Investigating the Use of Large Language Models for Alternative Word Substitution for Dyslexia

Laura Hoogenraad (s3165965)

Supervisors:
Dr. ir. D.J. Broekens & Dr. M.T. van der Meer

Bachelor Thesis– Informatica

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl July 14, 2026

Abstract

Dyslexia is a learning disability which affects a significant portion of the population. Dyslexic individuals have difficulty reading and writing accurately and fluently. This thesis investigated whether a Large Language Model (LLM) can be used in creating a compensatory dyslexia tool. The thesis investigated the effectiveness of a LLM in identifying specific linguistic features and if it is able to recommend alternatives to these features that are more "dyslexia-friendly". We determined the precision and recall of two different OpenAI models and tested two different system messages.

Our findings show that a LLM is not effective at making a text more "dyslexia-friendly". The tasks investigated in this thesis indicate that a LLM is too inconsistent and thus not effective at finding and replacing lexical features associated with dyslexia.

Contents

1 Introduction

Dyslexia is a learning disability that effects a persons ability to read and write. Dyslexia is believed to affect about 7% of the population, though this estimate is dependent on how one defines dyslexia.[PP12]. Dyslexia can negatively influence a persons ability to learn as they struggle to read and write properly. Fortunately, schools are able to support their students in a number of different ways. Whilst there is support that can help students improve their linguistic abilities, schools also offer a number of tools that are more focussed on minimising the limitations that dyslexic students might face.[Cen25a] So-called "dyslexia software", such as *Kurzweil 3000* or*Alinea Online* are commonly used. They, for example, offer support in the form of text-to-speech tools or spellcheckers.

In recent years the use of AI has become more prominent. However, a study by Barua et.al[BVG+22] which looked into the range and effectiveness of AI-assisted tools developed for children with specific learning difficulties, found that whilst AI tools can be used in an effective way in education there are disproportionately less AI-based tools designed for dyslexic students.

This thesis aims to investigate whether a Large Language can be utilised in developing a compensatory tool aimed specifically at dyslexic individuals. The development of a potential AI powered compensatory tool could help both dyslexic students and dyslexic people in general with reading and understanding written content. This thesis investigates if a LLM is able to identify specific lexical features in a text and whether it is then able to provide adequate substitutes to these features.

2 Theoretical frame and related work

2.1 What is Dyslexia

Stichting Dyslexie Nederland, an organisation who's goals it is to share scientific knowledge on the topic of dyslexia with professionals in education and care, defines dyslexia as a learning disability characterized by a persistent difficulty with reading and/or spelling accurately and fluently. These learning difficulties however, can not be attributed to any environmental factors and/or physical, neurological, or general intellectual disability. [SDN16] Many theories on the etiology of dyslexia exist and there seems to be no real consensus on what causes dyslexia. In this paragraph we will address three of the leading theories of developmental dyslexia.

2.1.1 The phonological theory

The phonological theory argues that dyslexics have poor reading skills because they failed to develop the necessary phonological awareness.[Ste23] The ability to perceive and manipulate linguistic sounds is essential to automating letter–sound correspondences, enabling accurate and fluent word recognition through phonological coding. [PP12] However, the theory has also been criticised, as it fails to provide an actual explanation as to the cause of dyslexia. [Ste23] Besides this, the theory is also a circular argument. Poor phonological skills can result in poor reading ability; conversely, limited reading ability may also contribute to impaired phonological skills. [Ste23][PP12]

2.1.2 The cerebellar theory

The cerebellar theory suggests that the problems which dyslexic people face can be attributed to an impaired cerebellum. An article by Nicolson et al [NFD01] hypothesises a causal chain in which cerebellar impairment can lead to phonological difficulties which can subsequently leads to a number of difficulties which dyslexic individuals struggle with e.g. reading, writing, and spelling. However, a study by Barth et al.[BDS+10] found little evidence to support the idea that cerebellar dysfunction is a major cause of dyslexia. Another study by Ashburn et al.[AFNE20] used functional magnetic resonance imaging (fMRI) to test the cerebellar theory and found no cerebellar activity specific to reading in both children with and without dyslexia.

2.1.3 The magnocellular theory

Another theory as to the cause of developmental dyslexia is the *magnocellular theory*. This theory proposes that people with dyslexia suffer from poor visual localization, due to impairments to the so called magnocellular system. In dyslexic individuals this system, which influences motion sensitivity, does not function accordingly. This in turn, could result in poor visual location which, when reading, can make is appear as if letters are moving around and overlapping. [Ste01] Besides having difficulty with reading this theory also suggests that people with developmental dyslexia struggle with auditory processing due to the underdevelopment of the magnocellular system, which would explain their issues with phonology.[Ste01].

2.2 Dyslexia support

Schools can offer a number of support to help students with dyslexia with their education. *Dyslexie Centraal*, a digital platform on which parents, students, and professionals can find information on dyslexia in education and healthcare, makes the following distinction between different types of support options: *stimulation*, *remediation*, *compensation*, and *dispensation*. [Cen25a]

Stimulation refers to support which is focussed on stimulating students to read. This can be accomplished through different means. For example by reading out loud to students, by making sure that there is a wide variety of books available to them, or by letting students choose themselves what book they want to read. [Cen25b]

Remediation focusses on providing task specific support. Improving skills such as technical reading and spelling. Dyslexic students, for example, might struggle with fully understanding a text as they are more focus on reading a text correctly. Thus, teaching students the appropriate reading strategies needed to understand a text is especially important for them in order to achieve text comprehension.[Cen]

Compensation are measures which aim to minimise the limitations which dyslexic students face. This can be measures such as extra time on tests, an increased font size, or software which can make reading easier and support writing. [dh]

Dispensation refers to measures which exempt students from certain subjects and/or course activities all together.

In this thesis we primarily focus on compensatory tool. In the following sections we look at software and AI powered tools developed for dyslexic support.

2.3 Dyslexia Software

Most dyslexia specific software is text-to-speech based software which offer some of the following features:

- An OCR or optical character reader
- Dictation
- Spellchecker
- Word predictor
- Dual highlighting cursor
- Homophone function

When students have to spend too much of their cognitive energy on accurately decoding words, the actual comprehension of the text suffers. [RPM+05] Text-to-speech can significantly improve the reading comprehension of dyslexic students.[KCS+23] [KCSHK20] Examples of dyslexia software include: *Alinea Online*, *ClaroRead*, and *Kurzweil 3000*.

2.4 Dyslexia and AI

2.4.1 Personalised learning

AI technologies can be used to create personalised education tailored to a students personal needs. A scoping review on the use of artificial intelligence in dyslexia research and education[YAC25] found that, for example, AI is being used to adapt educational materials to better suit a students cognitive profile.

A study by Athanaselis et al. [ABD+14] looked into an assistive reading tool which adjusts the reading environment. It does so though highlighting text, splitting words into syllables, putting emphasis on certain letters, and through text-to-speech support. Using speech recognition and face analysis a personal profile is build based on a learners reading error and personal preferences. The environment is then adjusted based on a learners personal needs. Over time, the user profile is adjusted based on a learners improvement. The study found that the use of this assistive tool had a positive influence on the reading skills of dyslexic students and noted an increase in both reading pace and reading accuracy.

In 2019, *Dytective*, an digital tool for children with dyslexia, won the *UNESCO King Hamad Bin Isa Al-Khalifa Prize for the Use of ICT in Education*. The application is able to both screen children for dyslexia, as well as provide them with personalized game-based exercises. [UNE]

2.4.2 LLMs

A study by Zhao et al. [ZXP+25] introduces a strategy called *Let AI Read First* (LARF) which has a LLM process a text and annotate important and/or relevant text. LARF aims to increase the readers focus by annotating important content, through highlighting, bolding, and underlining, whilst preserving the original text. The study found that LARF significantly improved both the reading performance and reading experience of dyslexic individuals.

A study by Piferi et al.[PCV+25] proposes a modular framework, ChatCare, combining LLMs and real-time emotion detection techniques construct a more personalized and interactive platform focused on dyslexic children. On the ChatCare platform children can work on exercises and get personalised tutoring from a conversational agent. The goal is to facilitate an environment in which children feel motivated and engaged with the educational material.

2.4.3 Browser extensions

Helperbird is a browser extension designed to make browsing more accessible to people with learning disabilities such as dyslexia, and offers number of features that can aid people reading. *Helperbird* can for example adjust the letter spacing, word spacing, and font size. [Hel25a]. Helperbird also offers a feature which can rewrite and simplify text using AI.[Hel25b]

2.4.4 Spellchecker

A study by Rello et al.[RBB15] designed a spellchecker specifically for people with dyslexia. The *Real Check* spellchecker presented in this study, focussed more on so called *real-word errors*. Real-word errors are spelling mistakes that accidentally end up resulting in another existing word. For example, the misspelling of *from* can lead to *form*. They found that the *Real Check* was able to recognise more errors in texts written by dyslexic people than other spellcheckers. People who used the spellchecker as they were writing text were able to construct sentences with more accuracy and in less time.

2.5 AI and Dyslexia Detection

Machine learning has been used to detect dyslexia in people. Early detection of dyslexia is essential as early intervention can improve reading and writing skills of dyslexic students.

A study by Robaa et al. [RBA+24] proposes the use of artificial intelligence to detect dyslexia through analysing handwriting. The study looked at datasets which included both uppercase and lowercase samples of dyslexic handwriting and defined three distinct categories: *normal*, *reversed*, and *corrected* letter. Normal letters refer to letters in its standard orientation. Reversed letters include letters which have been flipped, mirrored, or otherwise transformed to no longer fit the writing conventions. Corrected letter are letters that started out as reversed letters but during the writing process where adjusted to fit the standard writing conventions. They have found significant success with their explainable artificial intelligence framework which was able to achieve a test precision of 99.65%

Another study used machine learning to identify dyslexia in people by tracking their eye movement using a dataset consisting of the eye movements of 187 children (aged 9 to 10 years) of which 97 were dyslexic and 88 were not dyslexic.[PB20] The study was able to build a set of features related to eye movement which contributed most to the prediction of dyslexia. With this they were able to proposes a method that was able to distinguish between children with and without dyslexia with 95.6% accuracy.

2.6 Prompt Engineering

A large language model (LLM) is a computational model designed to predict what token should come next based on a textual input. Prompt engineering describes the process of crafting prompts aimed to guide LLMs into a certain direction. Determining the right prompt, which involves adjusting features such as length, writing style and structure, can all shape the output of a LLM.[Boo25]

2.6.1 Prompt Engineering Techniques

Prompt engineering techniques refer to methods used to guide LLMs into generating desired and relevant output, through prompting.[Gad25] There exists quite a large number of known prompting techniques. A systematic survey by Sahoo et al.[SSS+24] created a structured overview of the most recent advancements in the field of prompt engineering and was able compile a list of 41 different techniques. The two basic prompting techniques include *zero-shot prompting* and *few-shot prompting*.

Zero-shot prompting is a method which relies on a LLMs build-in pre-trained knowledge to perform a specific request. When prompting the model, one does not provide any example output and instead expects the model to be able to infer the desired input from the query alone, meaning that the performance is entirely dependant on the clarity of the prompt itself.[DSR+25]

Unlike zero-shot prompting, few-shot prompting does include examples in its prompt design. By including examples in the prompting itself, this methods aims to guide the model into a specific direction. The aim is to create a prompt that will generate output that is more accurate and relevant to the task.[DSR+25]

2.7 Parameters

Next to finding the right prompt the configuration of a LLM also greatly influences the output which it generates. For this thesis we will be focussing on the OpenAI Chat Completions API. The Chat Completion API allows you to adjust a number of these configuration allowing you to tailor the behaviour of the model. A number of these features include:

- **Model**
  The model parameter allows you to specify what model of GPT you want to use.[Yan24]
- **Messages** With the message parameter you can simulates a conversation with the model. Each message in this parameter consists of a *role* and *content*.[Yan24] The roles include *system*, *user*, and *assistant*:[Cha24]

  - **System**
    The system role provides the model with context and behaviour specific instructions.
  - **User**
    The user role represents the questions and tasks coming from the user.
  - **Assistant**
    The assistant role provides context by including output previously generated by the model itself.

- **Temperature**
  The temperature determines the degree of randomization that the LLM uses to generate an output. *Low temperatures* lead to a more deterministic output, whereas *high temperatures* will lead to more diverse and unpredictable output.[Boo25]

- **Top-p**
  The Top-P parameter is a sampling setting which, given a threshold p, selects the top tokens whose combined probability is not greater than p.[Boo25]

3 Research Question

This thesis aims to investigate how AI, specifically a Large Language Model (LLM), could be used to create a compensatory tool for dyslexic individuals.

The main research question of this thesis is:

Can a LLM effectively adjust a text to make it more dyslexia friendly?

To address this question the following 5 sub-questions were proposed:

Sub-question 1:
What lexical features do dyslexic people struggle with?

Sub-question 2:
Can an LLM be prompted to identify specific lexical features within a text?

Sub-question 3:
Can an LLM find appropriate alternative word substitutions to specific lexical features within a text?

Sub-question 4:
Will a newer, more advanced model perform better than an older lightweight model?

Sub-question 5:
What influence does a more detailed and task specific system message have on the performance of the model?

4 Methodology

To answer sub-question 1, a literature review was conducted, the results of which can be found in section 5.

Sub-question 2 was addressed using the findings from the literature review conducted for Sub-question 1. We focused on three specific lexical features: *low-frequency words*, *long words*, and *orthographic inconsistent words*. We opted not to investigate the two features *digits instead of words* and *percentages over fractions*. This is because finding a substitution for these features in a simple task which does not require a LLM. For example, the word *four* can easily be substituted for the number *4*. This change is not context specific, unlike the substitutions for the aforementioned three lexical features. An appropriate substitution for these features is one which does not alter the intended meaning of the text. This task is considerably more complex. With sub-question 3 we mean to investigate whether a LLM can be used to accomplish this task.

Experiment

To investigate how well the LLM is at identifying the three features we designed a *zero-shot* prompt for each of the features (Table 4.1). Besides using the zero-shot technique to prompt the LLM we also set the *temperature* setting to 0. In our testing of the model we found that the default temperature created inconsistent output, as can be observed in Tables A.1, A.3, A.5. With the temperature set to 0 we found the following results as can be seen in A.2, A.4, A.6. Therefore, when prompting the LLM, the temperature parameter was set to 0.

For each feature the LLM was prompted three times after which the intersection of the three outputs was taken to created one list of words identified by the LLM as matching a specific feature.

We tested the model on three different texts to ensure that the model was able to perform the task well for differing input. The texts that the LLM was tested on are three news articles on differing topics:

1. **Text 1** [VG26]
   An article by *Trouw* on the first international summit on stopping the use of oil, coals, and gas.

2. **Text 2** [Nie26]
   An article by *NRC* discusses the Artemis program and whether the missions are scientifically worthwhile.

3. **Text 3** [Sar26]
   An article by *NRC* on how lowering the blood pressure after a brain haemorrhage can significantly reduce the risk of a future stroke.

Feature	Prompt
Low-frequency word	"Low-frequency words are words that are rarely used in the Dutch language. Make a list of all the low-frequency words from the following text: `text`"
Long word	"A long word is a word which consists of N or more letters. Make a list of all the long words from the following text: `text`"
Orthographic inconsistent words	"Orthographic inconsistent words are words for which the pronunciation of the word is difficult to predict based on its spelling. Make a list of all the orthographic inconsistent words from the following text: `text`"

Table 4.1: Zero-shot prompts for identifying each lexical feature.

ID	System prompt
1	"You are a helpful assistant"
2	"You are used as a dyslexia tool. You need to adhere strictly to the instruction. And only output what is asked for. No extra text or comments, or questions."

Table 4.2: System messages for identifying each lexical feature.

To evaluate the response of the LLM we determined the *precision* and *recall* of the model.

To answer sub-question 4 we performed the experiment as described above for two different large language models. We used the OpenAI models *gpt-4o-mini* and *gpt-5.4-mini* . The gpt-4o-mini is an older model released in 2024 and was presented as a small and cost-efficient model.[Ope24] The gpt-5.4-mini was introduced in 2026 as model capable of taking on high-volume workloads.[Ope26]

Sub-question 5 was answered though repeating the experiment using two different system prompts as described in Table 4.2. The influence of the two system prompts was tested only on the gpt-4o-mini model.

Materials

In order to determine the precision and recall we had to create a ground truth for each of these three texts. These ground truth were establish as follows:

Low-frequency words ground truth

This ground truth was created through comparing the words in the text to a list consisting of the *9210* most frequent Dutch words.[FB22] Words in the text that are not included in this list were added to the *low-frequency words list*. We also removed any words from the *low-frequency words list* which we did not deem to be *low-frequency* words. We also removed names from the *low-frequency words list* (e.g. company names, country names, names of people).

Long words ground truth

Five ground truth lists were created to evaluate the LLM using a simple python script that filtered the three texts we used for the experiment:

- List of words consisting of words with a length ≥ 9.
- List of words consisting of words with a length ≥ 10.
- List of words consisting of words with a length ≥ 11.
- List of words consisting of words with a length ≥ 12.
- List of words consisting of words with a length ≥ 13.

Orthographic inconsistent words ground truth

In order to measure how well the LLM performs the task we developed an algorithm which is able to quantify the orthographic transparency of a word.

In an article by Neef et al.[NB11] a quantitive method is proposed to calculate the so called *graphematic transparency* of a language. The article uses two terms to describe the orthographic depth of a language, namely *graphematic transparency* and *orthographic transparency.* Graphematic transparency describes how reliably the pronunciation of a word can be deduced from its spelling. Orthographic transparency, in turn, describes how well the spelling of a word can be gathered from its phonological properties. The article's focus was on creating a means to calculate the graphematic transparency of a language by firstly establishing the relevant correspondence rules (a correspondence rule describes the relationship between a graphene and one or more phonemes) and then assigning numerical values to each of these rules based on its degree of transparency.

However, for the purposes of this thesis we will not make a distinction between different types of correspondence rules and instead treat each correspondence rule the same way: the numerical value assigned to a correspondence rule is determined based on how many phonemes map onto a specific graphene.

Tables A.7,A.8,A.9 describes how graphemes map onto phones. These tables were created using the *IPA-notatie voor het Nederlands*[IPA25].

Based on Tables A.7,A.8,A.9 we made the scoring as can be seen in Table 4.3.

Graphene	Score	Graphene	Score	Graphene	Score
<a>	3	<p>	1	<aa>	1
<e>	3	<q>	1	<ee>	2
<i>	3	<r>	2	<oo>	1
<o>	3	<s>	1	<uu>	1
<u>	3	<t>	2	<ij>	3
<b>	1	<v>	1	<ie>	2
<c>	2	<w>	1	<ei>	2
<d>	2	<y>	1	<eu>	2
<f>	1	<z>	1	<oe>	2
<g>	3	<vts>	1	<ou>	1
<h>	2	<ch>	3	<au>	1
<j>	3	<sh>	1	<ui>	2
<k>	1	<tsj>	1	<ao>	1
<l>	1	<ng>	1	<ail>	1
<m>	2	<nj>	1	<en>	1
<n>	2	<gn>	1	<ant>	1

Table 4.3: Grapematic transparency score for each graphene

Based on the scoring described in Table 4.3 we developed an algorithm that can determine the graphematic transparency of a word as can been seen in Listing 1. The algorithm breaks down a word into graphenes and sums up the score of each graphene in a word. This score is then divided by the square root of the word length, resulting in a graphematic transparency score (GTS) for that word.

The ground truth, for each text, is a list comprised of all the words in the text with a GTS greater than $Mean + 2SD$. The $Mean + 2SD$ is as follows for each of the texts:

- **Text 1**
  $3.96 + 2 * 1.36 = 6.68$
- **Text 2**
  $3.85 + 2 * 1.3 = 6.45$
- **Text 3**
  $3.85 + 2 * 1.3 = 6.45$

To answer sub-question 3 we investigated the LLM's ability to replace the lexical features for more "dyslexia-friendly" variants using the *zero-shot* prompt as can be seen in Table 4.4.

For this prompt we define a long word as a word which includes 11 or more letters. A study by Marian et al.[MBCS12] which looked into creating a database that would include so-called

```python
def GTscore (word: str, count: int) -> int:

    # all digraphs have been matched
    if len(word) == 0:
        return count

    else:

        if len(word) >= 3:
            if word[0:3] in ["vts", "tsj", "ail", "ant"]:
                return GTscore(word[3:], count+1)

        if len(word) >= 2:
            if word[0:2] in ["ch", "ij", "ee"]:
                return GTscore(word[2:], count+3)

            elif word[0:2] in ["ie", "ei", "eu", "oe", "ui"]:
                return GTscore(word[2:], count+2)

            elif word[0:2] in ["sh", "ng", "nj", "gn", "aa", "oo", "uu",
                               "ou", "au", "ao", "en"]:
                return GTscore(word[2:], count+1)

        if len(word) >= 1:
            if word[0] in ["a", "e", "o", "i", "u", "j", "g"]:
                return GTscore(word[1:], count+3)

            elif word[0] in ["c", "d", "h", "m", "n", "r", "t"]:
                return GTscore(word[1:], count+2)

            elif word[0] in ["b", "f", "k", "l", "p", "q", "s", "w", "y", "z", "v"]:
                return GTscore(word[1:], count+1)

    # digraph was not matched; add 1
    return GTscore(word[1:], count+1)

```

Listing 1: Algorithm which determines the graphematic transparency score (GTS) of a word.

”neighbourhood information” on commonly studied languages, found that the average word length of a Dutch words was 8.41 with a standard deviation of 2.79. This estimate was calculated using a corpus which included the 27,751 most frequently used Dutch words. We will therefore define a long word as consisting of 11 or more letters (which would be the average plus one standard deviation, $8.41 + 2.79 = 11.2 \approx 11$).

`list` refers to a list consisting of the union of the lists that were created to answer sub-question 2, 3, and 4 used to determine the precision and recall of the model.

Prompt
”Low-frequency words are words that are rarely used in the Dutch language. A long word is a word which consists of 11 or more letter. Orthographic inconsistent words are words for which the pronunciation of the word is difficult to predict based on its spelling. Given the following list of words: `list` Find a replacement words for each of these words that are not low-frequency words, are not long words, and are not orthographic inconsistent words. Make sure that each replacement words makes sense within the context of the following text: `text`”

Table 4.4: Prompt to substitute all the difficult words with more “dyslexia-friendly” words.

In order to evaluate the LLM's performance each word in these lists was given a *dyslexia-friendly score* (DFS) ranging from 0 to 3. A DFS for a word is determined by comparing that word to the ground truths that were used to evaluate the models performance for each of the three lexical features:

- **DFS = 3**
  The word does not appear in any of the three ground truths
- **DFS = 2**
  The word appears in one of the three ground truths
- **DFS = 1**
  The word appears in two of the three ground truths
- **DFS = 0**
  The word appears in all three of the ground truths

To evaluate the LLM's ability to find “dyslexia-friendly” substitutions for all the words the DFS of the original word was then compared to the DFS of its replacement word. If the DFS of a replacement words is greater than the DFS of the original we can conclude that the LLM was able to effectively substitute a word.

Table 5.24, 5.25, and 5.26 in section 5.2 show the percentage of words that were improved (had a higher DFS), and thus made more “dyslexia-friendly” by the LLM.

5 Results

5.1 Literature Review: Dyslexia Specific Reading Difficulties

The lexical features found through this review can be seen in Table 5.1. These features are specific word properties that dyslexic individuals struggle with to read according to the studies listed in the first column.

Next to the lexical features we also discovered a number of stylistic features which can be adjusted to make a text more suited to dyslexic readers. These features are listed in Table 5.2.

Author/Year	Feature	Explanation
2013[RBYDMS13]	low-frequency words	Low-frequency words are words which do not appear often in written language. A high rate of unfamiliar words can negatively affect the readability of a text for dyslexic individual.
2019[Pro19], 2013[RBYDMS13]	Word length	Dyslexic individuals read longer words much slower. This is due to Dyslexic people relying on letter-by-letter decoding.
2006 [BVVZ06]	Orthographic inconsistent words	Word consistency describe how predictable the relationship is between a word's spelling and its pronunciation. Words which are inconsistent are harder to read.
2013 [RBBY+13]	Digits instead of words	Readability improves in people with dyslexia when words are exchanged for digits.
2013 [RBBY+13]	Percentages over fractions	Readability improves in people with dyslexia when fractions are exchanged for percentages.

Table 5.1: List of lexical features dyslexic individuals struggle with reading.

Author/Year.	Feature	Explanation
2021[SK21], 2012[ZBF+12]	Increased letter spacing	Increasing the spacing between letters can improve reading speed.
2013[RBY13]	Font type	Font types, such as *Helvetica*, *Courier*, *Arial*, *Verdana* and *CMU*, can improve readability of people with dyslexia.
2017[RB17]	Background	Changing the background to a *warm* colour, such as peach, orange, and yellow with black fonts can increase reading speed.

Table 5.2: List of stylistic features that can be adjusted to accommodate dyslexic readers.

5.2 Experiment

Low-frequency Words

Text	Precision	Recall
1	52.94%	39.13%
2	48.39%	35.71%
3	100.0%	36.84%

Table 5.3: Precision and recall of the LLM given the low-frequency words prompt from Table 4.1 and system prompt 1 from Table 4.2, using model gpt-4o-mini.

Text	Precision	Recall
1	61.9%	28.26%
2	37.74%	47.62%
3	88.24%	39.47%

Table 5.4: Precision and recall of the LLM given the low-frequency words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Text	Precision	Recall
1	73.68%	30.43%
2	61.11%	13.1%
3	44.44%	63.16%

Table 5.5: Precision and recall of the LLM given the low-frequency words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-5.4-mini.

F1 score prompt 1 and 2 from Table 4.2, model gpt-4o-mini

Text	F1 score
1	45.0%
2	41.09%
3	53.84%

Table 5.6: F1 score given the low-frequency words prompt from Table 4.1 and system prompt 1 from Table 4.2, using model gpt-4o-mini.

Text	F1 score
1	38.80%
2	42.11%
3	54.54%

Table 5.7: F1 score given the low-frequency words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

F1 score model gpt-4o-mini and gpt-5.4-mini, prompt 2 from Table 4.2

Text	F1 score
1	38.80%
2	42.11%
3	54.54%

Table 5.8: F1 score given the low-frequency words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Text	F1 score
1	43.07%
2	21.58%
3	52.17%

Table 5.9: F1 score given the low-frequency words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-5.4-mini.

Long Words

Length \ Text	1	2	3
≥ 9	94.74% / 14.88%	85.19% / 17.16%	100.0% / 15.85%
≥ 10	94.74% / 21.69%	76.19% / 16.16%	100.0% / 15.52%
≥ 11	85.0% / 29.31%	72.22 % / 17.33%	90.91% / 31.25%
≥ 12	66.67% / 14.29%	47.62% / 20.0%	90.0% / 36.0%
≥ 13	70.0% / 24.14%	41.18% / 24.14%	71.43% / 38.46%

Table 5.10: Precision/recall of the LLM given the long words prompt from Table 4.1 and system prompt 1 from Table 4.2, using model gpt-4o-mini.

Length \ Text	1	2	3
≥ 9	86.96%/16.53%	92.31%/17.91%	85.71%/14.63%
≥ 10	91.67%/13.25%	86.21%/25.25%	88.89%/13.79%
≥ 11	81.82%/15.52%	66.67%/26.67%	100.0%/34.38%
≥ 12	58.33%/16.67%	50.0%/28.0%	87.5%/28.0%
≥ 13	50.0%/20.69%	45.45%/34.48%	80.0%/30.77%

Table 5.11: Precision/recall of the LLM given the long words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Length \ Text	1	2	3
≥ 9	97.1%/55.37%	88.89%/41.79%	83.33%/24.39%
≥ 10	74.07%/72.29%	83.64%/46.46%	81.58%/53.45%
≥ 11	60.0%/77.59%	68.18%/40.0%	46.0%/71.88%
≥ 12	41.82%/54.76%	50.0%/30.0%	84.62%/44.0%
≥ 13	36.0%/62.07%	45.83%/37.93%	85.71%/46.15%

Table 5.12: Precision/recall of the LLM given the long words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-5.4-mini.

F1 score prompt 1 and 2 from Table 4.2, model gpt-4o-mini

Length \ Text	1	2	3
≥ 9	25.72%	28.57%	27.36%
≥ 10	35.30%	26.66%	26.87%
≥ 11	43.59%	27.95%	46.51%
≥ 12	23.54%	28.17%	51.43%
≥ 13	35.90%	30.44%	50.0%

Table 5.13: F1 score given the long words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Length \ Text	1	2	3
≥ 9	27.78 %	30.0%	24.99%
≥ 10	23.15 %	39.06%	23.88%
≥ 11	26.09 %	38.1%	51.17%
≥ 12	25.93 %	35.9%	42.42%
≥ 13	29.27 %	39.21%	44.45%

Table 5.14: F1 score given the long words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4.o-mini.

F1 score model gpt-4o-mini and gpt-5.4-mini, prompt 2 from Table 4.2

Length \ Text	1	2	3
≥ 9	27.78 %	30.0%	24.99%
≥ 10	23.15 %	39.06%	23.88%
≥ 11	26.09 %	38.1%	51.17%
≥ 12	25.93 %	35.9%	42.42%
≥ 13	29.27 %	39.21%	44.45%

Table 5.15: F1 score given the long words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4.o-mini.

Length \ Text	1	2	3
≥ 9	70.52%	56.85%	37.74%
≥ 10	73.17%	59.74%	64.58%
≥ 11	67.67%	50.42%	56.1%
≥ 12	47.42%	37.5%	57.9%
≥ 13	45.57%	41.51%	60.0%

Table 5.16: F1 score given the long words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-5.4-mini.

Orthographic Inconsistent Words

Text	Precision	Recall
1	33.33%	3.57%
2	6.02%	61.54%
3	16.13%	22.73%

Table 5.17: Precision and recall of the LLM given the orthographic inconsistent words prompt from Table 4.1 and system prompt 1 from Table 4.2, using model gpt-4o-mini.

Text	Precision	Recall
1	40.0%	14.29%
2	7.45%	71.79%
3	35.71%	22.73%

Table 5.18: Precision and recall of the LLM given the orthographic inconsistent words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Text	Precision	Recall
1	6.96%	96.43%
2	20.0%	20.51%
3	12.9%	90.91%

Table 5.19: Precision and recall of the LLM given the orthographic inconsistent prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-5.4-mini.

F1 score prompt 1 and 2 from Table 4.2, model gpt-4o-mini

Text	F1 score
1	6.45%
2	10.97%
3	18.87%

Table 5.20: F1 score given the orthographic inconsistent words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Text	F1 score
1	21.06%
2	13.50%
3	27.78%

Table 5.21: F1 score given the orthographic inconsistent words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4.o-mini.

F1 score model gpt-4o-mini and gpt-5.4-mini, prompt 2 from Table 4.2

Text	F1 score
1	21.06%
2	13.50%
3	27.78%

Table 5.22: F1 score given the orthographic inconsistent words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-4o-mini.

Text	F1 score
1	12.98%
2	20.25%
3	22.59%

Table 5.23: F1 score given the orthographic inconsisten words prompt from Table 4.1 and system prompt 2 from Table 4.2, using model gpt-5.4-mini.

Dyslexia-friendly substitution

Text	Percentage of words improved	DFS original words	DFS improved words
1	69.77%	1.67	2.98
2	17.44%	2.59	2.87
3	48.78%	1.98	2.95

Table 5.24: Text improvement given system prompt 1 from Table 4.2, using model gpt-4o-mini.

Text	Percentage of words improved	DFS original words	DFS improved words
1	82.14%	1.54	2.96
2	5.21%	2.54	2.62
3	86.96%	1.09	2.91

Table 5.25: Text improvement given system prompt 2 from Table 4.2, using model gpt-4o-mini.

Text	Percentage of words improved	DFS original words	DFS improved words
1	16.11%	2.67	2.94
2	68.75%	1.77	2.96
3	30.0%	2.48	2.94

Table 5.26: Text improvement given system prompt 1 from Table 4.2, using model gpt-5.4-mini.

6 Discussion

This thesis meant to investigate if a large language model could be effectively make a text more "dyslexia-friendly". Firstly, we examined the LLM's ability to recognise specific lexical features.

The results indicate that an LLM, or at least the models we investigated in this thesis (gpt-4o-mini and gpt-5.4-mini) are not effective when it comes to our specific task. Namely, the identification of the lexical features *low-frequency words*, *long words*, and *orthographic inconsistent words*.

What makes them ineffective are the inconsistent results that we can observe per feature prompts (Table 4.1). Given the same prompt, the result across the three texts that we provided the LLM are significantly different. Where we would have expected the model to perform similarly for each texts given the same prompt and model, we found differing performance. This difference is most apparent when we look at the LLM's performance for the subtask *orthographic inconsistent words*, but can be observed across all feature specific prompts. Looking at Table 5.17 for example, we can see that both the precision and recall differ greatly across the three texts.

When we focus on comparing the performance of the different system prompts (prompt 1 and prompt 2 as defined in Table 4.2) by looking at the F1 score we observe the following:

- **Low frequency words**
  When we look at Tables 5.6 and 5.7 we can see that prompt 2 performs slightly better than prompt 1 for both text 2 and text 3. For text 1 however prompt 1 performs significantly better with a much higher F1 score.

- **Long words**
  Depending on the length of the word we ask the LLM to search for we find that either prompt 1 or prompt 2 results in a better F1 score for both text 1 and 3. The differences in percentage observed differ from slight to large differences, as can be seen in Table 5.13 and 5.14. For text 2, however, prompt 2 consistently performs better than prompt 1. Again the differences in percentage varies from small to large.

- **Orthographic inconsistent words** Tables 5.20 and 5.21 show that prompt 2 performs better overall with a higher F1 score across all three texts. Per text the difference in F1 score vary from small to large.

For *orthographic inconsistent words* we can observe that prompt 2 performs better overall. For *low-frequency words* and *long words* we cannot say definitively that either prompt 1 or prompt 2 result in a better performance. The influence of the two system prompts is different per feature specific prompt (as described in Table 4.1) but also across the three texts given the same feature specific prompt.

Focussing on the performance of the two models, gpt-4o-mini and gpt-5.4-mini, by looking at the F1 score, we again find varied results:

- **Low-frequency words**
  When we look at Tables 5.8 and 5.7 we can see that gpt-5.4-mini performs slightly better for both text 2 and text 3. For text 1 however gpt-4o-mini performs significantly better with a much higher F1 score.

- **Long words**
  In Tables 5.15 and 5.16 we can see that gpt-5.4-mini results in a higher F1 score for all lengths across all texts. The differences in percentage observed differ from slight to large.

- **Orthographic inconsistent words**
  Looking at Tables 5.22 and 5.23 we can see that model gpt-4o-mini performs better overall, with a higher F1 score, with the exception of text 1 where gpt-4o-mini performs significantly better.

For *long words* we can observe that gpt-5.4-mini performs better overall, with often times a much higher F1 score. For *low-frequency words* and *orthographic inconsistent words* we cannot say that either gpt-4o-mini or gpt-5.4-mini performed better.

Dyslexia-friendly substitution

Secondly, we investigated a LLM's ability to find appropriate substitution to the lexical features that were found in the first half of the experiment. Appropriate substitution are words that are more "dyslexia-friendly".

When looking at the results in Tables 5.24, 5.25, and 5.26 we find that the LLM was able to improve the dyslexia friendliness of all the words. The performance again differed substantially across the three instances (combination of prompt and model). Next to this the performance also differed per text given the same system prompt and model. And when comparing the performance per text across the three different instances (the three Tables) we again find no consistent performance.

For example, in Table 5.26 we see that the percentage of words improved for text 2 is much greater than the percentage of words improved for text 2 in Tables 5.24 and 5.25. This is likely due to the response that the LLM gave in the first part of the experiment where it had to find the specific lexical features (low-frequency, long words, and orthographic inconsistent words). For example, the LLM's response to the orthographic inconsistent words prompt (as described in Table 4.1) for system prompt 2 and model gpt-4o-mini as can be seen in Table 5.18 shows a relatively high recall and relatively low precision indicating that the LLM provided a lot of *false positives* in its output. This would explain why the percentage of improved words, which is 5.21% is so low. Because, whilst the percentage of improved words is low, the DFS of the improved words is fairly high. A great number of false positives added to the list that was given to the LLM when prompted with the prompt as described in Table 4.4 would skewer the percentage of words improved, as the DFS for the original words in that list would have already been high (due to the disproportionate amount of false positives included in the list).

7 Conclusion

From our finding we can conclude that a LLM is not effective at making a text more "dyslexia-friendly". Regarding the instances which we investigated in this thesis, namely the two system prompts and models, we found inconsistent results throughout regarding the precision and recall. We were not able to find a pattern within the LLM's response and the output which it generates seems to be entirely dependant on prompt, model, text, and task.

When we focus on the F1 score we find that the performance is more consistent. With regards to which system prompt (prompt 1 or prompt 2 as described in Table 4.2) results in a better performance we find that only for *orthographic inconsistent words* does prompt 2 perform better across all three texts. When we look at the performance in regards to which model performs better (gpt-4o-mini or gpt-5.4-mini) we can only say definitively that for *long words* gpt-5.4-mini performs significantly better across all three texts.

However, it is the inconsistency in performance based on precision and recall that lead to the varied results when it comes to the LLM's ability to generate more "dyslexia-friendly words. Despite, however, the LLM's performance on identifying the specific lexical features, the LLM was able to generate more "dyslexia-friendly" words in regards to both system prompts and models across all three texts.

Limitations and Future Research

Limitations include the construction of our ground truth for the *orthographic inconsistent words*. In order to create these ground truths we used our own metric to define an inconsistent word. For this thesis we attempted to create a metric that we could use to define orthographic inconsistent words. We did this through looking at the number of graphenes included in a word and the length of the word itself. However, due to our definition of the metric shorter but highly complex words could still end up not getting defined are orthographic inconsistent. We believe that this metric requires further scrutiny and could possibly be extended upon.

It is evident from the observations discussed above that the LLM's response is inconsistent as its performance varies greatly across all investigated tasks. It is however, not clear from this experiment alone what the exact cause of this variety is. In order to investigate this we would likely have to look in more detail at the exact response the models give. Future research could focus on investigating the inconsistency observed within this thesis.

Whilst in this thesis we did investigate whether the LLM was able to find substitute words that could be considered more “dyslexia-friendly” according to our metric (DFS), we did not investigate whether these words actually improved the readability of the texts. Future research could examine if the words suggested by the LLM actually made a text more dyslexia friendly.

References

[ABD+14] Theologos Athanaselis, Stelios Bakamidis, Ioannis Dologlou, Evmorfia N Argyriou, and Antonis Symvonis. Making assistive reading tools user friendly: A new platform for greek dyslexic students empowered by automatic speech recognition. *Multimedia tools and applications*, 68(3):681–699, 2014.

[AFNE20] Sikoya M Ashburn, D Lynn Flowers, Eileen M Napoliello, and Guinevere F Eden. Cerebellar function in children with and without dyslexia during single word processing. *Human brain mapping*, 41(1):120–138, 2020.

[BDS+10] Amy E Barth, Carolyn A Denton, Karla K Stuebing, Jack M Fletcher, Paul T Cirino, David J Francis, and Sharon Vaughn. A test of the cerebellar hypothesis of dyslexia in adequate and inadequate responders to reading intervention. *Journal of the International Neuropsychological Society*, 16(3):526–536, 2010.

[Boo25] Lee Boonstra. Prompt engineering. *Google, https://www. kaggle. com/whitepaper-prompt-engineering*, 2025.

[BVG+22] Prabal Datta Barua, Jahmunah Vicnesh, Raj Gururajan, Shu Lih Oh, Elizabeth Palmer, Muhammad Mokhzaini Azizan, Nahrizul Adib Kadri, and U Rajendra Acharya. Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review. *International journal of environmental research and public health*, 19(3):1192, 2022.

[BVVZ06] Anna MT Bosman, Wietske Vonk, and Margriet Van Zwam. Spelling consistency affects reading in young dutch readers with and without dyslexia. *Annals of Dyslexia*, 56(2):271–300, 2006.

[Cen] Dyslexie Centraal. Begeleiding bij het lezen en schrijven. https://dyslexiecentraal.nl/weten/ondersteuning-van-leerlingen-met-dyslexie/begeleiding-bij-het-lezen-en-schrijven.

[Cen25a] Dyslexie Centraal. Soorten ondersteuningsmogelijkheden. https://dyslexiecentraal.nl/weten/dyslexiehulpmiddelen/soorten-ondersteuningsmogelijkheden, April 2025.

[Cen25b] Dyslexie Centraal. Stimulerende leeromgeving. https://dyslexiecentraal.nl/weten/goed-lees-en-spellingonderwijs/stimulerende-leeromgeving, Februari 2025.

[Cha24] A Chandrasekar. Mastering prompt engineering: A guide to system, user, and assistant roles in openai api. Medium: https://medium.com/@csakash03/model-parameters-in-openai-api-fe6101d3f813, 2024.

[dh] Communiceer dyslexie hulpmiddelen. Begeleiding met dyslexiehulpmiddelen. https://www.dyslexiehulpmiddelen.com/files/203/wanneer+inzet+dyslexiehulpmiddelen.pdf.

[DSR+25] Tonmoy Debnath, Md Nurul Absar Siddiky, Muhammad Enayetur Rahman, Prosenjit Das, Antu Kumar Guha, Muhammad Rezaur Rahman, and HM Dipu Kabir. A comprehensive survey of prompt engineering techniques in large language models. 2025.

[FB22] Fonzie-Byte. 10_000-most-used-words-in-dutch/10.000 meestgebruikte woorden in het nederlands - uniek en zonder accenten.txt at main. GitHub: https://github.com/Fonzie-byte/10_000-most-used-words-in-Dutch/blob/main/10.000%20meestgebruikte%20woorden%20in%20het%20Nederlands%20-%20Uniek%20en%20zonder%20accenten.txt, April 2022.

[Gad25] Vrunda Gadesha. Prompt engineering techniques. https://www.ibm.com/think/topics/prompt-engineering-techniques, 2025.

[Hel25a] Helperbird. Accessibility productivity features for chrome, edge, firefox — helperbird. https://www.helperbird.com/features/, January 2025.

[Hel25b] Helperbird. Simplify text — make complex content easier to read - helperbird. https://www.helperbird.com/features/reword-and-simplify-websites/, January 2025.

[IPA25] Ipa-notatie voor het nederlands. https://nl.wikipedia.org/wiki/IPA-notatie_voor_het_Nederlands, 2025.

[KCS+23] Jennifer L Keelor, Nancy A Creaghead, Noah H Silbert, Allison D Breit, and Tzipi Horowitz-Kraus. Impact of text-to-speech features on the reading comprehension of children with reading and language difficulties. *Annals of dyslexia*, 73(3):469–486, 2023.

[KCSHK20] Jennifer L Keelor, Nancy Creaghead, Noah Silbert, and Tzipi Horowitz-Kraus. Text-to-speech technology: Enhancing reading comprehension for students with reading difficulty. *Assistive Technology Outcomes & Benefits*, 14(1):19–35, 2020.

[MBCS12] Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook. Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. 2012.

[NB11] Martin Neef and Miriam Balestra. Measuring graphematic transparency: German and italian compared. *Written Language & Literacy*, 14(1):109–142, 2011.

[NFD01] Roderick I Nicolson, Angela J Fawcett, and Paul Dean. Developmental dyslexia: the cerebellar deficit hypothesis. *Trends in neurosciences*, 24(9):508–511, 2001.

[Nie26] B Nieuwhof. Wat hébben we nu eigenlijk aan een menselijke maanlanding? https://www.nrc.nl/nieuws/2026/04/23/wat-hebben-we-nu-eigenlijk-aan-een-menselijke-maanlanding-a4925152, April 2026.

[Ope24] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.

[Ope26] OpenAI. Introducing gpt-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026.

[PB20] A Jothi Prabha and Renta Bhargavi. Predictive model for dyslexia from fixations and saccadic eye movement events. *Computer Methods and Programs in Biomedicine*, 195:105538, 2020.

[PCV+25] Francesco Piferi, Giovanni Caleffi, Giulia Valcamonica, Pietro Crovari, and Franca Garzotto. Integrating large language models into therapeutic education for children with dyslexia: a multimodal framework. In *Proceedings of the D-SAIL Workshop-Transformative Curriculum Design: Digitalisation, Sustainability, and AI Literacy for 21st Century Learning, CEUR-WS, Palermo, Italy*, pages 70–79, 2025.

[PP12] Robin L Peterson and Bruce F Pennington. Developmental dyslexia. *The lancet*, 379(9830):1997–2007, 2012.

[Pro19] Giofrè D. Adams A. M. Roberts D. J. Provazza, S. The clock counts - length effects in english dyslexic readers. frontiers in psychology, 2019.

[RB17] Luz Rello and Jeffrey P Bigham. Good background colors for readers: A study of people with and without dyslexia. In *Proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility*, pages 72–80, 2017.

[RBA+24] Mahmoud Robaa, Mazen Balat, Rewaa Awaad, Esraa Omar, and Salah Aly. Explainable ai in handwriting detection for dyslexia using transfer learning, 10 2024.

[RBB15] Luz Rello, Miguel Ballesteros, and Jeffrey P Bigham. A spellchecker for dyslexia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 39–47, 2015.

[RBBY+13] Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. One half or 50%? an eye-tracking study of number representation readability. In *IFIP Conference on Human-Computer Interaction*, volume 8120, pages 229–245, 2013.

[RBY13] Luz Rello and Ricardo Baeza-Yates. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, 2013.

[RBYDMS13] Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer, 2013.

[RPM+05] Timothy V Rasinski, Nancy D Padak, Christine A McKeon, Lori G Wilfong, Julie A Friedauer, and Patricia Heim. Is reading fluency a key for successful high school reading? *Journal of Adolescent & Adult Literacy*, 49(1):22–27, 2005.

[Sar26] K Saris. Houd na een hersenbloeding de bloeddruk laag en de kans op een nieuwe beroerte neemt aanzienlijk af. https://www.nrc.nl/nieuws/2026/04/28/houd-na-een-hersenbloeding-de-bloeddruk-laag-en-de-kans-op-een-nieuwe-beroerte-neemt-aanzienlijk-af-a4926216, April 2026.

[SDN16] P. F. De Bree E. H. Henneman K. Kleijnen R. Loykens E. H. M. Rolak M. Struiksma A. J. C Verhoeven L. Wijnen F. N. K. SDN, De Jong. Dyslexie: diagnostiek en behandeling. brochure van de stichting dyslexie nederland. https://dare.uva.nl/personal/pure/en/publications/dyslexie(a81d62b3-7fcb-4def-8283-109d6bf302e8).html, 2016.

[SK21] Steven D Stagg and Nathan Kiss. Room to read: The effect of extra-large letter spacing and coloured overlays on reading speed and accuracy in adolescents with dyslexia. *Research in Developmental Disabilities*, 119:104065, 2021.

[SSS+24] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1, 2024.

[Ste01] John Stein. The magnocellular theory of developmental dyslexia. *Dyslexia*, 7(1):12–36, 2001.

[Ste23] John Stein. Theories about developmental dyslexia. *Brain sciences*, 13(2):208, 2023.

[UNE] UNESCO. Spanish application using ai to help overcome dyslexia wins 2019 unesco ict prize. https://www.unesco.org/en/articles/spanish-application-using-ai-help-overcome-dyslexia-wins-2019-unesco-ict-prize.

[VG26] M Van Gestel. Eerste wereldwijde top over stoppen met olie, kolen en gas van start. 'op het best mogelijke moment'. https://www.trouw.nl/duurzaamheid-economie/eerste-wereldwijde-top-over-stoppen-met-olie-kolen-en-gas-van-start-op-het-best-mogelijke-moment~b248c782/, April 2026.

[YAC25] Jia Rong Yap, Thirishankari Aruthanan, and Mellisa Chin. Artificial intelligence in dyslexia research and education: a scoping review. *IEEE Access*, 13:7123–7134, 2025.

[Yan24] C Yan. A comprehensive guide to openai gpt model api parameters. Medium: https://chrisyandata.medium.com/a-comprehensive-guide-to-openai-gpt-model-api-parameters-c2c6dff9ce19, 2024.

[ZBF+12] Marco Zorzi, Chiara Barbiero, Andrea Facoetti, Isabella Lonciari, Marco Carrozzi, Marcella Montico, Laura Bravar, Florence George, Catherine Pech-Georgel, and Johannes C Ziegler. Extra-large letter spacing improves reading in dyslexia. *Proceedings of the National Academy of Sciences*, 109(28):11455–11459, 2012.

[ZXP+25] Sihang Zhao, Shoucong Carol Xiong, Bo Pang, Xiaoying Tang, and Pinjia He. Let ai read first: Enhancing reading abilities for individuals with dyslexia through artificial intelligence. In *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems*, pages 1–16, 2025.

A Tables

	1	2	3		1	2	3
ramp	x			consequente	x		
fossielebrandstoffencrisis	x	x	x	optelsom			
hameren	x			ontwikkeling	x		
hooggespannen	x	x	x	kippenvel	x		
denktank	x		x	allen	x		
atmosferisch	x			verdrag	x		
historisch	x	x	x	compromitteren	x		
momentum	x	x	x	investeringen	x		
omslagpunt	x	x	x	wereldwijde		x	
tijdperk	x			energiecrisis		x	x
transitie	x	x	x	optimistisch		x	x
krankzinnig	x	x	x	klimaatverandering		x	x
vooruitgang	x			geruchten		x	
aangemoedigd	x			vergadering		x	
vergezichten	x			gesteggel		x	x
milieuminister	x	x		verdrag		x	
kerosineschaarste	x			duurzaamheid		x	
afhankelijkheid	x	x		zorgvuldig		x	
export	x			afbouwpaden		x	x
divers	x			samenwerking		x	
eilandstaten	x			potentie		x	
gevaar	x			signalen		x	
onderhandelingen	x			overheidsvideo		x	
bezorgdheid	x			redmiddel			x
inspanning	x			VN-klimaattoppen			x
integreren	x			afhankelijkheid			x
bepleit	x		x	klimaatexperts			x
afscheid	x			klimaathoogleraar			x
schone	x			fossiele			x
economische	x			vergezichten		x	
malaise	x		x	slotverklaring			x
afbouwpaden	x			ngo's			x
vervuilende			x	schaarste			x

Table A.1: Output LLM to subtask 1 in Table 4.1, with temperature set to $default$, tested on **text 1**, model gpt-4o-mini.

	1	2	3		1	2	3
energiecrisis	x	x	x	slotverklaring	x	x	x
fossielebrandstoffencrisis	x	x	x	samenwerkingen	x	x	x
hooggespannen	x	x	x	vervuilende	x	x	x
optimistisch	x	x	x	bepleit	x	x	x
historisch	x	x	x	redmiddel	x	x	x
transitie	x	x	x	economische malaise	x	x	
krankzinnige	x	x	x	afbouwpaden	x	x	x
frustratie	x	x	x	druk	x	x	x
onderhandelen	x	x	x	inzet		x	x
vooruitgang	x	x	x	vertrouwen		x	x
klimaathoogleraar	x	x	x	aantrekkelijk		x	x
potentie	x	x	x	inzichten		x	x
onafhankelijker	x	x	x	samenhangt		x	x
acute	x	x	x	milieuminister		x	x
kerosineschaarste	x	x	x	schone energie		x	
vergezichten	x	x	x	economische			x
divers	x	x	x	malaise			x
afhankelijkheid	x	x	x	schone			x
eilandstaten	x	x	x	veiligheid			x
gevaar	x	x	x	geschiedenis			x
gesteggel	x	x	x				

Table A.2: Output LLM to subtask 1 in Table 4.1, with temperature set to 0, tested on **text 1**, model gpt-4o-mini.

	1	2	3		1	2	3
wereldwijde	x	x	x	afbouwpaden	x	x	
energiecrisis	x	x	x	afhankelijkheid	x		x
fossielebrandstoffencrisis	x	x	x	milieuminister	x		
verwachtingen	x	x	x	veiligheid	x		
optimistisch	x	x	x	geschiedenis	x		
historisch	x	x	x	hooggespannen		x	x
belangrijk	x			organisatie		x	
klimaatverandering	x	x	x	gesprekken		x	
vertegenwoordigt	x		x	aanwezigheid		x	
onafhankelijker	x			fossiele		x	
hoopvol	x		x	brandstoffen		x	
acute	x			aantrekkelijk		x	
vergezichten	x		x	onderhandelingen		x	
klimaatcrisis	x	x		succes		x	
Europese	x			samenwerkingen		x	x
afhankelijkheid	x	x		strijd		x	
vreemdelingen	x			malaise		x	
succesverhalen	x			Parijsakkoord			x
vervuilende	x	x	x	kwetsbaarheid			x
afspraken	x		x	nieuwe			x
verdrag	x			aanvaardbaar			x
overleven	x			eilandstaten			x
economische	x	x	x				

Table A.3: Output LLM to subtask 2 in Table 4.1, with temperature set to $default$, tested on **text 1**, model gpt-4o-mini.

	1	2	3		1	2	3
wereldwijde	x	x	x	afhankelijkheid	x	x	x
energiecrisis	x	x	x	fossiele	x	x	x
fossielebrandstoffencrisis	x	x	x	brandstoffen	x		
verwachtingen	x	x	x	klimaatcrisis	x	x	x
hooggespannen	x	x	x	aantrekkelijk	x	x	x
optimistisch	x	x	x	onderhandelingen	x	x	x
historisch	x	x	x	samenwerkingen	x	x	x
belangrijk	x	x	x	Milieudefensie	x		
transitie	x	x	x	vervuilende	x	x	x
krankzinnige	x	x	x	afspraken	x		x
klimaatverandering	x	x	x	eilandstaten	x		
vooruitgang	x	x	x	redmiddel	x	x	x
Parijsakkoord	x	x	x	economische	x	x	x
potentie	x	x		inzichten	x		x
onafhankelijker	x	x	x	afbouwpaden	x	x	x
conferentie	x	x	x	verenigen	x	x	
terughoudend	x	x	x	veiligheid	x	x	
acute	x	x		geschiedenis	x	x	
kerosineschaarste	x			organisaties		x	
vergezichten	x			analisten			x

Table A.4: Output LLM to subtask 2 in Table 4.1, with temperature set to 0, tested on **text 1**, model gpt-4o-mini.

	1	2	3
olie	x		
koolen	x		
gas	x		
fossielebrandstoffencrisis	x	x	x
klimaatexperts	x		
klimaathoogleraar	x		
veto	x		
Krankzinnig	x	x	x
milieuminister	x	x	x
afhankelijkheid	x	x	
afbouwpaden	x	x	
energiecrisis		x	x
optimistisch		x	x
zie		x	
transitie		x	
historisch			x
moment			x
overheidsvideo			x

Table A.5: Output LLM to subtask 3 in Table 4.1, with temperature set to $default$, tested on **text 1**, model gpt-4o-mini.

	1	2	3
energiecrisis	x	x	
fossielebrandstoffencrisis	x	x	
optimistisch	x	x	
transitie	x	x	
klimaatverandering	x	x	
Krankzinnig	x	x	
milieuminister	x	x	
afhankelijkheid	x	x	
vervuilende	x	x	
afbouwpaden	x	x	

Table A.6: Output LLM to subtask 3 in Table 4.1, with temperature set to 0, tested on **text 1**, model gpt-4o-mini.

Graphene	Phoneme	Graphene	Phoneme
<a>	[ɑ] [a] [aː]	<o>	[ɔ] [o] [oː]
<e>	[ɛ] [e] [ə]	<u>	[y] [yː] [ʏ]
<i>	[ɪ] [i] [ə]		

Table A.7: Graphene to phoneme correspondence, vowels

Graphene	Phoneme
<b>	[b]
<c>	[k] [s]
<d>	[d] [t]
<f>	[f]
<g>	[g] [ʒ], [ʑ] [d͡ʒ], [d͡ʑ]
<h>	[ɦ] [ʰ]
<j>	[j] [ʒ], [ʑ]
<k>	[k]
<l>	[l]
<m>	[m] [ɱ]
<n>	[n] [ŋ]
<p>	[p]
<q>	[k]
<r>	[r], ([ɻ], [ɽ] in some dialects) [ɹ]
<s>	[s]
<t>	[t] [t͡s]
<v>	[f]
<w>	[ʋ]
<y>	[i]
<z>	[z]

Table A.8: Graphene to phoneme correspondence, consonants

Graphene	Phoneme
<vts>	[t͡s]
<ch>	[ʃ], [ɕ] [t͡ʃ], [t͡ɕ] [ʒ], [ʑ]
<sh>	[ʃ], [ɕ]
<tsj>	[t͡ʃ], [t͡ɕ]
<ng>	[ŋ]
<nj>	[ɲ]
<gn>	[ɲ]
<aa>	[a:]
<ee>	[e] [ə]
<oo>	[o]
<uu>	[y:]
<ij>	[ə] [ɛɪ] [ɛ:]
<ie>	[i] [i:]
<ei>	[ɛɪ] [ɛ:]
<eu>	[ø] [œ]
<oe>	[u] [u:]
<ou>	[ʌu]
<au>	[ʌu]
<ui>	[œy] [œy:]
<ao>	[ʌu]
<ail>	[ɑɪ]
<en>	[ɑ̃]
<ant>	[ɑ̃]

Table A.9: Graphene to phoneme correspondence, digraphs