Face2Profile: A Face-and-URL Dataset for Open-Ended Profile Construction

Md. Mohibur Rahman Nabil[1], Afsara Waziha[1], Md. Nahin Alam[1], Md. Shakib Shahariar Junayed[1], Md. Sakib Sami[1], Tamim Ishrak Sanjid[1], Md. Fahim Morshed[1], Robin Krambroeckers[2], Labib Chowdhury[1], Mohammad Ruhul Amin[3], Nabeel Mohammed[1], Shafin Rahman[1]

[1]Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
[2]Robot Bulls, Geneva, Switzerland
[3]Department of Computer and Information Sciences, Fordham University, New York, USA

[1]{mohibur.nabil, afsara.waziha, nahin.alam, shakib.junayed sakib.sami, tamim.sanjid, fahim.morshed, labib.chowdhury, nabeel.mohammed, shafin.rahman}@northsouth.edu
[2]robin@robotbulls.com, [3]mamin17@fordham.edu

Abstract

Profile construction using human faces and probable URLs (likely containing biographical and professional information) is a critical challenge in multimodal AI. It involves extracting structured, identity-specific information by integrating visual and textual modalities. Existing datasets in this domain have poor demographic diversity, lack factual information in the real world, and minimal alignment between visual identity and textual evidence. These limitations make creating and evaluating vision language models (generating personalized summaries for individuals) challenging. To address these issues, we introduce a novel dataset, Face2Profile, which has approximately 10K publicly available facial images, names of the person, professional details of the person, curated sets of positive images that contain the person's information and negative or misleading web links, and peer-reviewed human-written summaries. In addition, the dataset emphasizes demographic diversity and includes challenging visual conditions such as poor lighting, occluded faces, or nonfrontal viewpoints to reflect real-world scenarios. It includes rigorous demographic stratification and annotation to ensure diversity, factual consistency, and relevance in real-world scenarios. We benchmark generative performance by evaluating GPT-4o and DeepSeek R1 using Bilingual Evaluation Understudy (BLEU) and a novel Custom-BLEU metric that penalizes missing identity elements such as names and occupations. Our analysis shows that GPT-4o and DeepSeek-R1 produce fluent summaries but frequently omit key factual content. We further evaluate lightweight language models—Phi-3 Mini and DeepSeek-1.5B—using an Entity Coverage Score (ECS) to assess the factual precision of structured output summaries by Small Language Models. This benchmark offers a novel perspective on identity-based profile construction by evaluating the models in a zero-shot setting without any model fine-tuning or task-specific training, and establishes a challenging benchmark dataset for future research on the multimodal profile construction task.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

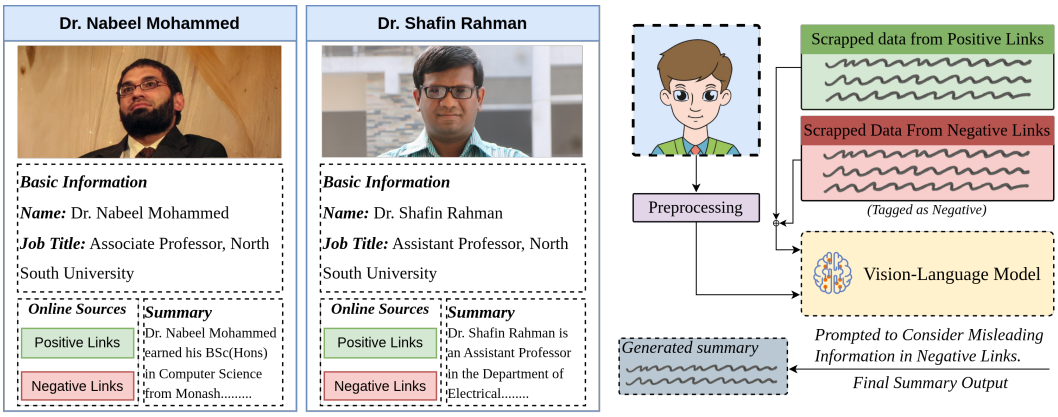


Figure 1: Illustration of our proposed dataset and approach for profile summary generation. **(a)** Given a human face image and some URLs that may contain biographical and professional information about the input face (maybe obtained from Pimeyes [3], FaceCheck [1], a Vision Language Model (VLM) can construct a profile summary of the individual. We propose a dataset comprising over 10,000 samples to train/validate/test such VLMs. **(b)** Overview of the proposed framework for bias-aware summary generation. The system scrapes both positive (accurate) and negative (misleading or irrelevant) online sources, preprocesses the content, and uses prompts to instruct the model to account for the nature of each source. This enables the generation of concise, contextually aware summaries.

1 Introduction

In an increasingly interconnected world, the ability to identify individuals from facial images and retrieve contextual professional information holds significant value across diverse applications. These include law enforcement, professional networking, recruitment, event personalization, and customer relationship management. However, most existing face recognition systems are optimized for controlled environments, which require high-resolution frontal-facing images. In real-world scenarios, such as conferences, public events, or group photographs, images often suffer from poor lighting, occlusions, non-frontal poses, or multiple individuals, posing significant challenges for accurate identification and profile generation. Recognizing unknown individuals in such unconstrained settings and generating detailed professional profiles (e.g., including name, job title) remains a complex, largely unresolved problem. Current technologies, such as Google Lens, Yandex [8], and Bing Visual Search [7], are primarily designed for object or product recognition and struggle with person-centric identification, especially in group photos. These systems often require carefully curated, portrait-style images and fail to deliver relevant or comprehensive results when processing low-quality or multi-person images. Moreover, they lack the ability to synthesize professional metadata or generate meaningful summaries, limiting their utility in applications such as social graph generation, talent scouting, or AI-driven networking.

This paper addresses the challenge of automated identity matching and the synthesis of professional profiles from unconstrained facial images. Our goal is to develop a system that not only identifies individuals in natural, real-world images but also aggregates relevant professional data (e.g., name, job title, publicly available image URLs, a set of curated positive links with target person information, negative links or misleading URLs and summary) into coherent, actionable profiles. This task includes several unique and realistic challenges. *Firstly*, a given URL may or may not contain information about the target individual and even if the link has information it may not be enough to get the identity of the person. *Secondly*, even in a positive or relevant URL, a single web page can include information about multiple people, making it difficult to isolate specific content for the person of interest. *Thirdly*, although some information about the target person may exist online in the positive link, it may not be easily extractable due to privacy restrictions. *Fourthly*, a negative or misleading link can hold picture of a person similar to the target person's face which might create confusion for face recognition and information extraction or it might contain the image of the target person but information to identify the person's affilation is missing. These challenges demand a system that can intelligently navigate noisy, multi-entity, and privacy-aware data sources to extract person-specific

knowledge with precision. By bridging facial recognition with interpretable information retrieval, this work enables advanced applications beyond mere identification, such as personalized event management and knowledge graph construction.

To achieve this, we propose a two-stage, data-driven system integrating facial recognition, clustering, information retrieval, and web mining. The system processes input images (containing one or more faces), extracts face embeddings using state-of-the-art models like DeepFace [20] or FaceNet [18], and performs local matching via cosine similarity against a curated dataset. If local matching is inconclusive, the system employs an online reverse image search, leveraging search engines and web scrapers to retrieve visually similar images. Textual metadata from these sources is then processed using natural language processing (NLP) and large language model (LLM)-based summarization to create structured professional profiles.

Central to our approach is an ethically sourced dataset of 9,472 professional profiles, compiled from trusted public sources such as Wellfound, Best Lawyers, university faculty directories, and legal networks. Each record includes high-resolution facial images and rich metadata, such as full name, job title, organization, education, nationality, and ethnicity. This dataset supports robust identity resolution under noisy or incomplete input conditions while adhering to ethical standards. To ensure privacy and fairness, we prioritize data from professional directories with implied consent for public use, avoid sensitive personal information, and implement safeguards to minimize bias in data collection and model outputs.

The key contributions of our research are as follows:

- **Curated Real-World Dataset**: We introduce a diverse, ethically sourced dataset of approximately 10,000 professional profiles with rich annotations, designed for facial identification and metadata extraction in unconstrained settings.
- **Hybrid Identity Matching Pipeline**: We propose a two-stage pipeline that combines fast local matching with face embeddings and clustering, and a web-based reverse image search and data scraping module for inconclusive cases. The pipeline supports both individual and group photographs.
- **Professional Profile Synthesis**: Using text metadata and LLM-based summarization (e.g., via GPT-4o), our system generates concise, human-readable profiles, enabling applications in networking, event organization, and knowledge graph population.

Preliminary experiments demonstrate promising results, achieving greater accuracy of identification than 86% on a standard set of real-world group photographs, significantly outperforming public services like Google Lens and Yandex. We also address ethical considerations, including consent, bias mitigation, and responsible data handling, to ensure the sensitive deployment of the technology. This research establishes a foundation for scalable, intelligent, and ethically aware facial identity systems that deliver actionable insights in complex, human-centric environments.

2 Related Work

Existing face recognition and profiling public Dataset: Several publicly available datasets [14, 5, 19, 23, 9, 12] have contributed significantly to facial recognition and demographic analysis by providing large-scale annotated image collections. The CelebA [14] dataset has around 200K celebrity face images, each with 40 binary descriptors including hair color, gender, and age, which makes it widely used for attribute prediction tasks such as smiling detection or gender classification. Megaface [12] is a collection of 690K identities with 1 million faces that were obtained from Flickr [21] under a Creative Commons license. It is a benchmark for commercial face recognition challenges and is one of the largest open-source face recognition datasets. Another similar dataset, Flickr-Faces-HQ (FFHQ) [11] crawled from Flickr, contains 70,000 high-quality images (1024×1024 resolution) of human faces. The data set includes more variation than CELEBA-HQ [10], which is a high-quality version of CelebA [14] in terms of age, ethnicity, and image background. It also has a much better coverage of accessories such as eyeglasses, sunglasses, hats, etc. The VGGFace2 [5] dataset includes 3.31 million photos from 9131 people retrieved from Google Image Search, which vary greatly in pose, age, illumination, ethnicity, and profession (e.g., actors, sports, politicians). To mitigate the race bias problem, the FairFace [9] dataset with 108,501 images collected from the YFCC-100M Flickr dataset with 7 race groups: White, black, Indian, East Asian, Southeast Asian, Middle Eastern, and

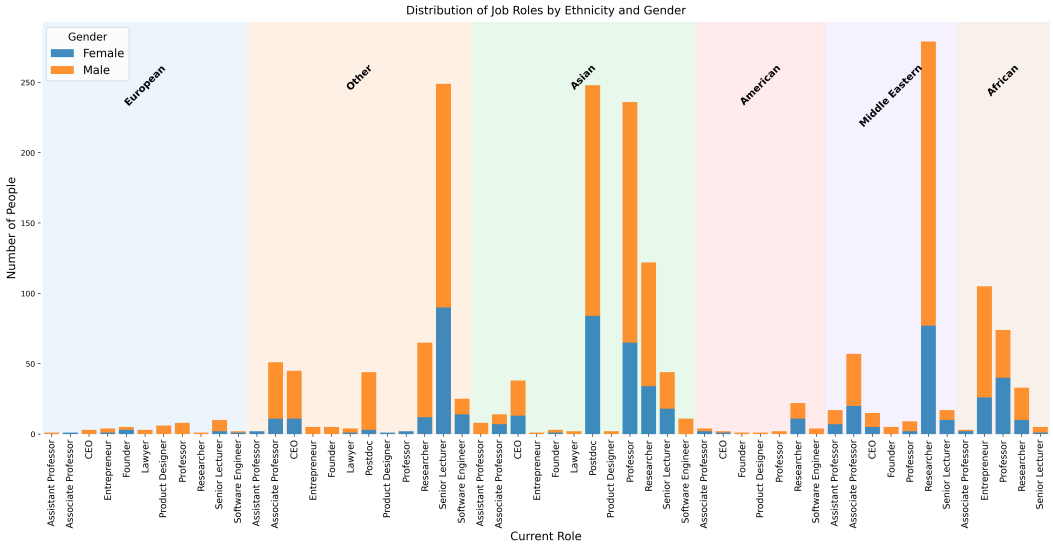
Latino were labeled with race, gender, and age groups. IMFDB (Indian Movie Face Database) [19] adds cultural and regional variety, enabling profiling in demographic contexts where Western-centric datasets fall short. The Tufts Face Database [15] exemplifies advances in multimodal recognition, including multiple imaging modalities (visible, near-infrared, thermal, computerized sketch, video, and 3D images) of the same subjects, allowing the development of more robust algorithms that work with different imaging types. It includes more than 10,000 images from 113 individuals from more than 15 different countries, various gender identities, ages, and ethnic backgrounds. Although these datasets focus on face recognition and a diverse demographic context, they lack multimodal identity profiling with expert-written summaries. On the other hand, our Face2Profile dataset includes not only face identification but also facial images with curated web-based evidence and peer-reviewed summaries that include fact-based biography generation and profile verification.

Retrieving personal information by private companies: In addition to public datasets, there are also proprietary web-scraped datasets built by private companies and platforms for large-scale face search and recognition applications, often used in open-source intelligence (OSINT) investigations. Facebook developed DeepFace [20], a deep learning facial recognition system that uses a facial alignment system based on explicit 3D modeling of faces. It greatly reduced the error rate, achieving an accuracy of 97.35% in the Labeled Faces in the Wild (LFW) dataset. A controversial face recognition tool, Clearview AI [6], is used by law enforcement, intelligence agencies, and private firms to identify people from images. It uses billions of images from public websites, including social media platforms such as Facebook, Instagram, LinkedIn, and YouTube, for criminal investigation, identification verification, and OSINT work. Similarly to Clearview, Pimeyes [3] also uses images scraped from public websites like blogs, news portals, personal pages, and social media profiles. It is a face search engine that allows users to upload a face photo and find other publicly available links containing the image of the same person on the Web. In addition, a search engine for information named FaceCheck [1] lets users upload a face image and search for similar photos online. Like PimEyes, it is made for OSINT and threat intelligence. It uses a proprietary dataset of publicly available web-scraped photos for face recognition and search. Another similar API for facial recognition is Kairos [2], which provides identity verification, emotion detection, gender, and age. It builds its algorithms on a proprietary dataset created from images collected from the internet. In contrast to these proprietary web-scraped datasets, we present Face2Profile, which will be a publicly available dataset that includes an expert-written summary of the person in the image from the curated links, which could be found using PimEyes or a similar face search engine.

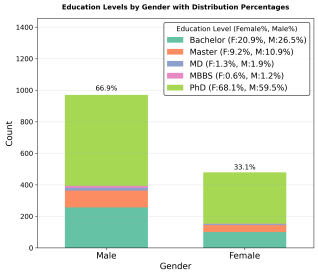
Personal profile generation: Personal profile generation involves identifying the person and structurally representing their identities by integrating data from various sources. Early efforts in this domain, such as [4], present a system for extracting personal data from the Internet to generate comprehensive user profiles of people in Spain, employing Big Data methodologies through web crawlers and scrapers. Addressing the challenge of extracting information, [13] proposes a convolutional neural network for data region location with a visual information-based segmentation algorithm, treating web pages as images to locate data regions with high accuracy. As profile generation tasks become more nuanced, recent approaches have incorporated large language models(LLMs) to enhance the extraction and synthesis of biographical information. For example, [17] addresses the challenge of disambiguation of names by using LLMs to accurately link the web profile. Moreover, to personalize guided profile generation [22] employs LLMs to summarize and extract significant, distinctive features from the personal context into concise, descriptive sentences, which results in closely tailoring their generation to the unique habits and preferences of an individual. Despite these advancements, most existing approaches focus either on isolated modality, text, or only images, or operate in constrained, domain-specific settings. Building on these foundations, our Face2Profile introduces a multimodal dataset explicitly designed to support identity-based profile generation from both facial imagery and web content.

3 Data Collection

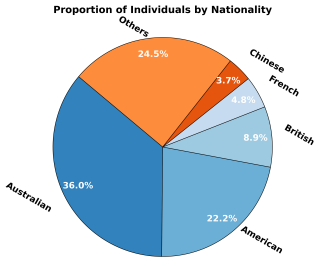
The construction of our Face2Profile dataset involved a meticulous profile collection process and systematic summary generation, aimed at developing a diverse and comprehensive resource for Human Information Extraction. During data collection, we ensured representation across a wide range of professions, genders, and ethnic backgrounds.



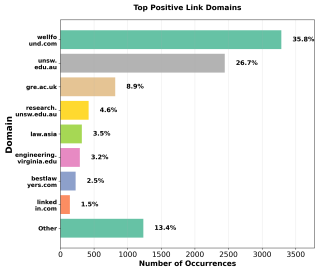
(a) Distribution of Job Roles by Ethnicity and Gender, Showing Variation in Occupational Representation Across Demographic Groups.



(b) Education level by gender distribution



(c) Proportion of inidivuals by Nationality



(d) Distribution of positive link domains

Figure 2: Statistics of the Face2Profile Dataset. More statistics and dataset examples are given in Appendix 7.

3.1 Dataset Collection Design

Data Composition The Face2Profile dataset is designed to facilitate research on structured information extraction from multimodal human-centric data. Each entry in the dataset consists of an image of a person, their name, job title, a collection of positive and negative web links, and a human-written summary. Positive links contain verifiable information about the individual depicted in the image, while negative links include irrelevant content or profiles of individuals who appear similar but are unrelated to the target subject. The Figure 2d gives a visualization of the positive web link domain distribution, where domains contributing less than 1.5% of individuals are grouped under "Other". The data reveal that the largest share of links is from wellfound.com, followed by unsw.edu.au, gre.ac.uk, research.unsw.edu.au, law.asia, engineering.virginia.edu, bestlawyers.com, angelinvestmentnetwork.us, linkedin.com and others respectively, which represent academic, legal and professional directories used in Face2Profile.

Diversity Considerations In constructing the Face2Profile dataset, we prioritized diversity by including individuals from a broad spectrum of professional fields, gender identities, and ethnic backgrounds. The pie chart in figure 2c highlights the nationality distribution and the bar chart in figure 2b shows the education level (PhD, Bachelor, Master, MD, and MBBS) grouped by gender distribution, ensuring that the dataset reflects the heterogeneity of real-world populations, which will enhance the robustness of the models trained on it. The goal was to mitigate demographic and occupational bias during both the training and evaluation phases.

Data Collector The dataset was collected by a team of four individuals. During the collection process, we ensured representation across multiple ethnicities, genders, and a wide range of professions to promote diversity. Additionally, we deliberately included challenging facial samples, such as partially occluded, low-resolution, or nonfrontal images, to increase the complexity of the dataset and better reflect real-world conditions.

Summary Writer To ensure high-quality and contextually relevant summary generation, we employed three dedicated summary writers for this project. Each writer was responsible for reviewing the image and positive links associated with a given individual to craft a concise and informative summary. These writers possess strong research and writing skills, with experience in synthesizing biographical and professional information from diverse online sources. Rather than relying on crowdsourced annotations, we opted for a smaller, expert-driven team to maintain consistency and accuracy across the dataset. All three writers were thoroughly briefed on the task guidelines to ensure alignment in style, tone, and factual precision.

Table 1: Dataset structure of Face2Profile. Each entry primarily consists of the following components.

Field	Description
Image	A clear photograph of the individual.
Name	Full name of the individual.
Job Title	Profession or primary role.
Positive Links	URLs with verified personal information.
Negative Links	Misleading or unrelated URLs.
Summary	Human-written paragraph synthesizing key biographical and professional details.

3.2 Dataset collection Process

The construction of our Face2Profile dataset followed a structured collection and annotation workflow designed to ensure both diversity and quality. In the following, we detail the key steps undertaken during the dataset creation process.

Demographic Stratification To promote representational fairness, the data collection team began by stratifying the subjects according to ethnicity. Within each ethnic group, individuals were selected across a range of professions and gender identities which is presented in figure 2a. This hierarchical approach ensured that the dataset captured a broad spectrum of human profiles reflecting real-world demographics.

Summary Generation and Evaluation Each data instance in the Face2Profile dataset includes a concise, human-written summary intended to capture key biographical and professional details of the individual, derived from their image and associated positive links. To maintain consistency and ensure high-quality annotations, a peer-reviewed system for summary generation was employed.

For each individual, three different summary writers independently generated candidate summaries. Once all three summaries were written, a scoring phase was started. Each writer was instructed to evaluate the other two summaries (excluding their own) using a 10-point scoring rubric. The evaluation criteria included accuracy, clarity, information, and alignment with the provided image and positive links. Let S_{ij} represent the score given by the writer i to the summary j, where $i \neq j$ and $i, j \in \{1, 2, 3\}$. The final score for a summary j is computed as:

$$\text{Score}(j) = \frac{1}{2} \sum_{\substack{i=1 \\ i \neq j}}^{3} S_{ij} \tag{1}$$

The summary with the highest average score was selected as the final reference summary for that data instance:

$$\text{Final Summary} = \arg \max_{j \in \{1,2,3\}} \text{Score}(j) \tag{2}$$

This peer review approach served two purposes: ***(1)*** it encouraged writers to be thoughtful and precise in their own summaries, knowing that they would also serve as reviewers; and ***(2)*** it created a lightweight but effective consensus mechanism to determine the highest-quality summary without the need for an external adjudicator.

This method of summary selection ensures that the final dataset includes only the most informative and well-articulated descriptions.

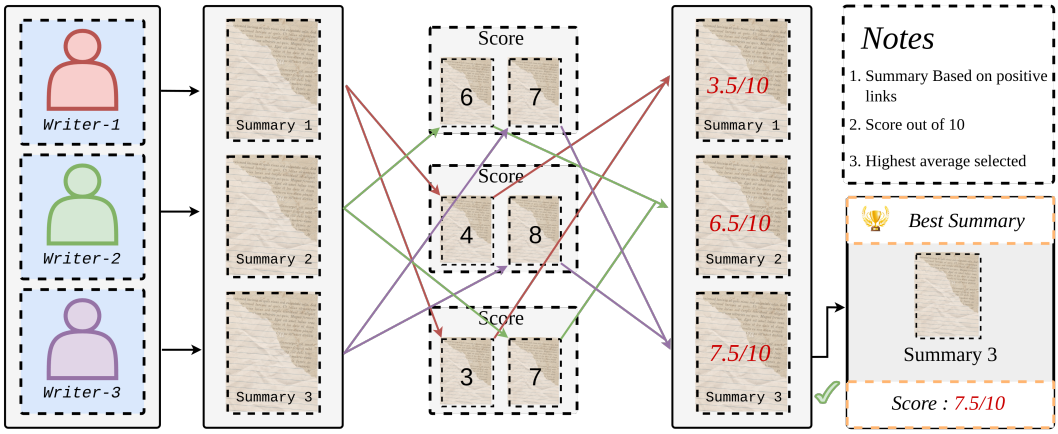


Figure 3: Overview of the summary generation and peer-review process in the Face2Profile dataset. For each individual, three independent writers produce candidate summaries based on the image and positive links. Writers then evaluate each other's work using a structured rubric. The highest-scoring summary is selected as the final reference, ensuring quality through collaborative assessment.

4 Experimental Setup

4.1 Summary Quality Evaluation with GPT-4o and Custom BLEU

To assess the effectiveness of human-generated summaries in our dataset, we performed an experiment using GPT-4o, where the model generated a summary for each individual based on their set of positive and negative links. We then compared the GPT-4o generated summary with the human-written reference summary using the BLEU score [16] and a modified BLEU score.

In addition to standard n-gram precision metrics (as in BLEU-4), we introduced two binary penalties: one for the presence of the individual's full name and another for the job title. If either was missing from the candidate summary, the BLEU score was penalized accordingly. The custom score is defined as:

$$\text{Custom-BLEU} = \text{BLEU-n} \times (\delta_{\text{name}} \times \delta_{\text{job}}) \tag{3}$$

Where BLEU-n is the standard BLEU score calculated over 1, 2, and 4-grams; $\delta_{\text{name}} = 1$ if the full name is present in the generated summary, otherwise 0.9; and $\delta_{\text{job}} = 1$ if the job title is present, otherwise 0.9.

This formulation ensures that summaries that omit crucial identity-related details receive a lower score, encouraging completeness and fluency.

4.2 Small LLM Evaluation: Phi-3 Mini and DeepSeek-1.5B

To assess the ability of lightweight language models to generate factually accurate summaries, we evaluated Phi-3 Mini and DeepSeek-1.5B using only structured input: the individual's full name and a set of positive links (the image was not passed to the model).

Prompt Design: Each model was given a prompt that included the individual's name and a formatted list of brief descriptions extracted from the positive links (e.g., headlines, meta-descriptions or first sentences). The task was to generate a concise summary that reflects the most important biographical and professional details.

Evaluation Strategy: Instead of BLEU, we used an **entity coverage score**, which measures how many key facts are correctly included in the generated summary. These facts, manually extracted from the positive links, are categorized as E_{name} (full name), E_{job} (job title or role), and E_{facts} (additional details such as organization, awards, location, or affiliations).

The *entity coverage score (ECS)* is calculated as:

$$\text{ECS} = \frac{\sum_{i=1}^{N} \left(\mathbb{1}_{\text{name}_i} + \mathbb{1}_{\text{job}_i} + \sum_{j=1}^{k_i} \mathbb{1}_{\text{fact}_{ij}} \right)}{\sum_{i=1}^{N} (1 + 1 + k_i)} \tag{4}$$

Where N is the number of evaluated samples, k_i is the number of additional factual entities in the i-th reference, and $\mathbb{1}_x$ is an indicator function that returns 1 if the entity x is present in the generated summary, and 0 otherwise.

This score ranges from 0 to 1, with higher values indicating better factual consistency and coverage. The evaluation emphasizes the ability of the model to recover named entities and key professional attributes from structured input.

Result Interpretation: This method offers a lightweight and interpretable way to assess small LLM performance without relying on pre-trained metrics or access to vision features, making it especially suitable for resource-constrained settings.

5 Results

This section evaluates language model performance on the Face2Profile dataset, with separate analyses for small and large language models. We use ECS to assess entity consistency in small models, and BLEU, along with a Custom-BLEU metric, to evaluate large models.

5.1 Performance of Small Language Models

Table 2 reports the ECS scores for two representative SLMs—Phi3-mini and DeepSeek-R1-1.5B—on the Face2Profile dataset. To disentangle the contribution of identity-bearing tokens, we compute ECS under three settings: $\text{ECS}_{\text{noname}}$, $\text{ECS}_{\text{nojob}}$, and full ECS (i.e., without masking).

Table 2: ECS Score Comparison across Small Language Models

Model	$\text{ECS}_{\text{noname}}$	$\text{ECS}_{\text{nojob}}$	ECS
Phi-3-mini-3.8B	0.294	0.577	0.435
DeepSeek-R1-Distill-Qwen-1.5B	0.299	0.599	0.449

A closer inspection of Table 2 reveals that both SLMs exhibit a significant discrepancy between $\text{ECS}_{\text{noname}}$ and $\text{ECS}_{\text{nojob}}$, suggesting a greater reliance on job titles than names for biographical inference. DeepSeek-R1-1.5B slightly outperforms Phi3-mini across all ECS variants, reflecting a marginally enhanced contextual encoding of identity-linked semantics. However, the overall ECS scores remain moderate, pointing to the inherent limitations of small-scale models in resolving nuanced biographical consistency, particularly when critical identifiers are suppressed.

5.2 Performance of Large Language Models

We further evaluate two large-scale models—GPT-4o and DeepSeek-R1 (full version)—using both standard BLEU and Custom-BLEU (C-BLEU) metrics. While BLEU assesses general n-gram overlap, C-BLEU imposes structured penalties for omitting key biographical entities such as full names and professional designations.

Table 3: Comparison of GPT-4o and DeepSeek-R1 using Standard BLEU and Identity-Aware Custom-BLEU (C-BLEU) Metrics.

Model	BLEU-1	BLEU-2	BLEU-4	C-BLEU-1	C-BLEU-2	C-BLEU-4
GPT-4o	0.287	0.140	0.069	0.247	0.120	0.059
DeepSeek-R1	0.252	0.108	0.049	0.218	0.094	0.042

From Table 3, GPT-4o consistently surpasses DeepSeek-R1 across all n-gram levels for both BLEU and C-BLEU metrics. The relatively high BLEU-1 score of 0.287 indicates robust lexical alignment at the unigram level, whereas the diminishing returns at BLEU-4 (0.069) reflect a loss of syntactic cohesion in longer sequences. The Custom-BLEU scores, in contrast, underscore the stringency of our identity-sensitive metric: GPT-4o's C-BLEU-1 drops to 0.247 and C-BLEU-4 to 0.059, highlighting penalties for omission or distortion of entity-rich tokens.

DeepSeek-R1 demonstrates a similar decay across increasing n-gram levels but underperforms GPT-4o in all configurations. Its C-BLEU-4 score of 0.042 suggests particular challenges in preserving identity coherence in longer spans of generated text. These disparities emphasize the superior fidelity of GPT-4o to biographical integrity under stricter semantic evaluation regimes.

5.3 Discussion

Comparing ECS and C-BLEU scores highlights a trade-off between model size and how well meaning is preserved. Smaller models show predictable behavior under controlled input changes (ECS), while larger models like GPT-4o are more expressive but need stricter metrics (like C-BLEU) to judge how well they maintain identity in summaries. GPT-4o performs best likely due to its size, training data, and tendency to preserve facts when summarizing biographical text. Still, it struggles with longer or more complex phrases, as seen in its lower BLEU-4 and C-BLEU-4 scores, showing that consistency in detailed summaries remains a challenge. These findings suggest that identity-aware metrics are not optional, they are crucial for evaluating biographical summarization. Future work could improve these metrics further, for example by adapting them to different languages or adjusting how they weigh different parts of the text.

6 Limitations

Although the Face2Profile dataset represents a significant advance in human-centric information extraction, we acknowledge several limitations that may affect its applicability and generalizability. The dataset relies on information scraped from public sources which implies availability and depth of information vary significantly between individuals, particularly for those with limited online presence. Moreover, Despite efforts to ensure diversity between professions, gender, and ethnicities, the dataset, with approximately 10,000 profiles, captures only a small fraction of the global population. This constraint limits the representation of certain ethnic groups and underrepresented communities, which could affect model performance when applied to broader or more diverse populations.

7 Conclusion

In this work, we introduced a large-scale dataset designed to advance human-centric information extraction research. By incorporating structured annotations—including facial images, biographical details, curated web evidence, and peer-reviewed summaries—we aim to provide a high-fidelity benchmark for evaluating factual consistency and identity-aware summarization. We also implemented some baseline methods using state-of-the-art LLM models like GPT-4o and small LLMs, Phi-3 Mini and DeepSeek-1.5B, which take a person's name and positive links (URLs that contain the person's information) as input and output a summary of the person. To evaluate the LLM-generated summaries, we proposed a novel Custom-BLEU metric that extends traditional BLEU scoring with penalties for missing core identity elements such as full name and job title. Our experiments with GPT-4o show that while the model demonstrates fluent generation capabilities, it frequently omits essential factual components, as reflected in the drop from BLEU to Custom-BLEU scores. This highlights the limitations of conventional n-gram metrics in identity-sensitive summarization tasks and underscores the value of structurally guided evaluation.

We hope that the proposed dataset is a valuable benchmark for developing and evaluating models that combine vision, language, and structured reasoning. Future research will look at fine-grained fact verification and bias reduction measures to ensure fairness between demographics.

References

[1] Facecheck.id: Reverse image search and facial recognition engine. `https://facecheck.id/`. Accessed: 2025-04-29.

[2] Kairos face recognition api documentation. `https://face.kairos.com/docs/api/`. Accessed: 2025-04-29.

[3] PimEyes: Facial Recognition Search Engine. `https://pimeyes.com/`. Accessed: 2025-04-22.

[4] Á. Bartolomé, D. García-Retuerta, F. Pinto-Santos, and P. Chamoso. Internet data extraction and analysis for profile generation. In *Ambient Intelligence–Software and Applications–, 10th International Symposium on Ambient Intelligence*, pages 112–119. Springer, 2020.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[6] C. Dul. Facial recognition technology vs privacy: The case of clearview ai. *QMLJ*, page 1, 2022.

[7] H. Hu, A. Sacheti, Y. Wang, L. Yang, P. Komlev, L. Huang, X. Chen, J. Huang, Y. Wu, and M. Merchant. Web-scale responsive visual search at bing. pages 359–367, 07 2018. doi: 10.1145/3219819.3219843.

[8] S. M. Jones and D. Oyen. Abstract images have different levels of retrievability per reverse image search engine. In L. Karlinsky, T. Michaeli, and K. Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 203–222, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25085-9.

[9] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[11] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[12] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[13] J. Liu, L. Lin, Z. Cai, J. Wang, and H.-j. Kim. Deep web data extraction based on visual information processing. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–11, 2024.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[15] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, et al. A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):509–520, 2018.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[17] P. Sancheti, K. Karlapalem, and K. Vemuri. Llm driven web profile extraction for identical names. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1616–1625, 2024.

[18] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

[19] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. Karure, R. Raju, B. Rajan, et al. Indian movie face database: a benchmark for face recognition under wide variations. In *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*, pages 1–5. IEEE, 2013.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[21] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1 (8), 2015.

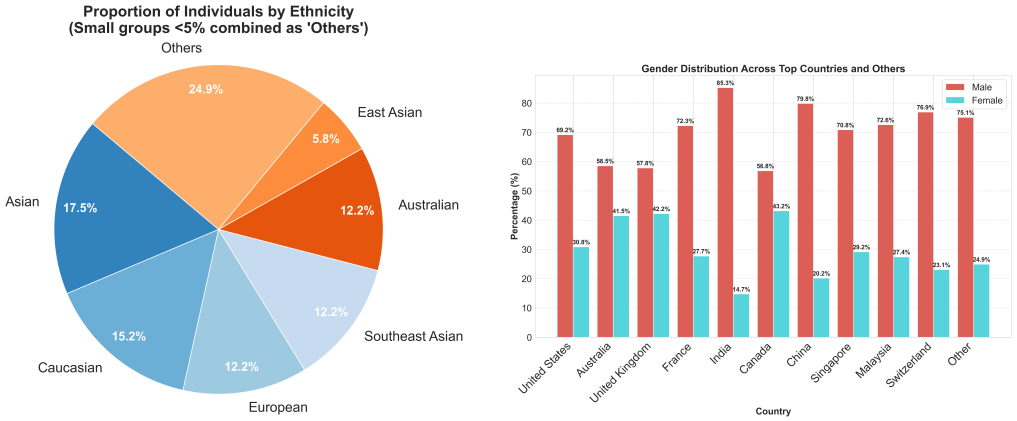
[22] J. Zhang. Guided profile generation improves personalization with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4005–4016, 2024.

[23] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

A Dataset Details

A.1 Different Statistics of the Dataset

The Figure 4a illustrates the proportion of individuals in Face2Profile grouped by ethnic categories, highlighting demographic diversity and ensuring balanced representation across ethnicity. In order to improve clarity and ensure statistical robustness, ethnic groups with fewer than 5% are categorized under the "Others" category, which makes up 24.9% of the dataset. The largest single group is Asian (17.5%), followed by Caucasian (15.2%), European (12.2%), Southeast Asian (12.2%), Australian (12.2%) and East Asian (5.8%), respectively. Besides, the bar chart 4b presents the percentage distribution of male and female individuals across different countries. It is observed that male individuals are overrepresented, with the highest disparity observed in India and China, respectively. In contrast, the United Kingdom and Australia show a comparatively more balanced gender representation. This demographic skew along gender lines shows, inspite of deliberate stratification efforts during dataset construction, real-world visibility and online presence biases likely contribute to this imbalance.



(a) Proportion of Individuals by Ethnicity

(b) Gender distribution Across Top Countries and Others

Figure 4: The Ethnic diversity and gender distribution across multiple countries in Face2Profile

A.2 Sample of Dataset

The Table 4 shows a qualitative snapshot of the Face2Profile dataset, which includes variation in demographics, professions, and image quality, reflecting real-world challenges. Each profile consist of publicly available facial images, the full name and professional history including job title, a set of positive links can contain verifialbe biographical information about the person, a set of negative links which contain link of person similar to target person or the links which doesn't hold enough info to identify the person and a peer reviewed summary. Besides, we also have a field that holds the URLs of the publicly available images.

Table 4: Sample Person Profiles from Face2Profile Dataset

Image	Name & Job Description	Positive Links & Negative Links	Summary
	**Name:** Adrian Rami **Job Title:** • C3 AI - Product Manager (2015 – Present, over 9 years)	**Postive Links:** • `https://rocketreach.co/adrian-rami-email_76749324` **Negative Links:** • `https://wellfound.com/u/adrian-rami`	Adrian Rami is an experienced Product Manager based in the San Francisco Bay Area, with a decade of expertise at C3 AI. Since 2015, Adrian has been instrumental in driving product development and innovation within the company, leveraging a solid foundation in engineering and business management. Adrian holds a Master of Business Administration (MBA) and a Bachelor of Science (BS) in Chemical Engineering, equipping them with a unique blend of technical and managerial skills. Adrian is committed to delivering impactful solutions that enhance organizational effectiveness and drive market success.
	**Name:** Rowena Guo **Job Title:** • Guideline - Operations • Manager - Insurance Huntington Beach	**Postive Links:** • `https://www.linkedin.com/in/rowenaguo/` • `https://rocketreach.co/rowena-guo-email_6723226` **Negative Links:** • `https://www.pinterest.com/rowguo/`	Rowena Guo is an Operations professional based in the San Francisco Bay Area, bringing expertise derived from her experience at Guideline. She holds a Bachelor of Science in Business Administration with a concentration in Finance from San Francisco State University, graduating in 2016. Known for her strong analytical skills and commitment to operational excellence, Rowena is dedicated to driving efficiency and effectiveness within her role.

Table 4: Sample Person Profiles from Face2Profile Dataset

Image	Name & Job Description	Positive Links & Negative Links	Summary
	**Name:** Professor A. K. M. Fazlul Haque **Job Title:** • Government Health Service: Worked from April 1982 to 2009 • International Experience: Registrar, Department of Surgery, Singapore General Hospital (1995–1996)	**Positive Links:** • https://profdrakmfaz..	Professor A. K. M. Fazlul Haque is a distinguished Bangladeshi colorectal surgeon with extensive experience in government health services, having served from 1982 to 2009. He gained international expertise as a registrar in the Department of Surgery at Singapore General Hospital from 1995 to 1996. Currently, he leads the Department of Colorectal Surgery at his own clinic, Dr. Fazlul Haque Colorectal Hospital Limited, renowned as one of the best healthcare facilities in Bangladesh for treating piles and related conditions. Professor Haque holds an M.B.B.S degree from Dhaka Medical College, complemented by a Secondary School Certificate and a Higher Secondary School Certificate. He is committed to delivering high-quality healthcare and has contributed significantly to the medical community in Bangladesh.

Table 4: Sample Person Profiles from Face2Profile Dataset

Image	Name & Job Description	Positive Links & Negative Links	Summary
	**Name:** Sandra Janus **Job Title:** • Harry's - Manager, Operations (2017 - Present (about 8 years)) • Morgan Stanley - Investment Banking Analyst (2014 - 2015 (11 months))	**Postive Links:** • `https://www.linkedin.com/in/sandrajanus/` **Negative Links:** • `https://medium.com/@sandrajanus/about`	Sandra Janus is an experienced Operations Manager based in New York City, currently leading operations at Harry's since 2017. She began her career as an Investment Banking Analyst at Morgan Stanley. Sandra holds a Bachelor of Arts in Honors Business Administration from the Richard Ivey School of Business and an Honors Specialization in Health Sciences from the University of Western Ontario, both completed in 2012.
	**Name:** Ms. Bridie Mary Moran **Job Title:** • Curator - Newcastle Museum (Full-time) • PhD Candidate - UNSW Art and Design (Full-time) • Executive Director and Acting Director - 4A Centre for Contemporary Asian Art (Full-time) • Co-Editor - The Journal of Australian Ceramics (Part-time) • Partnerships and Marketing Manager - The Walkley Foundation (Full-time)	**Positive Links:** • `https://www.unsw.edu.au/staff/bridie-moran` • `https://www.unsw.edu.au/hdr/bridie-moran` • https://research.unsw..	Bridie Mary Moran is a highly skilled researcher, curator, and arts manager with over 15 years of experience in the contemporary art and cultural sectors of Australia. Currently serving as the Curator at Newcastle Museum, Bridie is pursuing a PhD at UNSW Art and Design, focusing on the historical context and policy implications of craft in Australian arts and culture. She has held significant leadership roles, including Executive Director at 4A Centre for Contemporary Asian Art, where she greatly contributed to engagement and programming. Bridie is also a Co-Editor of The Journal of Australian Ceramics and has provided consultancy for notable organizations such as Creative Australia and Lake Macquarie City

Table 4: Sample Person Profiles from Face2Profile Dataset

Image	Name & Job Description	Positive Links & Negative Links	Summary
	**Name:** Ethan Ruby **Job Title:** • Craft Ventures - Venture Investor (2018 - Present (about 7 years)) • Zenefits - Business Operations Manager (2017 - 2018 (about 1 year))	**Positive Links:** • `https://wellfound.com/u/ethan-ruby` • https://www.linkedin..	Ethan Ruby is an accomplished venture investor at Craft Ventures, where he has been instrumental in identifying and nurturing innovative startups for approximately seven years. Previously, he served as a Business Operations Manager at Zenefits, contributing significantly to the company's operational strategies as one of its early employees. Ethan holds a Bachelor of Arts in Public Policy Studies, with a concentration in Asian and Middle Eastern Studies, from Duke University (2014). He is passionate about building impactful companies and leveraging his expertise to drive growth in the entrepreneurial ecosystem.