



M Ű E G Y E T E M 1 7 8 2

**Budapesti Műszaki és Gazdaságtudományi Egyetem**

Villamosmérnöki és Informatikai Kar

Irányítástechnika és Informatika Tanszék

Pütkösti Györk

**MULTIMODÁLIS AUTONÓM  
ÁGENSEK ALKALMAZÁSA  
JÁRMŰÍRÁNYÍTÁSI  
FELADATOKBAN**

KONZULENS

**Dr. Szántó Mátyás**

BUDAPEST, 2023



# Tartalomjegyzék

<b>Összefoglaló .....</b>	<b>4</b>
<b>Abstract.....</b>	<b>5</b>
<b>1. Bevezetés .....</b>	<b>6</b>
<b>2. Irodalomkutatás.....</b>	<b>9</b>
2.1. Vizuális nyelvi modellek .....	9
<b>3. Módszertan .....</b>	<b>14</b>
3.1. Pontozási rendszer hasonló modelleknél .....	18
<b>4. Kísérletek.....</b>	<b>19</b>
<b>5. Kiértékelés .....</b>	<b>27</b>
<b>6. Összegzés.....</b>	<b>31</b>
<b>7. Kitekintés.....</b>	<b>32</b>
<b>Köszönetnyilvánítás .....</b>	<b>34</b>
<b>Irodalomjegyzék.....</b>	<b>35</b>

# Összefoglaló

A mai napig nagy problémát jelent az autonóm járművek elterjedésében a rendszer környezetérzékelésének és -értelmezésének folyamata, illetve, hogy ezt a felhasználók számára nem érthető módon teszi meg. Ez manapság nagy gátat okoz az önvezető autóipar számára, hiszen az emberek nem bíznak meg ezen rendszerek pontosságában, mivel nem ismerik a pontos és érthető működésüket.

Sajnálatos módon nem könnyítik meg a helyzetet a napjainkban a környezetértelmezési feladatokban alkalmazott deep learning hálók sem, amelyek, habár nagy pontossággal, de rendkívül bonyolultan oldják meg az önvezetés feladatát. Erre a problémára jelenthet megoldást a kutatásom célja, amelynek kulcseleme a környezetértelmezés megvalósítása vizuális nyelvi modellekkel (Visual Language Models – VLM). A VLM-eket nagy siker övezi manapság: ezek a modellek bonyolult képi adatokat is képesek feldolgozni, majd természetes nyelven – mindenki számára érthetően – egy részletes leírást adni róluk. Munkám során olyan megoldást próbálok keresni az felhasználó-önvezetés feszültség oldására, ami jól működő önvezető autók létrehozásában is használható megoldást jelenthet.

A kutatásom során a MiniGPT4 vizuális nyelvi modell alkalmazásával olyan ágenseket fejleszték, amelyek képesek valós és mesterséges környezetből szintetizált képeket feldolgozni, és ezeken a jármű környezetét emberek számára is érthető módon leírni.

# Abstract

Until now, a significant challenge in the adaption of autonomous vehicles is the process of the system's environmental sensing and interpretation, which is not made in an understandable way to users. This, currently, makes a major barrier to the self-driving automotive industry, since people do not trust the accuracy of these systems, because as far as they are concerned, people do not understand the way of their precise and comprehensible functioning.

Sadly, the use of deep learning networks in environmental perception tasks does not make this situation any easier today. While they solve the task of self-driving with high precision, they do so in an extremely complex manner. The aim of my research is to address this problem, with a key focus on implementing environmental perception using visual language models (VLMs). VLMs are very successful nowadays. These models are able to process complex visual data and provide a detailed description of it in natural language that is understandable to everyone. In my work, I am going to try to find a solution to ease the tension of the human vs self-driving conflict, but also to create a useful solution of self-driving problems.

During my research, I am going to use the MiniGPT-4 visual language model in order to develop agents that can process real and artificially synthesized images from both natural and artificial environments and provide a comprehensible description of the vehicle's surroundings to humans.

# 1. fejezet

## Bevezetés

A szállítás problémája és az arra való megoldások kezelése már az ősidőkre eredeztethető vissza. Ahogy azt a felfedett párezeréves barlangrajzok is sugallják, az ősember vándorló életmódot folytatott, és ebből kifolyólag akár az élete is múlhatott azon, hogy képes-e egyik barlangból eljutni a másikba. Az idő múlásával viszont ahogy fejlődött a technológia, s folyamatosan továbblépett a kőkorszakból a bronz- majd pedig a vaskorba a szállítást is fejleszteni kellett. Az ősember munkaeszközei nagy értéket jelentettek a számára, ezek sokszor apáról fiúra szálltak, így nem engedhette meg magának, hogy minden egyes alkalommal amikor lakhelyet változtatott, újakat készítsen magának. A történészek számos bizonyítékot találtak arra, hogy a szállítás egyik legritimvebb segédeszköze a saru volt, ami megvédte a viselője lábát a különböző sérülésektől, ezzel elősegítve, hogy nagyobb távolságot tudjon megtenni rövidebb idő alatt.

Az ember számára mindig egy fontos cél volt az, hogy a legkevesebb munka befektetésével valósítson meg valamit. Éppen ezért jogosan tehető fel a kérdés, hogy mikortól beszélhetünk önvezetésről, és hova vezethető vissza ennek az eredete. Ehhez fontos ismernünk az autonóm járműirányítás definícióját, aminek az autóipar a következőt adta: „*képesség arra, hogy egy jármű részben vagy teljesen irányítsa önmagát limitált emberi közbelépéssel vagy e nélkül*”. Ha ezt a meghatározást vesszük alapul, akkor arra a következtetésre juthatunk, hogy az önvezetés alappilléreit még jóval a számítógép feltalálása előtt lefektették. Egyes feljegyzések alapján az ókori egyiptomiak a Nílus áramlatait használták ki, hogy mészkövet és márványt szállítsanak a gízai piramisok építéséhez. Ugyanúgy a hajósok is hasznosították a tengeráramlatokat, hogy azok „elvigyék” őket úticéljukhoz. Ezen módszerekben közös, hogy az irányítás javát az áramlatok végezték, a legénység feladata csupán minimális koordinálás volt. Mai szemmel így ezek tekinthetőek ősi robotpilótáknak.

Manapság, a XXI. században számos területen alkalmaznak már robotpilótákat. Ezek felhasználása dominál a levegőben történő szállítás során. A mai repülőgépek képesek maguktól felszállni, elrepülni a célig, majd ott biztonságosan leszállni, a pilóta

legfőbb dolga ezen rendszer felügyelete, szükség esetén korrigálása. Az első repülőgépben működő robotpilótát nemsokkal a repülőgépek feltalálása után alkalmazták. 1912-ben a Sperry Corporation készített egy olyan rendszert, ami képes volt arra, hogy egy repülőgép egyenesen repüljön pilóták irányítása nélkül. A második világháborúban Wernher von Braun, akinek nevéhez az Apollo-11 holdrakéta is fűződik, olyan hadászati rakétákat fejlesztett ki 1944-ben, amelyek kilövésük után képesek voltak maguktól eljutni egy előre betáplált célhoz, mindezt úgy, hogy akkoriban még nem létezett se GPS, se számítógép. Így nem is csoda, hogy a mai repülőek mikre nem képesek az okosrendszerekkel.

Napjainkban a szállítás megvalósulását tekinthetjük egy közvetett létszükségletnek is, hiszen nélküle a legtöbb minket ellátó szolgáltatás nem lenne képes működni. Az Európában megvalósuló szállításokat három különálló csoportra bonthatjuk fel: vízi, ami 46.0%-kal magaslik ki az összes kontinentális szállítás közül, légi, ami 20.7%-ot foglal magába és szárazföldi, ami 33.3%-ot tudhat a magáénak (a szárazföldi szállítás még tovább bontható három alrészre: úti – 19.5%, vasúti – 1.5% és egyéb, mint például a csőhálózatokon keresztül történő – 12.3%) [1]. E három főcsoport közül a légi és vízi transzportálás erősen automatizált, míg a szárazföldi szinte egyáltalán nem működőképes állandó emberi irányítás nélkül. Ez is bizonyíték arra, hogy az önvezető autók megoldásának kérdése nagy jelentőséggel bír jelenleg is. Dolgozatom során célom egy olyan megoldást megvalósítása, amely ezen arányokat nagyban javítja majd, elősegítve az autonóm járművek technológiájának fejlesztését.

Kihívást jelent viszont még mindig a biztonságos önvezető autók kifejlesztése. A földön történő járműirányítás jóval komplexebb, mint a levegőben történő, hiszen jelentősen több zavaró tényező található lenn, mint például a járókelők, az épületek, a forgalomirányítás elemei vagy akár váratlanul történő események. Ahhoz, hogy az önvezető járművünk balesetmentesen tudjon haladni, létfontosságú, hogy ezen tényezőkre rögtön pontos választ tudjon adni.

Az emberek bizalmatlansága sem javít az autonóm járművek helyzetén. A manapság is létező megoldások, mint a deep learning hálók alkotta nagy pontosságú környezetérzékelés és -értelmezés folyamata olyan módon történik, amit a felhasználók nem feltétlenül értenek, és így a jármű irányába elengedhetetlen teljes bizalmuk lehetősége csökken.

Dolgozatom során erre a problémára keresem a megoldást, kulcselemként használva a környezetértelmezés megvalósítását vizuális nyelvi modellekkel, amelyek képesek feldolgozni akár bonyolult képanyagokat is, majd természetes nyelven, mindenki számára érthető módon, részletes leírást adni róluk. Terveim szerint egy olyan megoldást fogok kidolgozni, amit majd később autonóm járművekbe is be lehet építeni. Ennek megvalósításához a MiniGPT-4 vizuális nyelvi modellt használom fel, párosítva a Llama-2 Chat 7B (7 milliárd paraméterrel betanított), a Vicuna V0 7B és a Vicuna V0 13B (13 milliárd paraméterrel betanított) nagy nyelvi modellekkel (LLMs – large language models), olyan ágensek fejlesztéséhez, amelyek valós és mesterséges környezetből szintetizált képeket képesek feldolgozni, majd emberek számára érthető módon leírni.



## 2. fejezet

# Irodalomkutatás

Az önvezetés technológiájának kutatása és felhasználása egy manapság széleskörben foglalkoztatott téma. Számos megközelítése létezik, ennek egy kis része már használatban is van, de javában még fejlesztés és kutatás alatt áll. A ma használt autonóm járművek többségének alapjául a deep learning hálók szolgálnak, melyek sok esetben nagy pontossággal képesek megoldani a problémát, viszont ezt nagyon nagy bonyolult módon teszik, az emberek számára sokszor érthetetlen módon. Továbbá számos olyan esetet különíthetünk el, amelyek során ezen rendszerek tévedtek, olyan ember számára egyértelmű jeleket értelmeztek félre, amik súlyos következményeket vonhatnak maguk után. Ilyen lehet például egy matricával teleragasztott stoptábla, amit egy sofőr bármikor felismer, viszont egy ágens akár egy 30-as táblával is felcserélheti.

Én a kutatásom során vizuális nyelvi modellek segítségével közelítem meg az autonóm vezetés problémáját. Ez egy aránylag új módszer, amivel még nem sokan foglalkoztak, és a 2022-ben elért nagy nyelvi modellek kutatása során elért áttörések tették lehetővé. Ezen megközelítés során egy olyan megoldást keresek, amely képes a fentebb említett deep learning hálók okozta problémák kiküszöbölésére is.

## 2.1. Vizuális nyelvi modellek

### 2.1.1 Nagy nyelvi modellek

A kommunikációnak egyik kulcspillére mindig a nyelv volt. A nyelv a legáltalánosabb jelekből álló jelrendszer. Mindennapi életünkben egy jelentős szerepet tölt be. Ugyanakkor elengedhetetlen szerepet tölt be a humán-gép interakciók során. A LLM-ek fő célja, hogy szövegeket dolgozzanak fel vagy írjanak, majd ezek segítségével tisztán érthető kommunikációt vigyenek véghez [2]. Szükségességüknek kiváltó okai közé sorolható a tény, hogy a napjainkban az egyre gyorsuló információáradatot a jelenlegi számítógépek minimálisan hasznosítható mértékben tudják feldolgozni, éppen ezért jelentős változást hozhatnak az olyan modellek, amelyek képesek komplex nyelvi

feladatok elvégzésére, mint például a fordítás, szövegösszefoglalás vagy beszélgetések lebonyolítása.

A nagy nyelv modellek hatalmas fejlődésen mentek keresztül az utóbbi években. Köszönhető ez annak, hogy a tanuló adathalmazok jóval nagyobbak lettek, így hatalmas mennyiségű paraméterrel lehet betanítani ezeket a modelleket, jelentősen javítva működésük precizitását. Az első jelentősebb nyelvi modellek, mint az OpenAI által kifejlesztett GPT-1 [3] vagy a Google készítette BERT 2018-ban jelentek meg, miután a Google kifejlesztette a transzformer architektúrát [5], ami egy multi-head attention mechanizmuson alapszik; egy enkóder-dekóder struktúrát használ, amely a válaszadáshoz nem alapoz se rekurrenciára, se pedig konvolúcióra. Ezek a modellek képesek voltak egyszerű chatbeszélgetésekre a felhasználókkal, ha olyan témákról volt szó, amilyen adatokon betanították ezeket. A későbbi modellek folyamatosan fejlődtek, a GPT-3 [6] (OpenAI) már a 175 milliárd paraméterével képes volt komplexebb feladatok elvégzésére is, mint cikkek, versek, történetek írása, mindezt úgy, hogy a megadott szövegben kiválogatta azokat a szavakat, amiket a legjelentősebbeknek tekintett. Akár egyszerűbb számolási feladatokat is meg tudott oldani. Jelenleg a legfejlettebb nyelvi modellek közé sorolható a GPT-4 [7] a maga 1760 milliárd paraméterével, a LLaMA [8] (Meta) vagy a Vicuna [9] (LMSYS Org), amelyek képesek komplex szövegek értelmezésére és megírására, valamint képek feldolgozására is természetes nyelven.

Habár a LLM-ek rendkívüli pontossággal képesek választ generálni, nem feltétlenül lehet egyelőre vakon megbízni a válaszadási képességükben, hiszen megtörténhet, hogy hihető formában rossz választ adnak a kérdésünkre. Ezt okozhatja a megfelelő tudás hiánya, vagy a modell valamilyen hibája abban az esetben, ha a szükséges tudás mégis megvan. Éppen ezért a nyelvi modelleket úgy pontozzák, hogy egyszerű kérdéseket tesznek fel nekik, majd megvárják, hogy helyes vagy helytelen választ ad a kérdésre, vagy azt feleli, hogy nem biztos benne. Egy ilyen teszt akár tizennyolcezer kérdés-válasz párból is állhat. [10]

### **2.1.2 Vizuális nyelvi modellek a mindennapokban**

Talán ki lehet jelenteni, hogy a XXI. században a adat a leghatalmasabb értékek közé tartozik. Ezt nem csak írásos formában jegyzi fel, rengeteg videó- és képanyagot is lementenek nap, mint nap. Évezredek óta az ember vizuálisan jut hozzá a legtöbb információhoz, mindent, amit lát megjegyez és tanul belőle. Éppen ezért a VLM-ek nagy

szerepet fognak betölteni a mindennapi életünkben, hiszen képesek leegyszerűsíteni azt a komplex információáradatot, amit különben nekünk kellene feldolgoznunk, ezzel leegyszerűsítve a mindennapi feladatainkat. Ilyen lehet például a kézírásfelismerő rendszer, ami képes a kézzel írott karaktereket felismerni és nyomtatott karakterekké alakítani [11]. Az utóbbi időben a VLM-ek nagy fejlődést mutattak a képek emberi nyelven való leírásában, vagy éppen szöveg utáni generálásban, szerkesztésében. Ez még egy aránylag fiatal kutatási terület, de a technológia és modellek továbbfejlesztésének hála robbanásszerűen fejlődik.

A VLM-eknek három általános kulcselemük van: képenkóder, szövegenkóder és egy stratégia, ami a két enkóder adatait összeköti; ezek szoros kapcsolatban állnak egymással, így javítva a veszteségfüggvény értékén. Ugyanakkor míg a korai modellek kézzel írt képleírásokkal és előre betanított szövektorokkal működtek, addig az újabb modellek a transzformer architektúrát alkalmazzák, a képek és szövegek különálló vagy egyesített feldolgozására. Fontos kiemelni, hogy a VLM-ek betanítása több módon is működhet, ebből fogok most néhányat röviden kiemelni [12]:

- **Kontrasztív tanulás:** a képek és a szöveg közös térben történő feldolgozása
- **PrefixLM:** egyesített kép-szöveg tanulás, a képek előtagként szerepelnek
- **Multimodális fúzió és cross attention:** a képanyag rétegekre való bontása a nyelvi modellben cross-attention mechanizmussal
- **MLM/ITM:** képrészek összerendelése a szöveggel maszkolt nyelvi modellek és kép-szöveg párosítás által
- **Nincs tanítás:** különálló kép- és nyelvi modellek használata iteratív optimalizálással

A VLM-ek betanítása során fontos szerepet játszanak az adathalmazok, amelyeken betanítják ezeket. Annak függvényében, hogy mit várunk el egy modelltől, különböző adathalmazokon taníthatjuk, finomhangolhatjuk őket. Ezek lehetnek már előre előkészített adathalmazok, mint a Laion-5b [13], a VIMA [14] vagy pedig a COYO-700M [15].

Napjainkban számos vizuális nyelvi modell van már használatban, ezek közül néhány ismertebb: az OpenAI fejlesztette DALL-E 3 [16] vagy CLIP [17], de ide sorolható a ViLBERT [18] és a FLAMINGO [19] is.

### 2.1.3 MiniGPT-4

Kutatásom során a MiniGPT-4-et [20] fogom használni. Azért választottam ezt a modellt, mert a nagyobbakkal ellentétben kevesebb GPU memóriára van szüksége, emiatt kisebb teljesítményű számítógépen is alkalmazható. A legtöbb nyelvi modellnek hatalmas méretű GPU-kat igényelnek, így laboratóriumi körülmények között legtöbbjüket lehetetlen lenne futtatni. Erre a problémára jelent megoldást a MiniGPT-4.

A MiniGPT-4 architektúrája szempontjából tartalmaz egy vizuális enkódert egy ViT-vel és Q-Formerrel összekötve, egy lineáris projektáló réteget és egy fejlett LLM-et, jelen esetben sorra a Llama-2 Chat 7B-t [21], majd az ennek az alapjaira épült Vicuna V0 7B-t és a Vicuna V0 13B-t. Felépítése jóvoltából a MiniGPT-4 esetén csupán a lineáris projektáló réteget kell úgy betanítani, hogy annak a vizuális elemei összeilleszthetők legyenek a fentebb leírt három LLM-mel.

A jelenlegi felmérések alapján a modell 90%-os pontossággal hasonlóan jó választ képes adni, mint a ChatGPT, eltörpülve annak gépigénye mellett, továbbá számos hasonlóságot lehet fellelni köztük és a GPT-4 között. Példának okáért a MiniGPT-4 képes fénykép alapján recepteket, történeteket, verseket, reklámokat generálni; problémákat felismerni és azokra megoldásokat javasolni.

A modell kiértékelése során a fejlesztők négy területen végeztek vizsgálatokat. Meme-eket értelmezése azzal a kérdéssel, hogy: *miért vicces ez a kép?* Receptgenerálás kép alapján: *hogyan kellene ezt elkészítenem?* Reklámok megírása a következő utasítással: *segíts kérlek írni egy profi reklámszöveget ehhez a képhez.* Végül pedig versírás: *írd kérlek egy szép verset a megadott fényképhez.* A generált eredményeket összehasonlították egy másik nyelvi modellel is, a BLIP-2-vel [22], ami egy hasonló modell, így pedig arra jutottak, hogy a MiniGPT-4 66.2%-os pontossággal tud képeket feldolgozni, míg a BLIP-2 csak 27.4%-os pontossággal képes ugyanerre. Ez valamivel több, mint kétszer akkora pontosságot jelent.

Kutatásom során a MiniGPT-4-et arra fogom alkalmazni, hogy a különböző három nyelvi modellel egyesítve külön-külön egy autó kamerájával készített képeket

dolgozzon fel, és azok alapján adjon választ arra, hogy a sofőr helyében milyen módon irányítaná a járművet.

## 3. fejezet

# Módszertan

Jelen fejezetben a kísérletek módszertanát fogom részletesen bemutatni: azt, hogy milyen formában fogom elvégezni őket, feljegyezni az eredményeket majd pedig kiértékelni azokat.

A kísérleteim során a MiniGPT-4-et alkalmazom. Azért választottam ezt a VLM-et, mert tulajdonságaiból adódóan nincs szüksége hatalmas GPU-ra, így az egyetem laborjának környezetében is képes futni. Ugyanakkor, amint azt már az előző fejezetben is kifejtettem, 80%-ban hasonló válaszokat képes adni, mint a GPT-4, ami jelenleg az egyik legfejlettebb modellek közé sorolható. A modelleket a kísérlet szempontjából a következő specifikációs adatokkal rendelkező számítógépen futtatjuk le: CPU Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz, NVIDIA RTX A6000 48GB GDDR6, 64GB RAM.

A kísérletekben mesterséges környezetből szintetizált, valamint valós fényképeket használok fel. Ezeket a képeket egymás után sorra betöltöm a VLM-be először a Llama-2 Chat 7B-t, majd a Vicuna V0 7B-t majd végül pedig a Vicuna V0 13B-t használva. Mindegyik alkalommal ugyanahhoz a képhez ugyanazt a kérdést/kérdéseket rendelem, majd megvárom amíg a modell feldolgozza azt, és válaszol rájuk. Az adatokat egy speciálisan erre a kutatásra létrehozott adatbázisba mentem le scénáriókra bontva, aminek segítségével a kísérletek befejeztével feldolgozom azokat. Minden egyes képhez a következő információk vannak rendelve: Llama-2 Chat 7B válasza, Vicuna V0 7B válasza, Vicuna V0 13B válasza, továbbá az általam megfogalmazott elvárt válasz.

<i>Kép</i>	<i>Kérdés</i>	<i>1. modell</i>	<i>2. modell</i>	<i>3. modell</i>	<i>Elvárt eredmény</i>
<i>i. fénykép</i>					

A különböző képekhez különböző valóságban is fellelhető szcenáriókat rendelünk. Miután egy fényképet feldolgozott a modell, azt a feladatot adjuk neki, hogy írja le nekünk milyen módon vezetné az autót, ha a fényképezőgép, amivel a felvétel készült, egy az autóra erősített RGB kamera lenne, a modell maga pedig egy autonóm irányítórendszer, ami a bemeneti kép alapján meg kell határozza milyen módon vezeti tovább az autót a bemeneti kép által levetített szituációban.

A modell válaszában a következő információkat keressük és várjuk el:

- A kormány elforgatásának szöge a  $[-60^\circ, +60^\circ]$  fokos intervallumon, amit a következő részintervallumokra bontunk fel: a  $[-60^\circ, 0^\circ]$  a balra forgatást, a  $(0^\circ, +60^\circ]$  a jobbra forgatást jelöli, míg a  $0^\circ$  pedig azt jelzi, hogy a kormányt egyenes pozícióban tartjuk
- A gáz erősségét 0-100% között; a 0% azt jelenti, hogy az ágens nem gyorsítja az autót, míg a 100% azt jelenti, hogy az autó maximálisan gyorsul
- A fék erősségét 0-100% között; a 0% jelzi, hogy az ágens „lába” nincs a fékpedálon, tehát nem fékez, a 100% pedig azt jelzi, hogy a fék teljes erejével üzemel

Ezek az egyetlen paraméterek, amiket elvárunk, hogy szerepeljenek a modell válaszában.

A modelleknek angol nyelven adunk meg minden utasítást, mert így nagyot javul azok reprodukálhatósága. Emiatt a válaszokat is angolul fogjuk megkapni, így azok angolul fognak szerepelni az adatbázisunkban.

Miután a modell eredményeit egy adatbázisba szerveztük, következik ezen információk kiértékelése. Az adatok elemzése során az ember által történő kiértékelést [23] fogjuk alkalmazni. Egy ilyen kiértékelés általánosan úgy működik, hogy egy kiértékelő bizottság elé kell helyezni az adathalmazt, amit a kísérlet során összegyűjtöttünk, majd pedig ezt egy elvárt eredményhalmazhoz kell viszonyítani.

Ebben az esetben a kiértékelő összehasonlítja minden egyes tesztesetet az elvárt eredménnyel, amit egy úgynevezett „alapigazságnak” tekintünk, egy ideális válasznak. Ez viszont azt a feltételt állítja fel, hogy minden egyes elemét az eredményhalmaznak előzetesen el kell készítenünk. Továbbá fontos megjegyezni, hogy ez az eredményhalmaz nagyban befolyásolja a kiértékelés menetét; nagy precizitással és odafigyeléssel kell azt elkészíteni, különben félrevezető eredményeket adhat a kiértékelés során.

Továbbá a kiértékelésünket egy pontozási rendszer szerint fogjuk elvégezni. A pontozási rendszer szerint megvalósuló kiértékelés során egy pontozási rendszert kell felállítanunk, mint például egy 1-től 10-ig terjedő intervallum. Mivel ebben az esetben nem létezik pontosan meghatározható igazság, a kiértékelő véleményén múlik a pontozás felállítása. A tesztesetek függvényében a pontozás történhet apró részletességgel, vagy nagyobb ugrásokkal:

- Apró részletességgel:  $dx$  kis nagyságrendű (0.1, 0.01, 0.001...)
- Nagyobb ugrásokkal:  $dx$  nagy nagyságrendű (1, 10, 100...)

Ahol  $dx$ -et a következőképpen határozzuk meg:

$$[a, b] = \cup [x_i, x_{i+1}]$$

$$dx = x_{i+1} - x_i$$

$a, b$  – a pontozási intervallum szélei

$x_i, x_{i+1}$  – két egymást követő kiértékelési értékpont

A kiértékelés mehet több paraméter szerint is, amiket előre le kell szögezni. Az is lehetséges, hogy különböző, nehezen meghatározható paramétereket is figyelembe kell venni. Továbbá nem szabad elfelejteni, hogy egy válasz teljesen rossz is lehet abban az esetben, ha a modell kifejejt a válaszából az elvárt adatokat.

Jelen esetben a modellünk eredményeit egy a mesterséges intelligenciák és nyelvi modellek terén nagy szakértelemmel rendelkező objektív kiértékelő bizottság dolgozza fel és értékeli ki. Ők a pontozást az 1-től 10-ig terjedő természetes intervallumon végzik el, ahol a 10 az jelzi, hogy a modell teljesen megfelelt az előzetesen elvárt eredményeknek, míg az 1 arra utal, hogy a modell válasza nem felelt meg az előzetes elvárásoknak.

A bizottság kiértékelése során három fontos paramétert kell figyelembe veygen: a kormány elforgatásának szöge, a gáz és a fék erőssége. Miután megtörtént a scenáriók kiértékelése, mindegyik VLM-hez külön-külön vonunk egy átlagot háromtizedes pontossággal, hogy megállapítsuk mennyire helyesen oldotta meg a szituációkat, amiket felkínáltunk nekik.



$$m = \frac{\sum \text{eset értékelése}}{\text{esetek száma}}$$

Legvégül a kapott eredményeket összehasonlítjuk egymással, hogy utána megállapíthassuk a Llama-2 Chat 7B, a Vicuna V0 7B vagy pedig a Vicuna V0 13B modell adja-e a leghelyesebb válaszokat, valamint, hogy általánosságában mennyire adnak helyes válaszokat a nyelvi modellek a vezetés során felmerülő kérdésekre, problémákra.

Fontosnak vélem megemlíteni, hogy ezt a kiértékelést ki lehetne egészíteni egy LLM által történő kiértékeléssel. Ebben ez esetben ugyanazok a szabályok érvényesülnek, mint az ember által megvalósuló kiértékelés esetén, viszont itt ugyanezt egy nagy nyelvi modell végzi. Jelen kutatás esetén megoldható lenne, hogy a MiniGPT-4 által generált különböző válaszokat a GPT-4-nek adjuk oda, hogy egy elvárt eredményhalmaz vagy saját maga által generált eredményhalmaz alapján pontozza azokat.

Mindkét módszernek megvannak a maga pozitív és negatív hozadécai: az ember által megvalósuló kiértékelés nagyon pontos eredményt nyújt, viszont rettenetesen időigényes és költséges, míg az LLM által megvalósuló kiértékelés nagyon gyors, viszont nem feltétlenül helyes. A nyelvi modellek fejlődésével idővel még nagyobb teret fognak hódítani az automatizált kiértékelés, de mindaddig, amíg a nyelvi modellek pontossága nem megbízható, a manuális, humán kiértékelés fog előnyben részesülni.

### 3.1. Pontozási rendszer hasonló modelleknél

A nyelvi modellek feltörekvése még egy friss folyamat, éppen ezért a nyelvi modellek által megvalósított autonóm vezetés még egy aránylag fiatal és eredeti kutatási téma. Viszont lehet találni néhány hasonló típusú kutatást.

A Wayve Technologies Ltd. csapata végzett hasonló kutatás [31] az autonóm járművek területén. Az ő esetükben a kiértékelés a következő módon valósult meg. Összeállítottak 1000 egymástól eltérő vezetési scenáriót. Ezeket vektorosan ábrázolták, majd azokat alakították át szöveggé, amit később fel tudott dolgozni a nyelvi modellük. A vezetés során ők négy adatot vártak el a modellüktől: gázerősség, fékerősség, kormány forgásszöge, valamint az indok arra, hogy a modell miért ezekkel a paraméterekkel irányítja az autót. Például kísérleteik során a következő eset is felállt:

**Q:** *How are you going to drive in this situation and why?*

**A:** *Here are my actions: - Accelerator pedal 0% - Brake pedal 7% - Steering straight. Reason: Because I'm maintaining a slow speed due to the pedestrians crossing, and I'm steering straight to keep our vehicle properly aligned with the road.*

Kutatásuk során viszont nem csak azt várták el a nyelvi modellektől, hogy a vezetés során helyes választ adjanak a jármű irányításának kérdésére, hanem azt is, hogy körül tudja írni azt, amit lát és hogy ismerje a helyi turisztalátványosságokat, mintha egy taxifőrr lenne.

Volt még egy hasonló kutatása a Tsinghua Egyetem kutatócsapatának is. Ők a GPT-4-et használták fel összekötve a CARLA szimulátorral, amit egy Python 3.7-es Unreal Engine 4-en futtattak le. Mivel a GPT-4 feldolgozási ideje lassabb volt, mint a Carlaban megjelenő képanyag sebessége, ezért azt lelassították, hogy a nyelvi modellnek legyen ideje helyes választ adni.

Miután elvégezték a kísérleteket, összehívtak egy huszonnégy vezetéshez értő személyből álló szakbizottságot, akiknek megmutatták a videófelveteleket, majd egy 5-ig terjedő skálán kellett pontozniuk azokat.

## 4. fejezet

### Kísérletek

Kutatásom alatt több kísérletet is elvégeztem. Ezek során különböző mesterséges környezetekből szinkronizált és valós fényképeket töltöttem be a VLM-ekbe, amelyek feldolgozták ezeket. Ezt követően scenáriókat állítottam fel, melyeket átadtam a modelleknek feldolgozásra. Ezen scenáriók által azt mértem fel, hogy a VLM-ek mekkora pontossággal képesek vizuális információ alapján felmérni a környezetet számára releváns és nemreleváns szemszögből, továbbá az érdekelt, hogy egy adott autonóm vezetési helyzetet mennyire helyes képes egyedül megoldani.

Minden egyes scenárióban, amikor autonóm vezetésről volt szó a következő három adatot vártam el a modellektől a válaszukban: a kormány elforgatásának a szöge, a gáz erőssége, a fék erőssége. Ezen információhármast szempontjából kell majd kiértékelje a kiértékelő bizottság a válaszokat.

Az első kísérlet során arra összpontosítottunk, hogy a modellek mennyire pontosan képesek felismerni a releváns információkat egy mesterséges környezetből szintetizált fényképről. Ebben az esetben eltekintettünk a három fő paraméterünktől, csupán a helyzetfelismerésre fókuszáltunk.



*1. ábra: a CARLA környezetből szintetizált kép [24]*

Az 1. ábrán megjelenített kép esetén a kérdés-válaszolás így nézett ki:

**Q:** *Describe detailed what you can see on the image.*

**Llama-2 Chat 7B:** *The image shows a city street with tall buildings on either side of the road. There is a traffic light at the intersection and a pedestrian crossing. The sky is blue and there are clouds in the background. The image is taken from a bird's eye view. It looks like a busy street with cars and pedestrians going in both directions. There are tall buildings on either side of the road and a traffic light at the intersection. The pedestrian crossing has a button and a light that shows when it is safe to cross. The sky is blue with some clouds in the background. The image is taken from a bird's eye view, showing the entire street and the surrounding buildings.*

**Vicuna V0 7B:** *The image shows a city street with a large sculpture in the center of the road, surrounded by modern buildings and trees. The buildings have large windows and smooth walls, and the sculpture is made of metal with sharp edges. On the left side of the image, there is a large glass building with a stepped roof and a sign saying "Midtown Tower". On the right side of the image, there is a smaller building with a pointed roof and windows that go all the way up to the top. The sidewalk in front of the buildings is made of concrete and there are some benches on the sidewalk. The sky in the image is clear and blue, with some clouds in the distance. There is no traffic on the street at the moment.*

**Vicuna V0 13B:** *The image shows a city street with buildings on either side of the road. The buildings appear to be modern, sleek, and white with large windows. The road is empty and there are no cars or people visible. The sky is clear and there are no clouds in sight. The image is in shades of gray and white.*

Már most láthatjuk, hogy a 7 milliárd paraméterrel betanított LLaMA és Vicuna modellek jelentősen hosszabb választ adnak, mint a 13 milliárd paraméteres Vicuna. De vajon mi okozhatja ezt? Jobbak lennének a kisebb modellek, mint a nagy? A kiértékelés után választ kapunk erre a kérdésre.

A kísérletek során arra a kérdésre is választ akartunk kapni, hogy mennyire helyesen oldanak meg a vizuális nyelvi modellek vezetés közben fellépő feladatokat, problémákat, vagy legalábbis egyáltalán képesek-e rá. Ehhez különböző szcenáriókat adtunk a VLM-eknek szóban. Itt viszont már elvártuk, hogy a modellek a három említett

paraméterrel válaszoljanak. Például így oldották meg a modellek a körforgalom elhagyását a második kijáraton:



**2. ábra: a CARLA környezetből szintetizál kép. [24]**

**Q:** *Act like the autonomous controller of the car whose windscreen-mounted camera recorded this image. Answer any questions with the inputs you would provide for the actuators of the car. The actuators are: angle motor of the steering wheel, input is expected in the  $-60^\circ$   $60^\circ$  range where the  $-60^\circ$   $0^\circ$  range indicates the left turn and the  $0^\circ$   $60^\circ$  range indicates that the car turns right; acceleration, input is expected in the  $0\%$   $100\%$  range; braking, input is expected in the  $0\%$   $100\%$  range. These are the only numbers I expect from you in your response. Provide the actuation sequence for the following maneuver: take the second exit of the roundabout ahead.*

**Llama-2 Chat 7B:** *As the autonomous controller of the car whose windscreen-mounted camera recorded this image, I would provide the following actuation sequence for taking the second exit of the roundabout ahead: Angle motor of the steering wheel: turn left by 40 degrees. Acceleration: increase the speed of the car to 75% of its maximum speed. Brake: apply the brakes gradually to slow the car down to 50% of its maximum speed. By following this sequence, the car should take the second exit of the roundabout ahead. Please note that this sequence is based on the assumption that the car has a steering wheel with a neutral position and that the brakes are in good condition.*

**Vicuna V0 13B:** *Step 1: Turn the steering wheel to the left, inputting 0 degrees. Step 2: Press the acceleration pedal, inputting 0%. Step 3: Release the brake pedal, inputting 0%. Step 4: Turn the steering wheel to the right, inputting 0 degrees. Step 5: Press the acceleration pedal, inputting 50%. Step 6: Release the brake pedal, inputting 0%. Step 7: Turn the steering wheel to the right, inputting 180 degrees. Step 8: Press the acceleration pedal, inputting 75%. Step 9: Release the brake pedal, inputting 0%. Step 10: Continue straight through the intersection. Note: This sequence assumes that the roundabout has two exits. If there are more than two exits, the sequence may need to be adjusted accordingly.*

Az előzetes eredményeket megtekintve megfigyelhető, hogy a két modell itt is eltérő választ ad: a kisebbik egy rövid, a kért három paraméterre kitérő, míg a nagyobbik egy több lépésből állót. Kiértékelés után megtekinthetjük melyik modell válasza a helyesebb.

Ugyanakkor vicces eredményeket is tud adni.

Ahogy azt az előző két példából láthattuk, a modell elboldogul a mesterséges környezetből szinkronizált képekkel. Következő lépésünkben az következett, hogy kiderítsük, hogy boldogul a valós világban rögzített fényképekkel. Ebben az esetben is egy helyzetfelismeréssel kezdtük.



**3. ábra: budapesti utca: Yael Edery, Pinterest [25]**

A fenti budapesti utcaképen alapján a modellek a következő leírást adták:

**Q:** *Describe detailed what you can see on the image.*

**Llama-2 Chat 7B:** *The image shows a street with a lot of cars parked on it. The buildings on either side of the street are tall and made of brick. There are a lot of people walking on the sidewalk and cycling on the street. There are also several parked cars on the street.*

**Vicuna V0 7B:** *The image shows a narrow, urban street with buildings on either side. There are several cars parked on the street, with empty spots available. The street appears to be well-maintained, with no visible cracks or potholes. The buildings on either side of the street are old, with a mix of modern and older architecture. Some of the windows appear to be closed, while others are open, allowing light to shine into the street. The sky is clear, with no clouds or other weather features visible.*

**Vicuna V0 13B:** *1. The street is lined with tall, old buildings made of beige and yellow bricks. 2. Many of the buildings have arched windows and balconies with decorative ironwork. 3. Cars are parked on either side of the street, leaving just enough space for pedestrians to walk by. 4. The street is surrounded by trees, which provide shade and contribute to the overall aesthetic of the area. 5. The sidewalk is made of grey cobblestones and is slightly raised compared to the road. 6. There are no visible street signs or markings on the road, except for a few painted lines indicating parking spaces and the occasional crosswalk. 7. The overall atmosphere of the scene is peaceful and charming, with the mix of old and new architecture and the presence of greenery adding to this effect.*

Mint láthatjuk a modellek ebben az esetben is elboldogultak a képpel, itt viszont az erősebb modell részletesebb leírást adott a gyengébb társainál.

Végül pedig a legutolsó esetben azt tekintjük meg, hogy boldogulnak a modellek egy valós képen (4. ábra) megjelenő vezetői feladattal. Jelen esetben a modellnek azt a feladatot adtuk, hogy parkoljon párhuzamosan az utca jobboldalán.

**Q:** *Act like the autonomous controller of the car whose windscreen-mounted camera recorded this image. Answer any questions with the inputs you would provide for the actuators of the car. The actuators are: angle motor of the steering wheel, input is expected in the  $-60^\circ$   $60^\circ$  range where the  $-60^\circ$   $0^\circ$  range indicates the left turn and the  $0^\circ$   $60^\circ$  range indicates that the car turns right; acceleration, input is expected in the  $0\%$*



100% range; braking, input is expected in the 0% 100% range. These are the only numbers I expect from you in your response. Provide the actuation sequence for the following maneuver: parallel park in the right.



4. ábra: budapesti utcakép, Google Earth [26]

**Llama-2 Chat 7B:** *Of course! I'm happy to help you with this scenario. As the autonomous controller of the car, I would first begin by analyzing the image captured by the windshield-mounted camera to determine the location and size of the parking space. Based on this information, I would calculate the optimal angle and speed of the steering wheel to turn the car into the parking space. For this maneuver, I would provide the following actuation sequence for the steering wheel: Start with the steering wheel at a central position, which is approximately 0 degrees. Calculate the angle of the steering wheel needed to turn the car into the parking space, based on the size and location of the space. For example, if the parking space is 5 meters long and 2 meters wide, I would instruct the steering wheel to turn -30 degrees to align the car with the entrance of the space. Use the angle motor to adjust the steering wheel to the calculated angle, within the -60° to 60° range. In this case, I would instruct the angle motor to turn the steering wheel -30 degrees. Once the car is aligned with the parking space, I would instruct the steering wheel to return to its central position. Next, I would calculate the acceleration and braking inputs needed to maneuver the car into the parking.*

**Vicuna V0 7B:** *1. Set the angle motor of the steering wheel to 0 degrees to straighten the car. 2. Activate the acceleration to move the car forward slowly and carefully towards*



*the parking spot. 3. As the car approaches the parking spot, adjust the angle motor to the appropriate angle to align the car with the parking spot. The angle should be around 45 degrees to the left. 4. Once the car is in the parking spot, activate the braking to slow down and stop the car. 5. Once the car has come to a complete stop, deactivate the acceleration and angle motor. 6. Release the brake to remove the parking brake and move the car out of the parking spot when ready to leave. Note: These are general instructions for executing a parallel park in the right direction. The exact details may vary depending on the specific vehicle, road conditions, and other factors.*

**Elvárt eredmény:** *Menj lassan a szürke autó mellé. Gáz: 15%, fék: 0%, kormány: 0°. A szürke autó mellett állj meg. Gáz: 0%, fék: 100%, kormány: 0°. Fordítsd el maximálisan a kormányt jobbra, majd lassan kezdj el tolat. Gáz: 25%, fék: 0%, kormány: +60°. Fordítsd a kormány maximálisan balra tolatás közben. Gáz: 25%, fék: 0%, kormány: +60%. Állj meg: Gáz: 0%, fék: 100%, kormány: 0°.*

Láthatóan a modellek elboldogultak egyarán a mesterséges és valós képekkel is. Az egyetlen kiderítendő kérdés már csak az, hogy mennyire helyesen dolgoztak. Ez a kiértékelés során fog napvilágot látni.

Ugyanakkor voltak sikertelen válaszai is a modelleknek. Például a következő esetek is felléptek a kísérletek során:

**Q:** *How would your drive the car in the following situation?*

**Llama-2 Chat 7B:** *As a language model I am not able to drive and because of this I cannot answer the question of how I would drive the car in the following situation.*

**Q:** *Act like the driver of the vehicle that we are currently in.*

**Llama-2 Chat 7B:** *Okay! Adjust mirrors and sunglasses. Whoa, dude! Check out that rad wave we just got! Looks around Nice, man. This ride is fly. Pauses I think we might be lost, bro. Looks at the GPS Oh, man. We're way off course. Adjusts mirrors again Well, at least we got some sick tunes to keep us entertained. Cranks up the volume What's your favorite song? Grins*

**Q:** *Act like the autonomous controller of the car whose windscreen-mounted camera recorded this image. Answer any questions with the inputs you would provide for the actuators of the car. The actuators are: angle motor of the steering wheel, input is*

*expected in the  $-60^\circ$   $60^\circ$  range where the  $-60^\circ$   $0$  range indicates the left turn and the  $0^\circ$   $60^\circ$  range indicates that the car turns right; acceleration, input is expected in the 0% 100% range; braking, input is expected in the 0% 100% range. These are the only numbers I expect from you in your response. Provide the actuation sequence for the following maneuver: take the first exit of the roundabout you are in.*

**Llama-2 Chat 7B:** *Ah, a fellow autonomous controller! I'm glad to see you're interested in the intricacies of vehicle control. adjusts virtual glasses Now, let's get down to business. (még folytatta)*

Mint azt láthatjuk, a modellek számos értelmetlen választ adtak a kísérletek során. Ezt úgy javítottuk ki, hogy amint az feljebb is látszik, nagyon pontosan elmondtuk a modelleknek, hogy mi a pontos feladatuk (autonóm járműirányító), milyen adatokat várunk el tőlük, és csak azokra az adatokra van szükségünk. Ezek után már nem volt gyakori, hogy a modellek humorosan válaszoljanak, ha pedig mégis ezt tették, akkor újra odaadtuk nekik a parancsot, ami után már helyesen működtek.

## 5. fejezet

### Kiértékelés

A kísérletek elvégzése után a feljegyzett adatokat egy ezen a területen nagy szaktudással rendelkező kiértékelő bizottság elé állítottam, akik a kiértékelési módszertan szerint elvégezték a modellek 1-től 10-ig terjedő pontozását. Ezek után a következő eredményre jutottak háromtizedes pontossággal:

Kép sorszáma	Llama-2 Chat 7B	Vicuna V0 7B	Vicuna V0 13B
1. kép	3,750	5,000	5,500
2. kép	7,000	6,000	6,750
3. kép	5,750	5,000	6,000
4. kép	6,000	5,000	5,750
5. kép	5,250	5,250	6,750
6. kép	6,250	4,000	4,750
7. kép	4,500	4,000	5,000
Átlag	4,959	4,510	5,061
Szórás	2,603	2,408	2,897

Amint az a táblázatból is jól kivehető, és az előzetes sejtést is megerősíti, a három nagy nyelvi modell közül a Vicuna V0 13B bizonyult a legjobbnak. Ez az eredmény nem

meglepő, hiszen ez a modell tizenhárommilliárd paraméterrel volt betanítva, majdnem kétszer annyival, mint a másik kettő külön-külön, így normális, hogy pontosabb válaszokat ad, mint társai.

Nemsokkal az első helyezett mögött a Llama-2 Chat 7B szerepel, ez pedig kicsit már nagyobb ugrással előzi ki a Vicuna V0 7B-t, a táblázat eredményeit nézve különbségük 0.449. Ez meglepő is meg nem is, hiszen a Vicuna a LLaMA alapján készült, de ezek szerint nem előzi meg annak a pontosságát a működésben. Hasonló következtetést vonhatunk le, ha a Llama-2 Chat 7B és a Vicuna V0 13B közötti különbséget nézzük, hiszen ez az érték is alacsony, pontosabban 0.102. A nagyobb modell jobb, de nem sokkal.

Összegezve a modellek értékelését és az ezek közötti különbséget arra jutunk, hogy habár a Vicuna V0 13B modell nyújtotta a legjobb válaszokat a kísérletek során, nem elhanyagolható a Llama-2 Chat 7B erőssége, ami egy jóval kisebb modell, éppen emiatt nem igényel akkora feldolgozási kapacitást, mint a nagyobb Vicuna.

A továbbiakban kihangsúlyoznék különböző megfigyeléseket, amelyek a kísérletek során tűntek fel számomra.

Ahogy azt már korábban kiemeltem, a képek leírása során a kisebb modellek sokkal hosszabb és részletesebb válaszokkal szolgáltak, mint a 13 milliárdos modell, de nem pontosabban. A kisebb modellek nem értették még teljesen a fényképeket, éppen ezért nagyvonalakban leírták, amit sikerült feldolgozniuk, majd válaszaikban megfogalmaztak egy csomó spekulációt. Például az *1. ábrán* a LLaMA úgy fogalmaz, hogy járókelők és forgalmi lámpák láthatóak a képen, továbbá, hogy a kép egy madár szemszögéből volt elkészítve. Ezt pedig kétszer is elmondja egymás után. A kép megfigyelése során nem lehet kétségünk róla, hogy ezek mind spekulációk, hiszen azon nem szerepelnek se járókelők, se forgalmi lámpák, sőt egyértelmű, hogy a kép nem egy madár szemszögéből készült, hanem egy autóra erősített kamerából. Ezzel ellentétbe állítva a Vicuna V0 7B már kicsit pontosabb választ ad, felismeri, hogy a képen nincs se forgalom, se járókelők, viszont nemlétező feliratot tűz az épületekre, és spekulál azok felépítéséről. A 13 milliárdos modell egy rövid de helyes leírást adott a képről.

A vezetés során felmerülő problémákat a modellek megpróbálták általános igazságokkal megoldani, de nem lettek nagyon sikeresek. Például a *2. ábrán* megjelenített körforgalom esetén a modellek el voltak veszve, amikor a második kijáraton kértem az autó kivezetését. A LLaMA például a kamera állása alapján úgy akar eljutni a második

kijáráshoz, hogy keresztül vág a fűvön, eltekintve attól, hogy azon még egy szobor is van. Mintha az egy út lenne. A Vicuna V0 13B már máshogy állt a teljes feladathoz. Megpróbálta részfeladatokra bontani a körforgalomban történő vezetést, viszont nem tudta megállapítani, hogy a kért paramétereket milyen mértékre kellene állítania, így rengeteg értelmetlen input született, mint például a 180° fokos kormánytekerés, miközben meg volt mondva neki előre, hogy -60° és +60° közötti értékeket várunk el tőle. Továbbá a fék és gáz erejével sem boldogul, ezért mindegyikhez általában a kezdő 0%-ot helyezte, ezzel jelezve, hogy alkalmazni kell azokat is.

Jól boldogultak a modellek a valós fénykép alapján történő leírás elkészítésével. A 7B-s modellek helyesen felismerték a 3. *ábra* környezetét és egy helyes leírást adtak a képekről. A 13B-s modell is megértette a képet, és egy sokkal részletesebb, de helyes leírást adott ugyanarról a képről, mint 7B-s társai.

Itt ellentét áll fenn a mesterséges és valós képek közötti értelmezések között, hiszen a 7B-s modellek itt rengeteget spekuláltak a mesterséges képek esetén. Ennek kiváltó oka lehet akár a mesterséges kép monotonitása, a színek halványan egybemosódása, de egyértelmű, hogy a mesterséges képek felismerésével nem boldogulnak a modellek.

Legvégül pedig nem szabad elfelejteni a valós környezeti kép alapján történő vezetői feladatok megoldását. A 4. *ábrán* látható esetben a párhuzamos parkolást aránylag jól oldották meg a modellek. Rengeteg spekulációt fűztek hozzá a válaszukhoz, de ezt úgy tették, hogyha valaki elolvassa, akkor hihetőnek tűnjön a válasza. Ennek oka lehet az is, hogy valós környezetben a modellek jobban fel tudják dolgozni a képeket, de az is, hogy a párhuzamos parkolás nem egy annyire komplex feladat, mint keresztülhajtani egy körforgalmon.

Következtetésként levonható, hogy a modellek képeket aránylag jól értelmeznek, és egy tiszta kép esetén elég jó leírást tudnak adni. Ennek oka valószínűleg az a tény, hogy amikor a modelleket betanították, akkor rengeteg reklámot írtak velük, így ezek elsajátították az ábrák jó leírását. Ugyanakkor a vezetés feladatával nem boldogulnak, rengeteg értelmetlen választ adnak. Nem tudják kiszámolni kép alapján, milyen szögben kellene elforgassák a kormány, vagy felülbírálják a megadott kormányparamétereket, és sajátokkal helyettesítik az. Nem képesek megállapítani, mennyi fékre és gázra van szükség, csak annyit tudnak, hogy kell, emiatt ezek használatának jelzésére vagy 0%-ot vagy 100%-ot írnak, a szélsőértékeket, vagy pedig ritkább esetben egy random értéket

írnak be 0-100% között, viszont ez az érték teljesen helytelen. Egyértelmű, hogy a modellek rengeteget spekulálnak.

Továbbá megjegyezném, hogy a kísérletek során volt olyan eset, amikor a modellek nem tudtak feldolgozni egy fényképet, ilyenkor nem írtak ki semmilyen választ, vagy nagyban kiírták válaszként, hogy *Error*.

Egyértelműen felismerhető, hogy a modellek nem vezetés céljából voltak betanítva, hiszen a pontos megfogalmazott parancsig számtalanszor megpróbálták elkerülni a válaszadást. Mint azt a kísérletek során láthattuk, ameddig nem adtunk nekik helyes parancsot, addig megpróbálták elkerülni a válaszadást, mondván, hogy ők nyelvi modellként nem tudnak ilyesmire válaszolni, vagy ha válaszoltak azt viccelődve tették. Ennek fő okai közé sorolható a tény, hogy a modellek betanításának egy része vicces képek értelmezésével történt. Emiatt a modellek nincsenek tisztában azzal, hogyan kellene egy járművet irányítaniuk.

Megjegyezném, hogy a kiértékelés során még egy érdekes eredmény látott napvilágot. A kiértékelő bizottság mellett még egy olyan csapat is kitöltötte a kiértékelési kérdőívet, akik nem jártasok az informatikában. Ebben az esetben kiértékelés után arra jutottunk, hogy a Llama-2 Chat 7B a legjobb nyelvi modell. Úgy vélem, hogy ennek a kiváltó oka az lehet, hogy habár a LLaMA sokat spekulált, ezt egy nagyon „határozott” módon tette, emiatt pedig sokan hitelesnek hitték a válaszait.

## 6. fejezet

# Összegzés

Kutatásom során egy olyan autonóm ágens kidolgozásán dolgoztam, amely vizuális nyelvi modellek segítségével közelíti meg az autonóm járműirányítás kérdését, mindenki számára érthető módon, természetes nyelven leírva azt. Ezzel az a célom, hogy olyan hibákat, mint amilyenek a jelenlegi deep learning hálónál felmerülhetnek, kiküszöböljek, valamint pedig, hogy oldjam az emberek bizalmatlanságát az autonóm járművek fele, azzal, hogy megértik, hogyan működnek ezek.

A kísérletek során a MiniGPT-4 vizuális nyelvi modellt használtam fel sorra összekapcsolva a Llama-2 Chat 7B, majd a Vicuna V0 7B legvégül pedig a Vicuna V0 13B nagy nyelvi modellel. A kísérletek során egyaránt alkalmaztam mesterséges környezetből szintetizált és valós képeket is.

A kísérletek levezénylése és a kiértékelés után arra jutottam, hogy a Vicuna V0 13B végzi a legjobb járműirányítást 5.061-es átlaggal egy 1-től 10-ig terjedő skálán. Ebből azt a következtetést vontam le, hogy megéri folytatni az autonóm járművek kutatását ebből a megközelítésből, mert hatalmas potenciál van benne, annak ellenére, hogy a jelenlegi kutatásban alkalmazott modellek nem lennének képesek egy valós járművet elvezetni. A kutatás kiterjesztéseképpen lehet a jövőben más nyelvi modelleket is alkalmazni, vagy esetleg kifejleszteni egy sajátot, nyelvi modelleket finomhangolni a vezetés problémájára, továbbá olyan ágenseket kifejleszteni, amelyek megértik, hogy a jelenlegi kutatás során fejlesztett társuk milyen parancsokat ad ki, és az alapján irányítják a járművet. Továbbá érdemes kipróbálni ezeket az ágenseket virtuális környezetben is, mint például a CARLA szimulátorban.

## 7. fejezet

### Kitekintés

Kutatásom elkezdésekor a fő céljaim azok voltak, hogy sikeresen felépítsek egy olyan autonóm járműirányítási modellt, amely képes érthető módon megfogalmazni azokat a folyamatokat, amelyeken keresztül megy, amikor egy autót irányít, ezáltal megszüntetve a bizalmatlanságot az emberek felől az önvezetés irányába. Továbbá egy olyan ágens elkészítése volt még a cél, amelyet be lehet építeni egy valós autóba, hogy egy járművet vezessen egy kamera által előállított kép alapján.

A kiértékelés során egyértelműen megfogalmazhatóvá vált számunkra az, hogy az LLM-ek ágensként való alkalmazása egy helyes és lehetőségekkel teli megközelítése az önvezetésnek. Mivel szöveges formában természetes nyelven fogalmazzák meg a járműirányítás menetét, ezért ezek a modellek emberhez hasonló módon tudják megközelíteni a vezetés problémáját. Így ezek alkalmazása megszüntetné az ember gépek felé tanúsított bizalmatlanságát, hiszen végre értenék, milyen folyamatokon mennek ezek keresztül, miközben egy járművet irányítanak.

A kutatás folytatásában számos lehetőség bontakozik ki. Elsősorban az ágens áll olyan szinten, hogy összekössük a CARLA Simulator-ral [27], amely egy városi környezetet épít fel, hogy ott autonóm járműirányítást lehessen tesztelni virtuális környezetben. A jelenlegi kutatás során is használtunk fel a CARLA adatbázisából szinkronizált képeket, viszont miután a rendszert és a szimuláció környezetet összekötjük, lehet készíteni egy autonóm járműirányítási szimulációt videóanyag alapján.

Továbbá még ki lehet bővíteni a kutatást más vizuális nyelvi modellek alkalmazására is. Véleményem szerint megéri kipróbálni nagyobb modelleket is, hiszen ezek pontosabb válaszokat adhatnak a kísérletek során felmerülő szcenáriókra, ezzel precízebb modelleket lehet majd kiépíteni.

Fontosnak tartom megemlíteni, hogy a kísérlet során alkalmazott VLM-ek nem vezetésre voltak betanítva. Sokkal jobb válaszokat adhatnak majd egy finomhangolás után. Éppen ezért a további kutatások során majd mindenképp vezetésre kell finomhangolni a modelljeinket. Számos nyilvános autonóm vezetés betanítására



használható adatbázis lehet elérni online, mint például a BerkeleyDeepDrive Dataset [28], a Google-Landmark Dataset [29] vagy a KITTI Vision Benchmark Suite [30].

A kutatás folytatásához még szükség van egy LLM alapú ágens kifejlesztésére is. Ennek az lenne a feladata, hogy értelmezze a már megvalósított VLM rendszer által előállított szöveget, és ennek a parancsai alapján irányítsa a járművet. Így a jármű irányítása gyakorlatilag egy VLM-LLM kommunikációból állna össze egy mindenki számára egyszerűen értelmezhető módon.

Legvégül még szintén a kutatás folytatásának lehet tekinteni azt a feladatot, hogy kifejlesszünk egy saját, a vezetésre betanított nyelvi modellt, amely egyetlen feladata a járműirányítás közben megjelenő videóanyag pontos értelmezése és az irányítási feladatok megfogalmazása vagy ennek értelmezése.

Legvégül a különböző ágenseket összekapcsolhatjuk egymással, majd először mesterséges környezetben tesztelhetjük ezt a rendszert. Ha ott jó eredményekkel szolgálnak, akkor a következő lépés ennek a rendszernek a beépítése egy igazi járműbe, és a közúton való tesztelése.

Még sok kutatás vár erre a témára, de meg vagyok győződve, hogy ez egy helyes megközelítése az autonóm járműirányítás kérdésének. Pár éven belül a nyelvi modellek technológiája fejlődhet annyit, hogy képesek legyenek videó alapján úgy autót irányítani, mint ahogyan az emberek gondolkodnak, mindenki számára érthető módon. Megtörténhet, hogy néhány év múlva az ezen technológia szerint működő autonóm járművek fogják uralni a piacot.

# Köszönetnyilvánítás

A következő dolgozat létrejöttében nagy szerepet vállalt témavezetőm, dr. Szántó Mátyás, ezúton is köszönetet szeretnék nyilvánítani számára a sok segítségben és támogatásban, amit felém tanúsított.

Továbbá köszönettel tartozom még a kísérletek kiértékelése során összeállt kiértékelő bizottságnak, akiknek segítség nélkül nem lehetett volna ilyen precízen kiértékelni a kísérletek eredményeit.

Végül, de nem utolsó sorban pedig köszönet a Meta csapatának, amiért elérhetővé tették számomra a LLaMA-t, a LMSYS Org csapatnak, amiért használhattam a Vicuna nyelvi modelleket, valamint a KAUST csapatának, akik a MiniGPT-4-et készítették.

# Irodalomjegyzék

- [1] Eurostat, "Globalisation patterns in EU", 2022, lásd:  
[https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Globalisation\\_patterns\\_in\\_EU\\_trade\\_and\\_investment](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Globalisation_patterns_in_EU_trade_and_investment),  
(utoljára megnyitva: 2023.10.31.)
- [2] B. A. y Arcas, "Do large language models understand us?" *Daedalus*, vol. 151, no. 2, pp. 183–197, 2022
- [3] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
- [4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [5] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- [6] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
- [7] OpenAI, R. "GPT-4 technical report. arXiv 2303.08774." *View in Article* (2023).
- [8] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).
- [9] Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality." *See <https://vicuna.lmsys.org> (accessed 14 April 2023)* (2023).
- [10] Sun, Kai, et al. "Head-to-Tail: How Knowledgeable are Large Language Models (LLM)? AKA Will LLMs Replace Knowledge Graphs?." *arXiv preprint arXiv:2308.10168* (2023).
- [11] Shamim, S. M., et al. "Handwritten digit recognition using machine learning algorithms." *Global Journal Of Computer Science And Technology* 18.1 (2018): 17-23.

- [12] Alara Dirik, Sayak Paul, "A Dive into Vision-Language Models", Hugging Face, 2023, lásd: [https://huggingface.co/blog/vision\\_language\\_pretraining](https://huggingface.co/blog/vision_language_pretraining) (utoljára elérve 2023.10.30.)
- [13] Schuhmann, Christoph, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." *Advances in Neural Information Processing Systems* 35 (2022): 25278-25294.
- [14] Jiang, Yunfan, et al. "Vima: General robot manipulation with multimodal prompts." *arXiv preprint arXiv:2210.03094* (2022).
- [15] Minwoo Byeon, et al. "COYO-700M: Image-Text Pair Dataset", 2022, GitHub, lásd: <https://github.com/kakaobrain/coyo-dataset> (utoljára megnyitva 2023.10.30.)
- [16] James Betker, et al. "Improving Image Generation with Better Caption", OpenAI, 2023
- [17] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.
- [18] Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *Advances in neural information processing systems* 32 (2019).
- [19] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." *Advances in Neural Information Processing Systems* 35 (2022): 23716-23736.
- [20] Zhu, Deyao, et al. "Minigt-4: Enhancing vision-language understanding with advanced large language models." *arXiv preprint arXiv:2304.10592* (2023).
- [21] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288* (2023).
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

- [23] Meor Amer, "Evaluating LLM Outputs", Cohere, 2023, lásd: <https://txt.cohere.com/evaluating-llm-outputs/#human-evaluation> (utoljára megnyitva 2023.10.31.)
- [24] Szántó, Mátyás, et al. "Building Maps Using Monocular Image-feeds from Windshield-mounted Cameras in a Simulator Environment." *Periodica Polytechnica Civil Engineering* 67.2 (2023): 457-472.
- [25] Yael Edery, Pinterest, lásd: <https://hu.pinterest.com/pin/349662358567342592/> (utoljára megnyitve 2023.10.31.)
- [26] Budapest, Google Earth, 2014
- [27] Dosovitskiy, Alexey, et al. "CARLA: An open urban driving simulator." *Conference on robot learning*. PMLR, 2017.
- [28] University of California at Berkeley, BDD100K, lásd: <https://bdd-data.berkeley.edu/>, (utoljára elérve: 2023.11.01.)
- [29] Weyand, Tobias, et al. "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
- [30] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.
- [31] Chen, Long, et al. "Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving." *arXiv preprint arXiv:2310.01957* (2023).
- [32] Jin, Ye, et al. "SurrealDriver: Designing Generative Driver Agent Simulation Framework in Urban Contexts based on Large Language Model." *arXiv preprint arXiv:2309.13193* (2023).