//

🐞 Gemini   ジェミニ

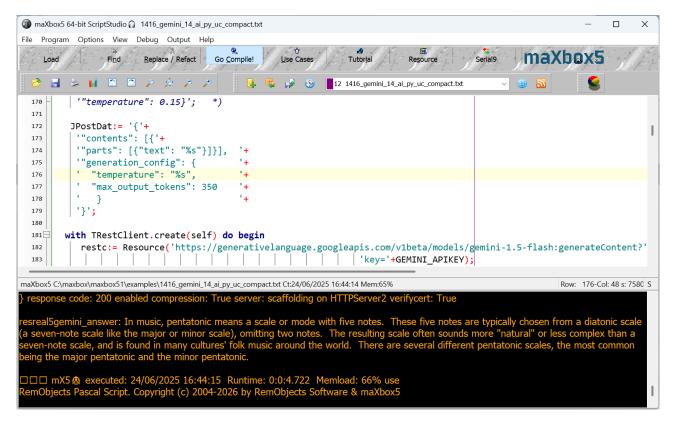maXbox Starter 148 – Get a Language-model.

"Non ridere, non lugere, neque detestari, sed intelligere.[1] –
Spinoza.

Source: **1416_gemini_14_ai_py_uc_compact.txt**

https://sourceforge.net/projects/maxbox5/files/examples/
1416_gemini_14_ai_py_uc_compact.txt/download

Google Gemini is a family of large language models (LLMs) offering state-
of-the-art AI capabilities created by Google AI.

The answer generated by Gemini we get trough an API call.

1416_Gemini2025-06-24_164737.png

In the following we use the flash model. Gemini 1.5 Flash is Google's
freely available Gemini product, and it was built for speed, efficiency,
and cost-effectiveness while still maintaining a high level of
performance and reasoning. We can also use 2.0 Flash with a structured
answer, for example the question what is diatonic in music:
Diatonic refers to a musical scale or mode with seven notes, where the

---

1 (ethik. 9,13) - Not to laugh, not to cry, not to hate, but to understand. -

intervals between the notes follow a specific pattern.
  - This pattern typically consists of two whole steps, a half step,
    three whole steps, and a half step.
    ◦ Notation: 2212221
  - It's the most common and familiar type of scale in Western music,
    forming the basis for most major and minor keys.

The current implementation is my third approach to the challenges I have
faced during several years of dealing with a lot of unexpected scenarios
in the field of unexpected REST-Api error like that:

```
@addr:TRestResource@5A19F070
debug: 60-No argument for format '{"contents": [{"parts": [{"text' 876 err:20
EWI_Exc:Exception: No argument for format '{"contents": [{"parts": [{"text'
resreal5:
```

That means in the second parameter was no value passed as JSON format:

```
JPostDat:= '{'+
    '"contents": [{'+
    '"parts": [{"text": "%s"}]}],  '+
    '"generation_config": {       '+
    '  "temperature": "%s",       '+
    '  "max_output_tokens": 350   '+
    '  }                          '+
    '}';
```

The second parameter in that POST call is the temperature, in the context
of large language models (LLMs) like those used in "Flash" attention or
other "Flash" algorithms, "temperature" refers to a hyperparameter that
controls the randomness of the model's output.
It's a crucial setting that influences the creativity and diversity of
the generated text or answer.
A higher temperature makes the probability distribution flatter. This
means less likely words have a higher chance of being selected.

```
function TRestClient4_AskChatGPTGemini(askstream, tempera: string;
                          aResponseHeader:TRestResponseHandler):string;
var JPostdat: string; jo: TJSON;
    restc: TRestResource; stream: TStringStream; jArray:TJsonArray2;
begin
with TRestClient.create(self) do begin
    restc:= Resource('https://generativelanguage.googleapis.com/v1beta/models/
                      gemini-1.5-flash:generateContent?'+'key='+GEMINI_APIKEY);
    println('@addr:'+objtostr(restc))
    restc.ContentType('application/json');
    ConnectionType:= hctWinInet;
    OnResponse:= @TRestOnResponseEvent2;
    try
      jo:= TJSON.Create();
      stream:= TStringStream.create('');
      stream.writestring(format(JPostDat, [askstream,tempera]));
      jo.parse(restc.Post(stream))
      jArray:= jo.values['candidates'].asarray[0].asobject['content'].
                        asobject['parts'].asarray;
      result:= JArray[0].asobject['text'].asstring;
    finally
      Free;  //restclient
```

```pascal
    jo.Free;
    stream.free;
  except
    writeln('EWI_Except:'+ExceptiontoString(exceptiontype,exceptionparam));
  end;
  end; //with
end;
```

The model tends to generate more common, predictable, and safe responses.
It sticks closer to the training data.

```pascal
{@To get access weather API you need own API key whatever account you chose!!}
https://home.openweathermap.org/api_keys
 {
  "message": "Not Authorized",
  "error_detail": "Direct access not allowed"
  }
```

```pascal
Const AGEMINI_APIKEY = 'AIzaSyA_use/set your own key first_____';
```

Gemini 1.5/2 Pro is Google's paid version of Gemini, and its API version
has a context window of 2 million tokens, compared to Flash's 1 million.

Note: when no internet is available you get:
Dec: without internet fail safe
debug: 332- 4294967295 err:0; debug: 333-Socket Error # 11001
Host not found. 856 err:20

The same goes also in Python:

```pascal
    execstr('os.environ[''GOOGLE_API_KEY'']='
                        +loadfile3(exepath+'geminiapikey.txt')+'');
     //execstr('os.environ[''GOOGLE_API_KEY'']=
                        "'+loadfile3(exepath+'geminiapikey2.txt')+'"');
    execstr('genai.configure(api_key = os.environ[''GOOGLE_API_KEY''])');
    //execstr('model = genai.GenerativeModel(''gemini-pro'')');
    execstr('model = genai.GenerativeModel(''gemini-1.5-flash'')');
    execstr('response = model.generate_content("Why is the earth blue?")');
    execstr('print("Pythonback:",response.text) ');
    //execstr('(read_http_input()) ');
    println(stringreplace(evalstr('output.getvalue().strip()'),
                        #10,#13#10,[rfReplaceAll, rfIgnoreCase]));
```

Conclusion

Multimodal: As always with Gemini models, 1.5 Flash can natively process
multiple data types, including text, images, audio, and video, within a
single conversation and/or prompt.

- Fast: The closest model to Gemini 1.5 Flash in terms of speed is
  GPT-4o Mini, which has a mere 99 output tokens per second compared
  to 1.5 Flash's 194, more than double.
- Large context window: While the publicly available Gemini 1.5 Flash
  only has a 32k context window, the API version boasts a context
  window of 1M tokens; that's equivalent to an entire repository of
  code or a multi-hour lecture.
- Cheap: The Gemini 1.5 Flash API is priced at 0.1 USD per 1M tokens,
  which is three times cheaper than the second cheapest, GPT-4o Mini.

Gemini 1.5 Flash stands out as a highly efficient and cost-effective model from Google, designed to balance speed, performance, and affordability. While it doesn't lead in reasoning benchmarks, its multimodal capabilities, speed, cheap pricing, and context window make it a robust model for everyday tasks.
The temperature parameter is applied to the probability distribution of the next predicted token. Before the model selects the most likely next word, the logits (the raw scores output by the model) are divided by the temperature value.
Gemini Ultra — The largest and most powerful model excels in complex tasks like coding, logical reasoning, and creative collaboration. Available through Gemini Advanced (formerly Bard).

Script:
https://sourceforge.net/projects/maxbox5/files/examples/
1416_gemini_14_ai_py_uc_compact.txt/download

References:
Google Gemini API – Code Blog

Gemini 1.5 Flash

Google Gemini updates: Flash 1.5, Gemma 2 and Project Astra

Maps | Mapbox

🐞言語モデル.

Doc and Tool: maXbox5 - Manage Files at SourceForge.net

Max Kleiner 24/06/2025