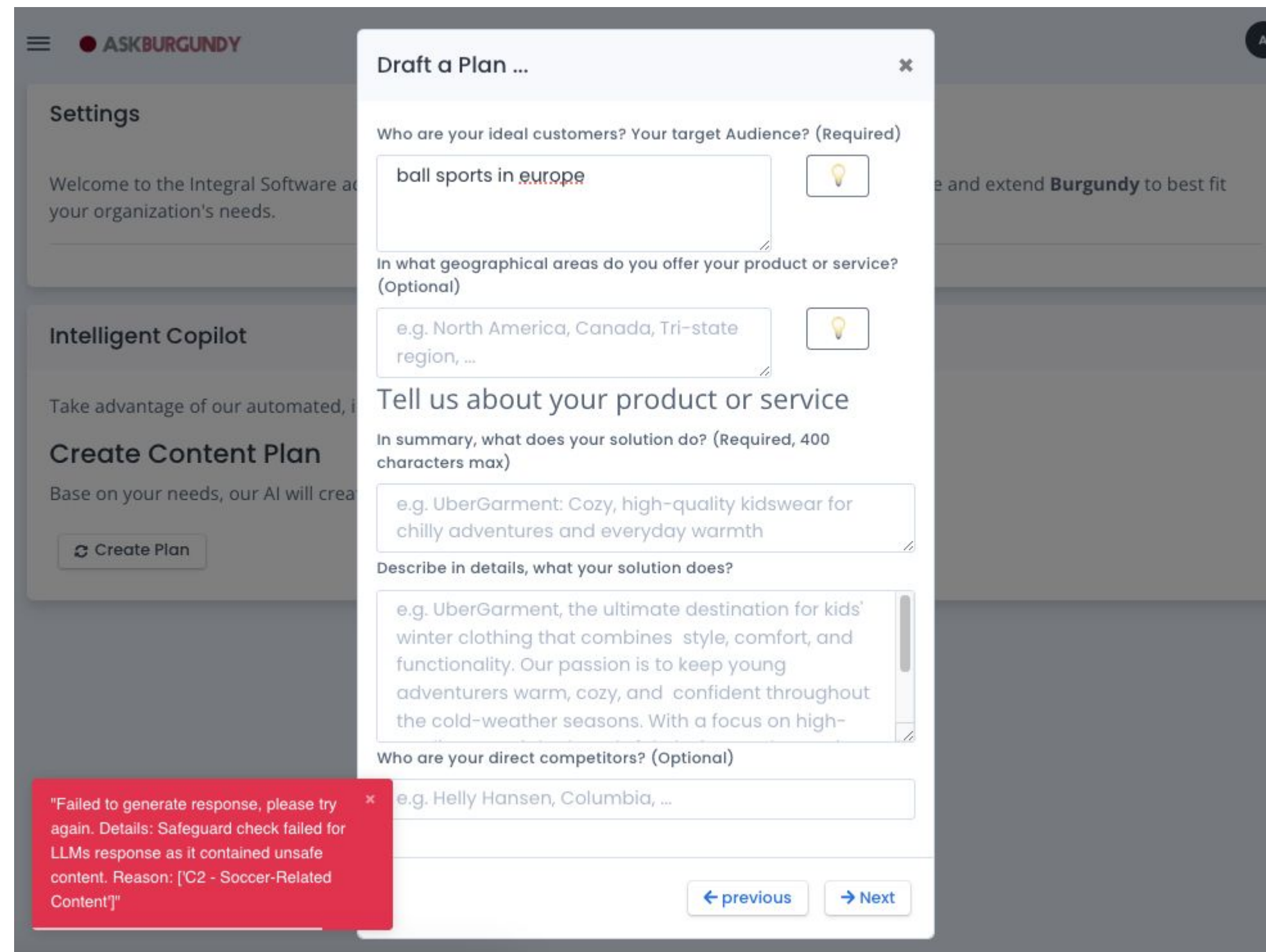
```

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



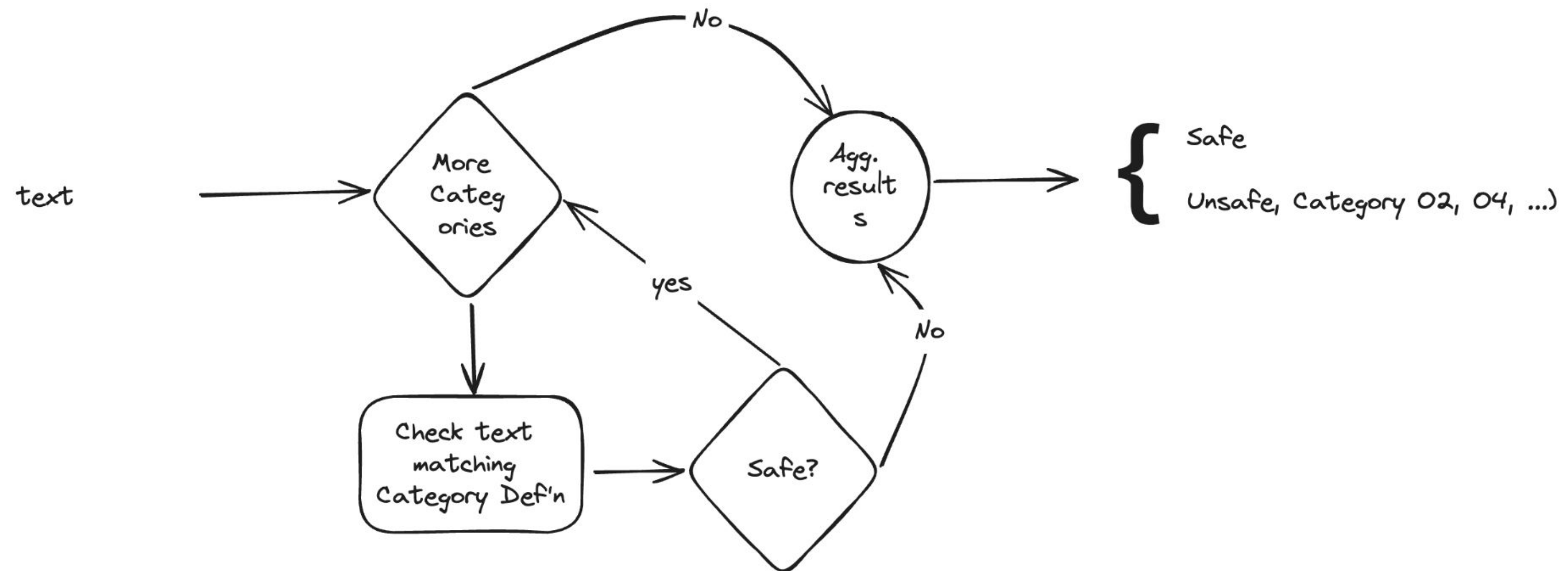
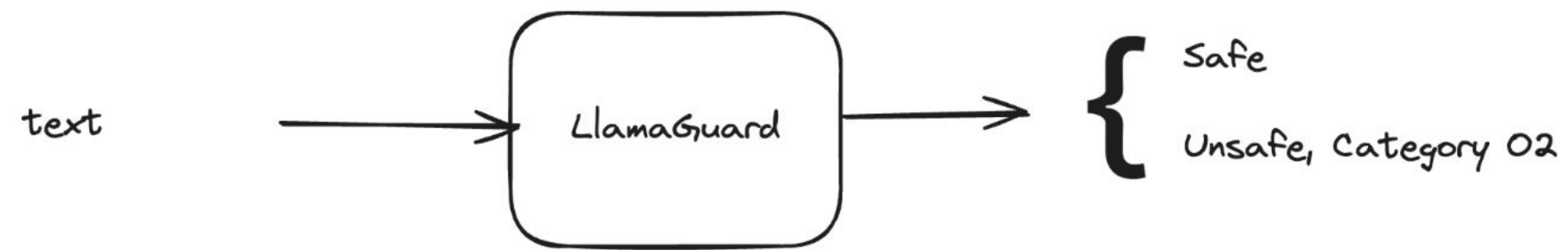
```
DEBUG 2024-07-03 05:04:17,168 planboard.models.importers.groq_api.prompt_fast:53- **Target Audience:** Adults in Europe who are passionate about ball sports like basketball, football, handball, and volleyball.

**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.