Factuality GPT Classification Prompts

Gili Goldin Shira Wigderson Ella Rabinovich Shuly Wintner

July 24, 2025

GPT Classification Prompts

This section presents the prompts used to classify sentences for check-worthiness using GPT-based models.

The *User Prompt* used in all experiments is as follows:

> *Classify the following sentence into one of the following categories: 'worth checking', 'not worth checking', or 'not a factual proposition'. Respond with only the label. Sentence:* {`sentence_input`}

The *System Prompts* for each prompting technique are described below:

Zero-Shot Prompt

> *You are a helpful assistant that strictly classifies sentences into 'worth checking', 'not worth checking', or 'not a factual proposition'. Always respond with only the label and nothing else.*

Instruction Based Prompt

> *You are a helpful assistant that strictly classifies sentences into 'worth checking', 'not worth checking', or 'not a factual proposition'. Always respond with only the label and nothing else. Follow the definitions provided strictly and always respond with only the label. Here is what each category means: 1. 'worth checking' - Sentences that include claims or propositions that can be factually verified or debunked. For example: 'The Earth is flat.' 2. 'not worth checking' - Sentences that include obvious truths or widely accepted facts, or subjective opinions or claims that cannot be verified or are not important to check. For example: 'The sun rises in the east.' or 'Chocolate ice cream is the best dessert.' 3. 'not a factual proposition' - Sentences that do not propose a factual claim. This includes questions, commands, or exclamations. For example: 'Do you think this is true?' or 'Please close the door.'*

Few-Shot Prompt These examples were originally in Hebrew, but are presented here in English (translated to English using gpt-4o).

> *You are a helpful assistant that strictly classifies sentences into 'worth checking', 'not worth checking', or 'not a factual proposition'. Always respond with only the label and nothing else. Follow the definitions provided strictly and always respond*

with only the label. Here is what each category means: 1. 'worth checking' - Sentences that include claims or propositions that can be factually verified or debunked. Examples:

'- Additionally, during the Eighteenth Knesset, the Ministry of Justice promoted reforms in various legislative areas: criminal law, security, civil law, economic-fiscal law, administrative law, and international law.'

'- The budget is the best way for a government to express its vision, priorities, and its approach to shaping society.'

'- Seventy members of my family were murdered for the sanctification of God's name in that horrific place called Auschwitz-Birkenau.'

'- The State of Israel is now 66 years old, I believe.'

2. 'not worth checking' - Sentences that include obvious truths or widely accepted facts, or subjective opinions or claims that cannot be verified or are not important to check. Examples:

'- I am saying here that we have a shared responsibility with the Ministry of Finance to advance this.'

' - As the Chairperson of the Public Petitions Committee, you surely know how burdensome it is for the public when they feel they are paying different amounts for the same service.'

'-I don't even want to imagine.'

' - He is afraid of everything.'

3. 'not a factual proposition' - Sentences that do not propose a factual claim. This includes questions, commands, or exclamations. Examples:

'- To you, members of the Nineteenth Knesset, I will recommend and request: focus on significant legislative proposals, substantive issues, and comprehensive, important reforms.

'- Where is the Minister of Finance to come up and respond?'

'- Thank you.'

'- So do me a favor, stop with this whole 'he's a handsome guy, a nice guy,' and that he provided cellular security for people in Israel.'